

EXPLORATION AND DISCOVERY OF DIFFERENTIALLY EXPRESSED GENE SIGNATURES: A COMPREHENSIVE DATA-DRIVEN APPROACH UNVEILING POTENTIAL BIOMARKERS***¹Ms. T. Sangeetha**

Research Scholar,

²Dr. K. Manikandan

Associate Professor,

³Dr. D. Victor Arokia Doss

Associate Professor,

^{1,2}Department of Computer Science, ³Department of Biochemistry^{1,2,3} PSG College of Arts & Science, Coimbatore - 641014, Tamil Nadu, India¹thangarajusangeetha@gmail.com, ²prof.k.manikandan@gmail.com,³victordoss64@gmail.com

*Corresponding Author

Abstract

Purpose: Transcriptomics has been revolutionized by the development of microarray technology, which makes it possible to simultaneously measure thousands of genes' levels of gene expression. This innovation holds an immense potential in understanding cardiovascular diseases such as Ischemic Cardiomyopathy (ICM) and Non-Ischemic Cardiomyopathy (NICM), which present substantial health concerns on a global scale implying the need for studying ICM and NICM exhaustively. The primary objective of this proof-of-concept paper aims at uncovering potential biomarkers and learn using data-driven method to identify important genes that are differentially expressed.

Methods: Microarray data from Gene Expression Omnibus (GEO) repository provided the dataset, which includes expression data from peripheral blood mononuclear cells (PBMC) of patients with ischemic and non-ischemic cardiomyopathy as well as a control group that was age and gender-matched. This research paper endeavours to conduct comprehensive microarray data analysis for transcriptomic profiling aimed at the identification of differentially expressed genes (DEG) associated with cardiomyopathy. Leveraging a data science process model, this study delves into the exploration and interpretation of a specific dataset, GDS3115, curated for its relevance to cardiomyopathy.

Results: In total, five DEG showing significant differences in their Gene Expression Profiles to make diagnostic / prognostic analysis were identified. The inferences are tabulated and plotted the DEG in volcano plot as an interpretation of result obtained.

Conclusion: Candidate biomarker genes such as CX3CR1C, HSPA1L///HSPA1B///HSPA1A, JUN, ZNF331, RORA are ICM's therapeutic targets. This study identified several DEG that may be involved in the pathogenesis of ICM/NICM. This abstract synthesizes the research idea, workflow, methodologies

employed, and the potential implications of the study in identifying cardiomyopathy related genes via Biological analysis using the GDS3115 dataset.

Keywords: ischemic cardiomyopathy, drugs, non-ischemic cardiomyopathy, Robust Multi-Array(RMA), IQR, Gene Filtering.

Introduction:

The last stage of coronary artery diseases are ICM which have coronary artery constrictions, reactive cellular hypertrophies, myocyte deaths, and ventricular scars as characteristics [9]. This kind of cardio myopathy carries a significant danger to one's health because of the high rate of sudden cardiac death among ICM patients worldwide [1]. Surgical vascular bypass, interventional angioplasty, and medication therapy are the three main traditional treatment modalities for ICM [2]. Nonetheless, certain patients' vascular lesions reveal tiny vessel illnesses that are unsuitable for vascular obstructive intervention or surgery [12]. Therefore, novel treatments for ICM that meet present clinical requirements are required.

The etiology of cardiomyopathy has been successfully predicted by gene expression profiling [5]. Furthermore, the topological structure of biological networks has been used to identify a few putative disease-related gene markers [7, 8]. Hence, applying a bioinformatics method might help identify new biomarkers for treating cardiomyopathy. This work used microarray data analyses based on gene expression profile (GDS3115) in order to explore and identify novel biomarkers. Additionally, it was envisaged that by screening the new biomarkers, additional understanding of the molecular underpinnings of ICM would be obtained. This might aid in the development of new medicines for ICM as well as the selection of a suitable treatment plan. Patients are not [5] receiving treatment in a timely manner and there are very few treatment options because to the uncertain mechanism behind ICM. Novel biomarkers are therefore important to research and find because they can help with ICM/NICM diagnosis and preventive care.

GEO database provided microarray data for this investigation where DEG associated with ICM/NICM were determined. Figure 1 displays the flowchart of this work's suggested design.

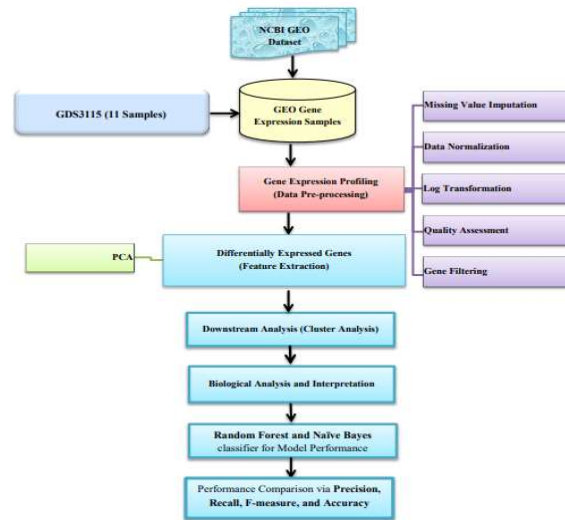


Figure 1: Proposed Workflow Architecture

In the study, the research methodology involves the application of robust statistical algorithms, notably the Robust Multi-Array (RMA) algorithm and interquartile range (IQR), to pre-process and eliminate noise from the microarray data, ensuring the reliability of subsequent analyses. Hierarchical clustering, specifically using the complete linkage method, is utilized to reveal distinct gene expression patterns among different patient cohorts, potentially illuminating biomarkers or pathways relevant to different forms of cardiomyopathy. Principal component analysis, downstream analysis, and functional enrichment analysis were used in investigations of ICM's underlying mechanism. Future research may find it useful to refer to medication predictions associated with gene identifications and obtain information required for this study i.e. identifying new biomarkers and treatment targets for ICM/NICM.

Materials and Methods

Data pre-processes: Affy package [4] in R was used to pre-process raw CEL format data which included background corrections and normalisations. Limma package [4] in R statistically compared gene expression patterns of ICM and control groups. Genes were deemed substantially different if its $-\log_{10}(p \text{ value}) > 5$ and $\log_2\text{FC}$ (fold change) > -1.1 and < 2 . Using pheatmap of R[11], hierarchical clustering [15, 3] were carried out on DEG expression levels based on Euclidean distances.

Proposed Architecture According to the architecture proposed in the Figure 1, the workflow began with gene expression profiling and data pre-processing utilizing the GDS3115 dataset consisting of 11 samples. This phase encompassed missing value imputation, data normalization, \log_2 transformation, quality assessment, and gene

filtering to ensure robustness and reliability in subsequent analyses. The following phase is focused on identifying DEG as a critical feature extraction process. A top table was generated, resulting in the identification of five DEG, comprising two up-regulated and three down-regulated genes, providing initial insight into potential molecular signatures associated with the studied cardiovascular conditions. Finally a comprehensive biological analysis and interpretation were performed. Principal Component Analysis (PCA) was utilized for dimensional reduction and visualization, providing an overview of sample relationships. Subsequently, downstream analyses, including Hierarchical Clustering, offered insights into potential gene expression patterns and clustering within the dataset.

R Programming: R is a versatile statistical programming language commonly used in bioinformatics for analyzing microarray data, offering a rich ecosystem of packages like limma and DESeq2 that facilitate differential gene expression analysis and visualization through statistical models and graphical representations.

Python: Python, with libraries such as pandas, NumPy, and scikit-learn, is increasingly utilized in transcriptomic analysis for preprocessing microarray data, conducting statistical tests, and performing machine learning algorithms, providing a flexible and powerful environment for gene expression studies.

Hierarchical Clustering: Hierarchical clustering, using the complete linkage method, organizes genes or samples based on their similarity, forming clusters by considering the maximum distance between all possible pairs of elements from two different clusters, commonly employed in microarray data analysis to identify distinct expression patterns among genes or experimental conditions.

Robust Multi-Array (RMA) Algorithm: RMA is a robust preprocessing algorithm widely applied in microarray data analysis, known for its ability to normalize and summarize probe-level intensities, reducing technical variations across arrays, and enhancing the accuracy of detecting DEG by improving data quality.

Interquartile Range (IQR) in Microarray Data Analysis: The IQR, a measure of statistical dispersion, is used in microarray data analysis to identify outliers and filter noise by calculating the range between the first and third quartiles of expression values, aiding in the identification of significantly DEG by minimizing outlier value impacts for improved robustness of analyses.

The integration of R programming facilitated the workflow, allowing for statistical analyses and visualization, while Python aided in generating essential visualizations like scatter plots and volcano plots. The utilization of multiple analytical tools and methodologies provided a holistic understanding of the molecular landscape associated with ICM, NICM, and healthy controls.

Data Science Process Model

Using the above concepts data science process model is designed to identify the differential expression genes. The data science model of the study is presented in Figure



2.

Figure 2: Data Science Process Model

Objective of the model: The primary objective of this proof-of-concept aimed to uncover potential biomarkers and gain insights into the molecular mechanisms underlying using a comprehensive data- driven approach.

Description of Data Collection: The dataset GDS3115, obtained from the GEO Repository and originating from human (*Homo sapiens*) samples, comprises 11 distinct samples intended for analysis. These samples were segregated into three patient groups for study:

1. ICM: This subset involved 12 individuals diagnosed with ischemic cardiomyopathy.
2. Non - Ischemic Cardiomyopathy (NICM): Consisting of 12 individuals diagnosed with non-ischemic cardiomyopathy, specifically NYHA III-IV CHF patients.
3. Control Group: Comprised of 12 age- and gender-matched individuals, this group served as controls for comparative analysis.

This dataset, with its comprehensive analysis of gene expression in PBMCs from individuals with heart failure and matched controls, provides valuable insights into the molecular mechanisms underlying different forms of cardiomyopathy. It serves as a resource for researchers investigating genetic and molecular underpinnings of heart failures and potentially contributes to the identification of novel therapeutic targets or diagnostic biomarkers for this prevalent cardiovascular condition.

Exploratory Data Analysis (EDA) is the third step in the data science process model as represented in Figure 2. It is the statistical approach of analyzing the data. This step plays a vital role in understanding of the data and summarizing the data through visuals. The graphs and plots provide crisp insights about the data as represented in Figure 3.

Dimensionality Reduction: At this stage, an analysis of the relationship between variables is very essential. The significance of using this step before model building is to analyze the relationship between variables, identifies the underlying patterns, and aims at reducing the number of dimensions by replacing them with latent variables called factors. The output of this step is represented in Figure 7 and 8.

Model Building: The Naïve Bayes algorithm and Random Forest is used as a classifier to measure the performance of the dataset GDS3115. The Classifiers both used are supervised learning technique that can be used for both classification and regression. It is considered one of the fastest and most accurate algorithms used in prediction, especially for large datasets. The Naïve Bayes algorithm

technique could be used for multi-class classification (when the response variable is not binary and has more than two classes) also. This is based on Bayes Theorem. In the machine learning context, a is the response variable and b_i are the predictor variables representing $b_1, b_2, b_3 \dots b_n$. The Naïve Bayes formula is as follows:

$$P\left(\frac{a}{b}\right) = P\left(\frac{b}{a}\right) * \frac{P(a)}{P(b)} \quad (1)$$

Where : $P(a)$ implies probabilities of response variables, $P(b_i)$ represents probabilities of predictors, $P(a/b_i)$ stands for conditional probabilities of response variables assuming variables for predictors (predictions), $P(b_i/a)$ implies conditional probabilities of occurrence of predictor variables given response variables (training data).

Future directions for investigation: In the near future, many longitudinal samples will need to undergo additional experimental validations of presented results. The expression levels of biomarkers identified in this investigation should be assessed in people with ICM or in high-risk persons for novel gene treatments and preventions of the illness [33].

Statistical Analysis The significant differences between two groups were examined using t-test. Corrected P-values were computed using Benjamini and Hochberg approach to control error rates. Statistical significances were defined as an adjusted P-value < 0.05 [5].

Results

The study initiated by pre-processing and quality-checking the GDS3115 dataset, ensuring the reliability and consistency of the data. Subsequently, differential expression analysis is done to identify genes that exhibit significant changes in expression levels under specific experimental conditions like t-test, Benjamini and Hochberg. The results of each phase are represented in the form of visualization as follows.

Gene Expression Profiling (Data Pre-processing)
--

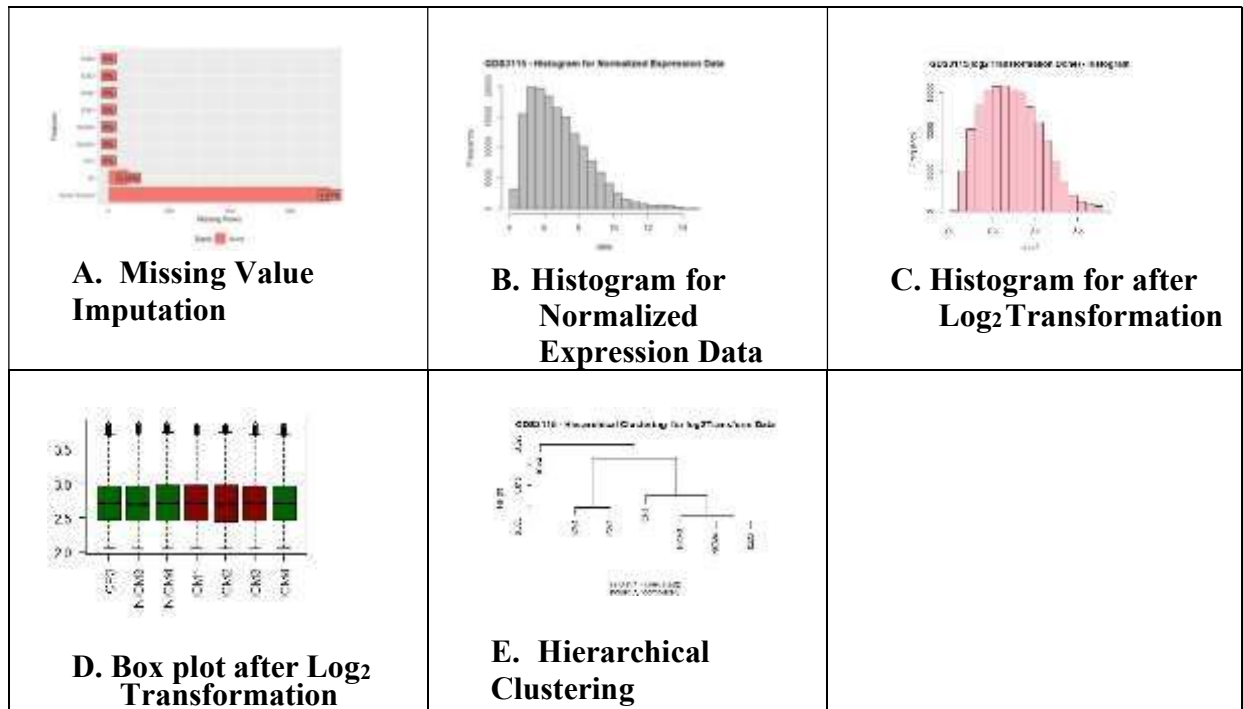
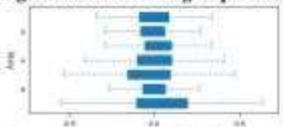
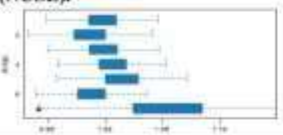
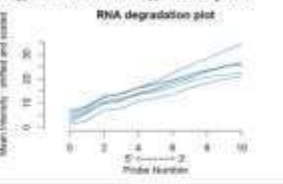
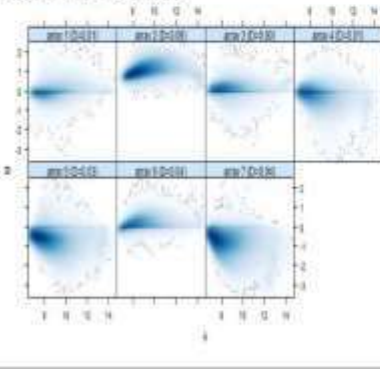
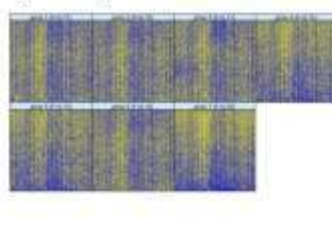


Figure 3: A. Missing Value Imputation B. Histogram for Normalized Expression Data C. Histogram for after Log₂ Transformation D. Box plot after Log₂ Transformation E. Hierarchical Clustering for log₂ Transformation Data F. Quality Assessment Report before and after Normalization using RMA Algorithm.

<p>Quality Assessment before Normalization (Raw Data): This report has been created with arrayQualityMetrics 3.52.0 under R version 4.2.2</p>	
<p>Section 1: Between array comparison</p>	
<p>Figure 1: Distances between arrays.</p>	<p>Figure 1: shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance d_{ab} between two arrays a and b is computed as the mean absolute difference (L_1-distance) between the data of the arrays (using the data from all probes without filtering). In formula, $d_{ab} = \text{mean} M_{ai} - M_{bi}$, where M_{ai} is the value of the i-th probe on the a-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \sum_i d_{ai}$ was exceptionally large. One such array was detected, and it is marked by an asterisk,*.</p>
<p>Figure 2: Principal Component Analysis.</p>	<p>Figure 2 shows a scatterplot of the arrays along the first two principal components. This plot is used to explore if the arrays cluster, and whether this is according to an intended experimental factor (such a factor is indicated by color using the 'intgroup' argument). Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.</p>
<p>Section 2: Array intensity distributions</p>	
<p>Figure 3: Boxplots.</p>	<p>Figure 3 shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic K_i between each array's distribution and the distribution of the pooled data.</p>
<p>Figure 4: Density plots.</p>	<p>Figure 4 shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right tail. A bulge at the upper end of the intensity range often indicates signal saturation.</p>
<p>Section 3: Variance mean dependence</p>	
<p>Figure 5: Standard deviation versus rank of the mean.</p>	<p>Figure 5 shows a density plot of the standard deviation of the intensities across arrays on the y-axis versus the rank of their mean on the x-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm (\log-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic saturation of the intensities.</p>

Section 4: Affymetrix specific plots	
<p>Figure 6: Relative Log Expression (RLE).</p> 	<p>Figure 6 shows the <i>Relative Log Expression (RLE)</i> plot. Arrays whose boxes are centered away from 0 and/or are more spread out are potentially problematic. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic R_n between each array's RLE values and the pooled, overall distribution of RLE values.</p>
<p>Figure 7: Normalized Unscaled Standard Error (NUSE).</p> 	<p>Figure 7 shows the <i>Normalized Unscaled Standard Error (NUSE)</i> plot. For each array, the boxes should be centered around 1. An array where the values are elevated relative to the other arrays is typically of lower quality. Outlier detection was performed by computing the 75% quantile N_n of each array's NUSE values and looking for arrays with large N_n.</p>
<p>Figure 8: RNA digestion plot.</p> 	<p>Figure 8 shows the <i>RNA digestion</i> plot. The shown values are computed from the pre-processed data (after background correction and quantile normalisation). Each array is represented by a single line; move the mouse over the lines to see their corresponding sample names. The plot can be used to identify array(s) that have a slope very different from the others. This could indicate that the RNA used for that array has been handled differently from what was done for the other arrays.</p>
<p>Figure 9: Perfect matches and mismatches.</p>	<p>Figure 9 shows the density distributions of the log₂ intensities grouped by the matching type of the probes. The blue line shows a density estimate (smoothed histogram) from intensities of perfect match probes (PM), the grey line, one from the mismatch probes (MM). We expect that MM probes have poorer hybridization than PM probes, and thus that the PM curve be to the right of the MM curve.</p>
Section 5: Individual array quality	
<p>Figure 10: MA plots.</p> 	<p>Figure 10 shows MA plots. M and A are defined as: $M = \log_2(I_1) - \log_2(I_2)$ $A = 1/2 (\log_2(I_1) + \log_2(I_2))$, where I_1 is the intensity of the array studied, and I_2 is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the $M = 0$ axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation). Outlier detection was performed by computing Hoeffding's statistic D_n on the joint distribution of A and M for each array. The value of D_n is shown in the panel headings. 0 arrays had $D_n > 0.15$ and were marked as outliers.</p>
<p>Figure 11: Spatial distribution of M.</p> 	<p>Figure 11 shows false color representations of the arrays' spatial distributions of feature intensities (M). Normally, when the features are distributed randomly on the arrays, one expects to see a uniform distribution; control features with particularly high or low intensities may stand out. The color scale is proportional to the ranks of the probe intensities. Note that the rank scale has the potential to amplify patterns that are small in amplitude but systematic within an array. It is possible to switch off the rank scaling by modifying the argument scale in the call of the <i>aqm</i> spatial function. Outlier detection was performed by computing F_n, the sum of the absolute value of low frequency Fourier coefficients, as a measure of large scale spatial structures. The value of F_n is shown in the panel headings.</p>

<p>Quality Assessment After RMA Normalization (Normalized Data): This report has been created with arrayQualityMetrics 3.52.0 under R version 4.2.2</p>	
<p>Section 1: Between array comparison</p>	
<p>Figure 1: Distances between arrays.</p>	<p>Figure 1 shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance d_{ab} between two arrays a and b is computed as the mean absolute difference (L₁-distance) between the data of the arrays (using the data from all probes without filtering). In formula, $d_{ab} = \text{mean} M_{ai} - M_{bi}$, where M_{ai} is the value of the i-th probe on the a-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \sum_b d_{ab}$ was exceptionally large. No such arrays were detected.</p>
<p>Figure 2: Principal Component Analysis.</p>	<p>Figure 3 shows a scatterplot of the arrays along the first two principal components. This plot is used to explore if the arrays cluster, and whether this is according to an intended experimental factor (such a factor is indicated by color using the 'intgroup' argument). Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.</p>
<p>Section 2: Array intensity distributions</p>	
<p>Figure 3: Boxplots.</p>	<p>Figure 3 shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic K_i between each array's distribution and the distribution of the pooled data.</p>
<p>Figure 4: Density plots.</p>	<p>Figure 4 shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right tail. A bulge at the upper end of the intensity range often indicates signal saturation.</p>
<p>Section 3: Variance mean dependence</p>	
<p>Figure 5: Standard deviation versus rank of the mean.</p>	<p>Figure 5 shows a density plot of the standard deviation of the intensities across arrays on the y-axis versus the rank of their mean on the x-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.</p>

F. Quality Assessment Report

Gene Filtering

Filtering Low Variance Probe Sets in Microarray Data Analysis

Microarray data analysis often involves preprocessing steps to filter out probe sets with low variability across arrays. In this study, filtering step is performed to remove probe sets exhibiting low variance, specifically those falling below the 0.25 quantile threshold across the entire dataset. Initially, the dataset is comprised a total of 22,283 probe sets. To ensure the reliability of subsequent analyses and to focus on probe sets exhibiting substantial variability, a filtering criterion based on variance is implemented. Probe sets with variance values below the 0.25 quantile threshold were deemed to have low variability and consequently removed from the dataset.

Upon implementing the filtering criterion, the dataset underwent a reduction in the number of probe sets. The total number of rows post-filtering amounted to 16,712, indicating the removal of 5,571 probe sets with lower variance levels.

Quantification of Filtering:

The reduction from 22,283 to 16,712 probe sets equates to a filtering percentage of approximately 75%. This quantifies the impact of the filtering process, illustrating that approximately three-fourths of the probe sets were eliminated due to their lower variance levels falling below the 0.25 quantile threshold.

Significance of Filtering:

Filtering out probe sets with low variance is a crucial preprocessing step in microarray data analysis. This step helps in streamlining the dataset by focusing on probe sets with higher variability, which are often more informative and likely to represent genes or genomic regions displaying meaningful differences across experimental conditions or samples.

Filtering Probe Sets without Gene Annotation Information in Microarray Analysis

Accurate gene annotation is crucial in microarray data analysis to ensure the relevance and interpretability of results. In this study, a filtering step is performed to eliminate probe sets lacking gene annotation information from the dataset. Initially, the dataset consisted of 22,283 probe sets derived from microarray data. Recognizing the importance of gene annotation for meaningful analysis, filtering process is initiated to exclude probe sets that lacked associated gene annotation information.

Quantification of Filtering:

The reduction from 22,283 to 15,992 probe sets signifies a filtering percentage of approximately 72%. This quantifies the impact of the filtering process, illustrating that nearly three-fourths of the probe sets lacking gene annotation information were excluded from the dataset.

Significance of Filtering:

Filtering out probe sets without gene annotation information is essential to ensure that subsequent analyses focus on annotated genes, facilitating meaningful interpretation of gene expression data. Genes with proper annotations provide valuable insights into biological functions, pathways, and associated molecular mechanisms.

Filtering Probe Sets with Multiple Gene Symbols in Microarray Data Analysis

Accurate gene annotation and assignment of probe sets to specific genes are crucial for meaningful interpretation in microarray data analysis. In this study, filtering process is initiated to exclude probe sets associated with multiple gene symbols, aiming to ensure clarity and specificity in gene identification. Initially, dataset contained 22,283 probe sets derived from microarray data. Recognizing the need for precise gene attribution, a filtering criterion is implemented to remove probesets linked to multiple gene symbols, as these cases might introduce ambiguity in gene identification.

Quantification of Filtering:

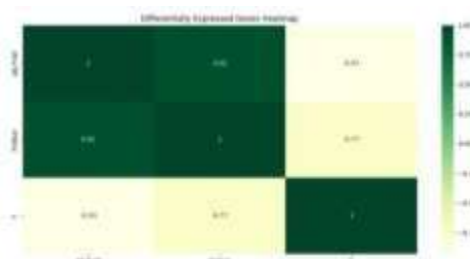
The reduction from 22,283 to 9,982 probe sets represents a filtering percentage of approximately 45%. This quantifies the impact of the filtering process, demonstrating that nearly half of the probe sets linked to multiple gene symbols was eliminated from the dataset.

Significance of Filtering:

Filtering out probe sets associated with multiple gene symbols enhances the specificity and accuracy of gene attribution in microarray data analysis. The removal of ambiguous or non-specific gene assignments ensures that downstream analyses focus on probe sets uniquely associated with individual genes, thereby improving the reliability of biological interpretations.

DEG(Feature Extraction)

Heat maps are used to identify genes which are regulated and associated with particular condition[14]. A heat map displaying the differential expression of 250 genes from the top table has been generated, employing various statistical measures such as adjusted P-values, P-values, and F-values. The values ranged from -0.75 to 1.00, with negative values indicating down regulated genes, positive values signifying up regulated genes, and the magnitude representing the extent of expression changes. The heat map of 250 DEG's



from 22283 rows of data are depicted in Figure 4, which illustrates distinct gene expression profiles between ICM samples and normal controls [13].

Heat map for 250 DEG's from 22283 rows of data

Figure 4: Heat map illustrating DEG. The color gradient, ranging from yellow to green, indicates the gene expression values relative to the ischemic cardiomyopathy group

compared to control groups, representing down to up regulations, respectively.

Distribution of Statistical Significance ($-\log_{10}(\text{P-Value})$) Among Top Genes

The analyses of differential gene expression often involves the identification of statistically significant genes. In this investigation, a top table comprising 250 genes extracted from a dataset containing 22,185 rows of data were analysed. The visual representation of the distribution of statistical significance using a histogram plot based on the negative logarithm of the p-values ($-\log_{10}(\text{P-Value})$) derived from the top table is

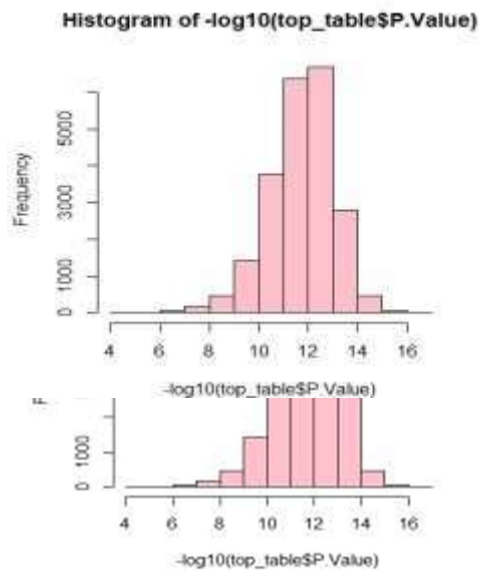


Figure 5: Histogram Plot using top table P values

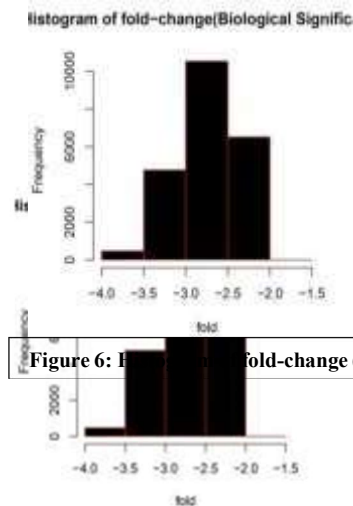


Figure 6: Histogram of fold-change (biological)

Figure 6: Histogram of fold-change (biological)

The histogram depicts a distribution of statistical significance among the top 250 genes. The x-axis range from 4 to 16 signifies varying degrees of statistical significance, with higher values indicating increased significance. The y-axis illustrates the frequency of genes falling within these significance levels. The peak in the histogram between the $-\log_{10}(\text{P-Value})$ range of 12 to 13 indicates a concentration of genes with extremely high statistical significance in the dataset. When the value exceeds 13 and reaches towards 16, it indicates that these genes have extremely low p-values.

This visualization aids in understanding the range and frequency distribution of statistical significance levels, contributing to the identification and prioritization of genes with higher levels of statistical confidence in the context of differential expression analysis.

1. Likelihood of True Association: The peak in this range suggests that a substantial number of genes within the top 250 genes exhibit exceptionally strong evidence of differential expression or association with the studied conditions or factors. These genes are likely to be highly relevant or crucial in the context of the

biological phenomena being investigated.

2. **Potential Biological Significance:** Genes with such high statistical significance (represented by $-\log_{10}(\text{P-Value})$ in the range of 12 to 13) might indicate important regulatory mechanisms, key molecular players, or biomarkers associated with the experimental conditions. These genes could potentially serve as targets for further functional validation or exploration in subsequent studies due to their robust statistical support.

In summary, the peak in the histogram within the 12 to 13 $-\log_{10}(\text{P-Value})$ range signifies a notable concentration of genes displaying exceptionally strong statistical significance, suggesting their potential importance and relevance in the biological context under investigation.

Exploring Biological Significance through Fold-Change distribution among top genes

Biological significance in gene expression analysis is often assessed through fold-change values, indicating the magnitude of differential expression between experimental conditions. In this analysis, the dataset comprising 22,185 rows were utilized and extracted a subset of the top 250 genes based on differential expression. Histogram graphic was used to show distributions of fold-change values amongst these top genes. The distributions of fold-change values amongst top 250 genes were shown in histogram graphic. The x-axis ranges from -4.0 to -1.5, representing different fold-change intervals, while the y-axis illustrates the frequency distribution, spanning increments of 2000 from 0 to 10000.

This describes the significance of the histogram plot in visualizing distributions of fold-changes amongst top genes and emphasized the purpose of using fold-change as a measure to assess the magnitude of differential gene expression between experimental conditions. A higher absolute fold-change value indicates a greater magnitude of differential expression between experimental conditions. In this case, the peak in the range of -3.0 to -2.5 signifies that a substantial number of genes among the top 250 exhibit moderate to moderately high differences in expression levels. These genes are likely to have notable biological significance and may serve as potential candidates for further investigation in understanding their roles in the studied conditions or biological processes.

Below describes the histogram plot generated from the fold-change values of the top 250 genes derived from a dataset of 22,185 rows.

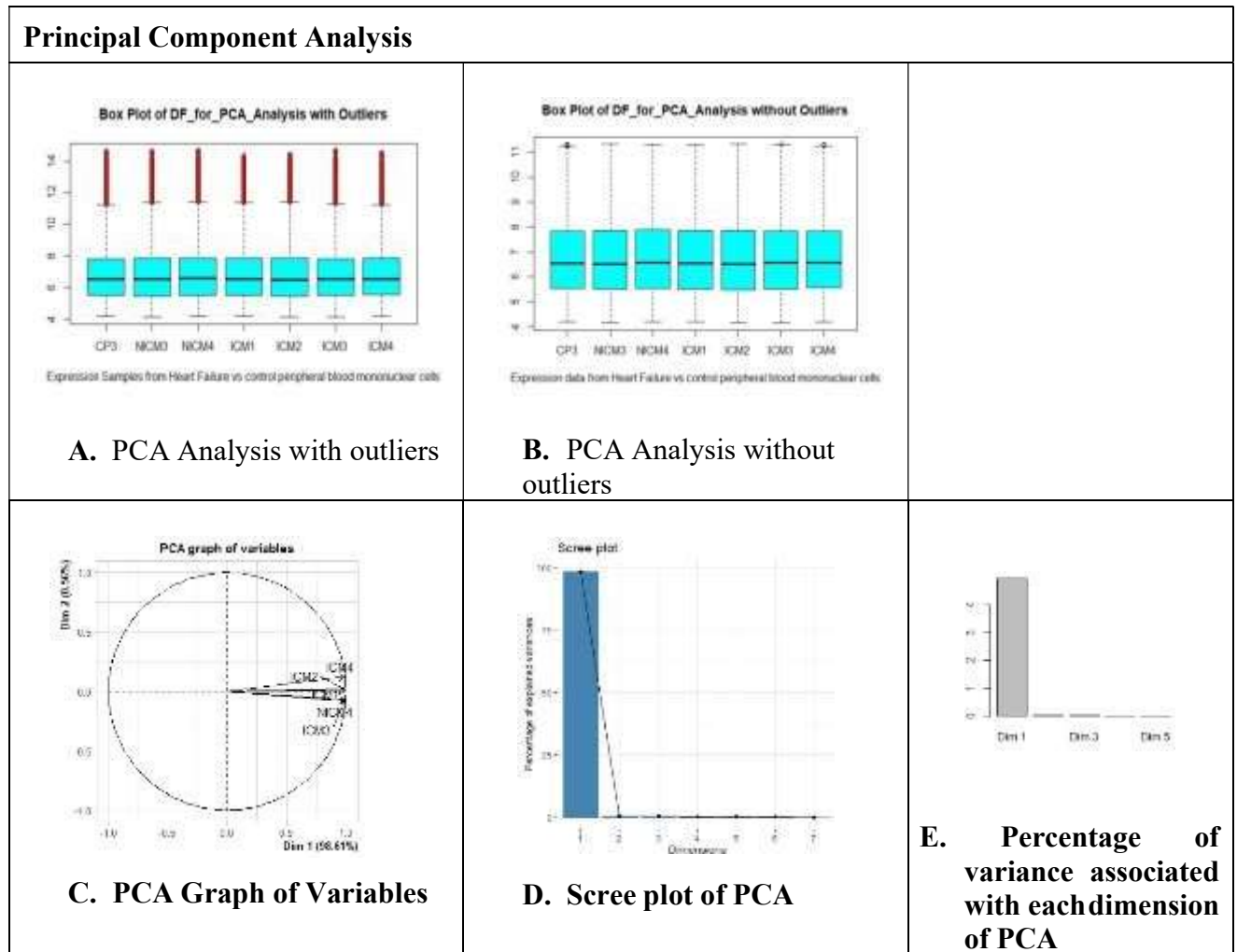


Figure 7: A. PCA Analysis with outliers B. PCA Analysis without outliers C. PCA Graph of Variables D.

Scree plot of PCA **E. Percentage of variance associated with each dimension of PCA**

Scatter Plot: Positive correlation exists: When the “y” variable tends to increase as the “x” variable increases, we say there is a **positive correlation** between the variables. Few outliers are there.

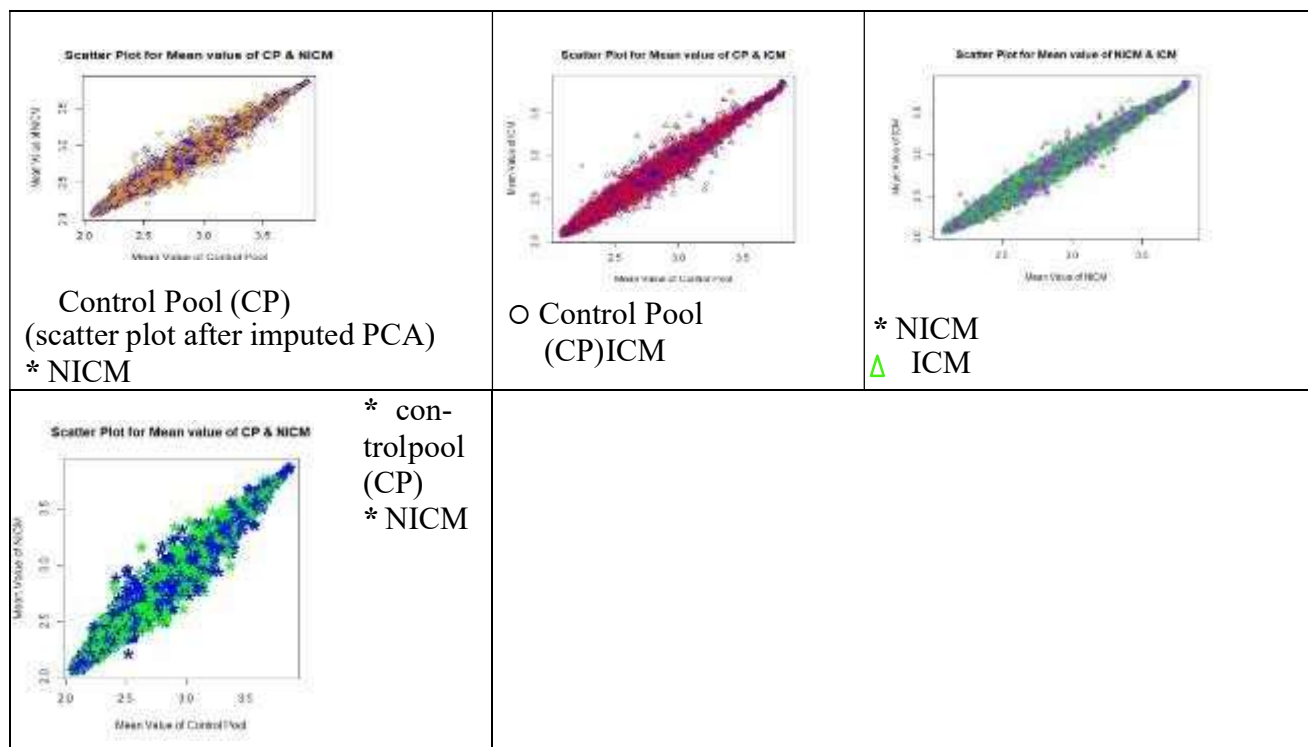


Figure 8: Scatter Plot between Control pool, NICM and ICM.

Results Identification of DEG's in ICM/NICM samples compared with normal controls: Between the ICM, NICM, and control groups, a total of 5 genes with DEG were filtered out; 2 of these genes had up-regulated expression and 3 had down-regulated expression [14]. Table 1 displays the DEG that were found to have significantly different gene expression patterns. Figure 9 plots these DEG using a volcano plot to highlight the different gene expression profiles between the ICM sample and the normal controls.

S. No	ID	Gene Symbol	Gene Title	Log2 FC Value	-log10 (P Value)	Up/Down Regulated Gene
1	205898_at	CX3CR1C	C-X3-C motif chemokine receptor 1	2.164	5.303	Up
2	200800_s_at	HSPA1L//HSPA1B//HSPA1A	Heat Shock Protein Family A (HSP 70) Member 1 like// Heat Shock Protein Family A (HSP 70) Member 1B// Heat Shock Protein Family A (HSP 70) Member 1A	1.745	5.052	Up
3	201466_s_at	JUN	Jun proto-oncogene, AP-1 transcription factor subunit	-1.158	5.425	Down
4	219228_a	ZNF331	Zinc finger protein 331	-1.667	5.192	Down

	t					
5	Z10426_xat	RORA	KAR related orphan receptor A	-1.819	5.582	Down

Table 1: Identified DEG’s showing significant differences in their gene expression profiles

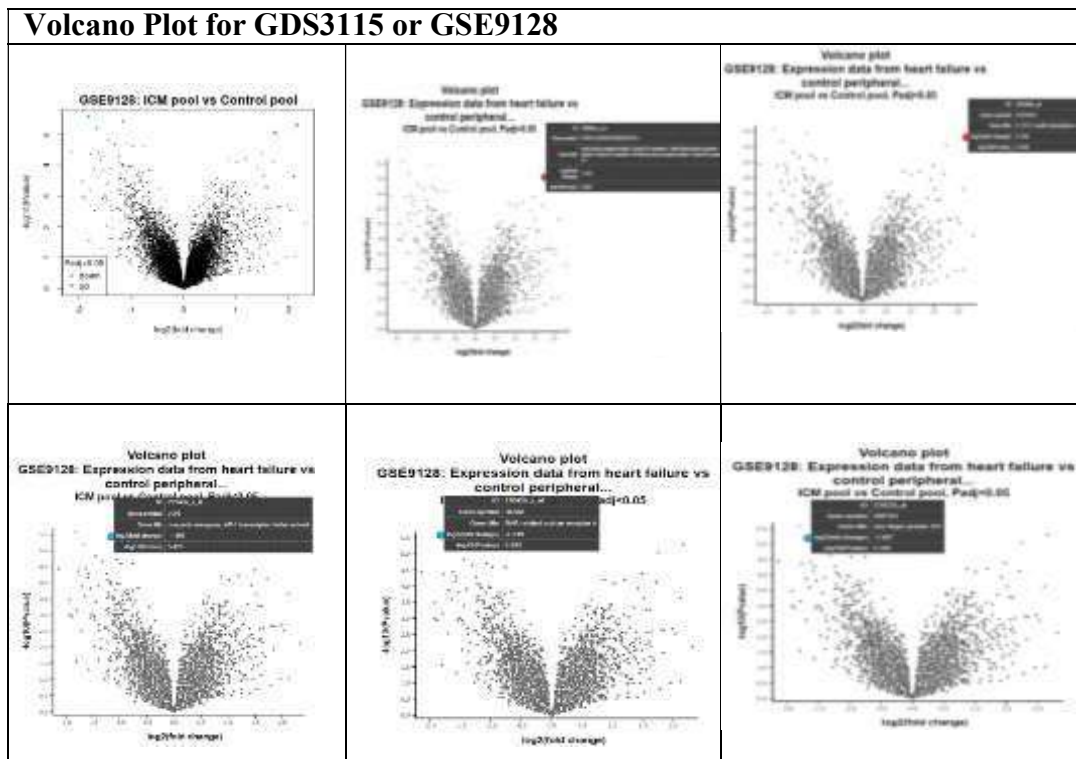


Figure 9: Volcano plot illustrating the up and down-regulated genes between ICM and control pool

Performance measures: Precision, Recall, F-measure, Accuracy and Error have been used to evaluate the classifiers in this study. The outline of the confusion matrix is revealed in Table 2.

Actual Class	Prediction Class	
	P	N
P	TP	FN

N	FP	TN
---	----	----

**TABLE 2 CONFUSION MATR
IX IN THIS STUDY.**

The performance of classifiers was assessed using the following metrics: accuracy, f-measure, recall, specificity, sensitivity, and precision. With the direction of the following equations (1–5),

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

$$\text{Recall}(R)/ \text{Sensitivity}(\text{Sen}) = \frac{TP}{TP + FN} \tag{3}$$

$$F - \text{measure} = 2 \cdot \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Accuracy}(\text{Acc}) = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Error} = \frac{FP + FN}{TP + TN + FP + FN} \tag{6}$$

True Positive (TP) signifies that the positive samples' diagnoses were accurate. False Negative (FN) denotes an inaccurate diagnosis of the positive samples. False Positive (FP) denotes erroneous diagnosis of the non-positive samples. True Negative (TN) signifies that the non-positive samples diagnoses were accurate.

DATASE T	Classifiers	Precisi on (%)	Reca ll (%)	F - Measu re (%)	Accura cy (%)	Erro r (%)
GDS3115	Random Forest	87.2	87.2	93.2	87.2	14.83
	Naïve Bayes	83.3	85.6	84.4	85.6	15.90
	GNN+SVM(Existing)	82.66	80.04	81.35	84.47	15.53

TABLE 3 COMPARATIVE PERFORMANCES OF THE PROPOSED CLASSIFIERS WITH EXISTING METHODS

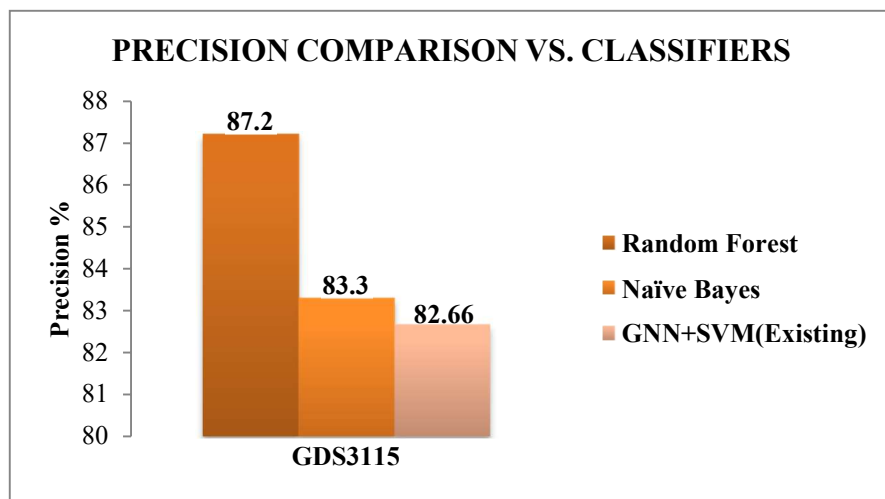


FIGURE 10 PRECISION COMPARISON VS. CLASSIFIERS

Figure 10 shows the precision comparison of classifiers like GNN+SVM, Naive Bayes, and Random Forest with respect to gene dataset (GDS3115). Random Forest classifier has produces highest precision results of 87.2 and GNN+SVM have lowest precision of 82.66% for GDS3115 dataset (Refer Table 3).

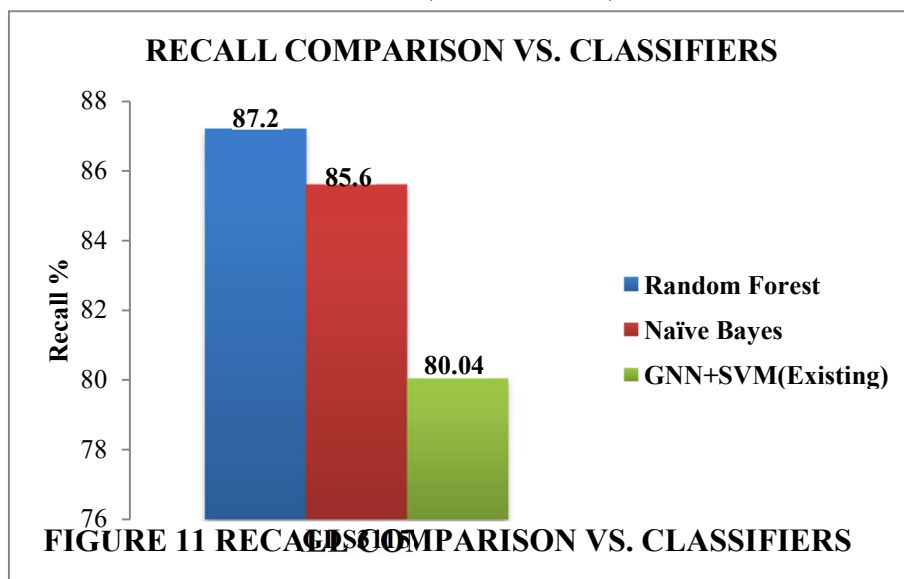


FIGURE 11 RECALL COMPARISON VS. CLASSIFIERS

Figure 11 shows the recall comparison of classifiers like GNN+SVM, Naive

Bayes, and Random Forest with respect to gene dataset (GDS3115). Random Forest classifier has produces highest recall results of and GNN+SVM has lowest recall of 80.04% for GDS3115 dataset (Refer Table 3).

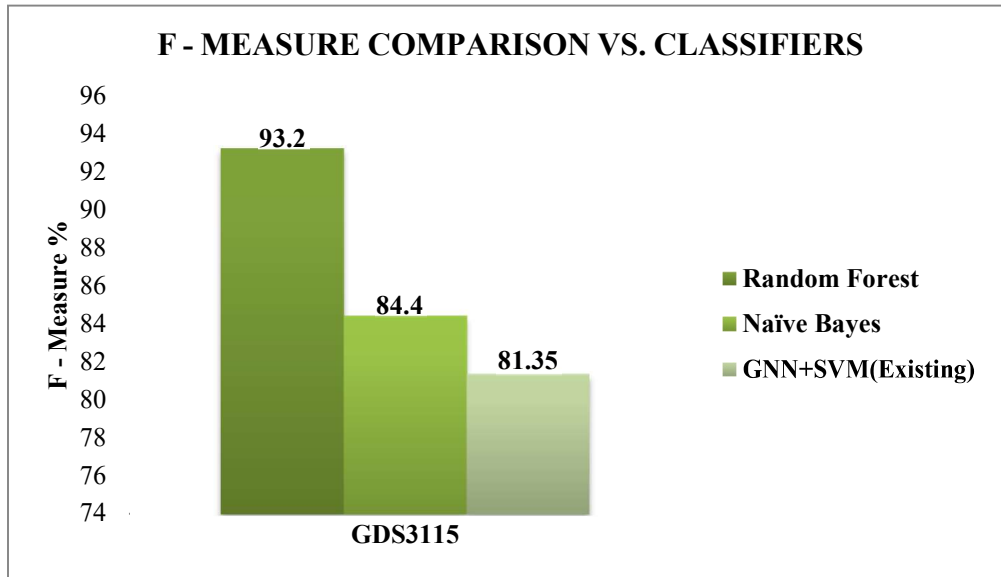


FIGURE 12 F – MEASURE COMPARISON VS. CLASSIFIERS

Figure 12 shows the F – Measure comparison of classifiers like GNN+SVM, Naive Bayes, and Random Forest with respect to gene dataset (GDS3115). Random Forest classifier has produces highest F – Measure results of 93.2 and GNN+SVM have lowest precision of 81.35% for GDS3115 dataset (Refer Table 3).

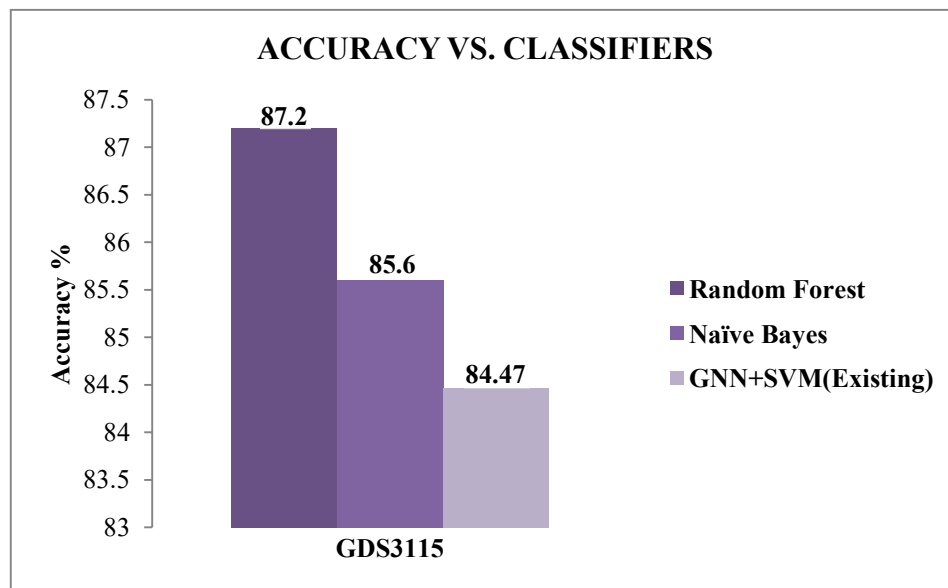


FIGURE 13 ACCURACY COMPARISON VS. CLASSIFIERS

Figure 13 shows the accuracy comparison of classifiers like GNN+SVM, Naive Bayes, and Random Forest with respect to gene dataset (GDS3115). Random Forest classifier has produces highest accuracy results of 87.2 and GNN+SVM have lowest accuracy of 84.47% for GDS3115 dataset (Refer Table 3).

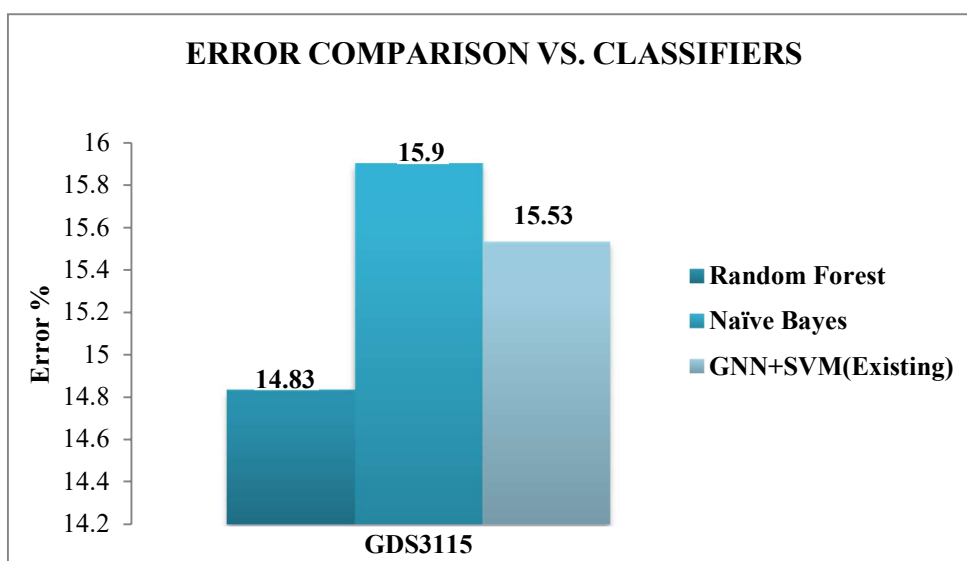


FIGURE 14 ERROR COMPARISON VS. CLASSIFIERS

Figure 13 shows the error comparison of classifiers like GNN+SVM, Naive Bayes, and Random Forestwith respect to gene dataset (GDS3115). Naive Bayes classifier has produces highest error results of 15.9% and Random Forest has lowest error of 14.83% for GDS3115 dataset (Refer Table 3).

Conclusion

"In conclusion, this proof-of-concept-based investigation successfully identified DEG that differentiate between patient groups and controls in the GDS3115 dataset, employing a rigorous statistical framework and utilizing R programming and Python for differential gene expression analysis. The meticulous data-driven approach exemplified in this research emphasized systematic analysis methodologies, highlighting the significance of identifying several DEGpotentially involved in the pathogenesis of ICM or NICM. The exploration of transcriptomicdata sheds light on these genes, suggesting their potential roles in clinical therapeutic strategies.

These findings not only provide biological interpretation and functional context but also lay the groundwork for future investigations. This groundwork holds promise for

discovering novel therapeutic targets and putative biomarkers in various biological contexts related to cardiovascular disease research. Ultimately, this study's outcomes pave the way for the development of innovative diagnostic or therapeutic approaches based on the identified genes associated with ICM/NICM pathogenesis."

References

1. Albert R, Jeong H, Barabasi A. Error and attack tolerance of complex networks. *Nature*. 2000; 406: 378–382, doi: 10.1038/35019019, indexed in Pubmed: 10935628.
2. Beltrami CA, Finato N, Rocco M, et al. Structural basis of end stage failure in ischemic cardiomyopathy in humans. *Circulation*. 1994; 89(1): 151–163, doi: 10.1161/01.cir.89.1.151, indexed in Pubmed: 8281642.
3. Candell-Riera J, Romero-Farina G, Aguadé-Bruix S, et al. Ischemic cardiomyopathy: a clinical nuclear cardiology perspective. *Rev Esp Cardiol*. 2009; 62(8): 903–917, doi:10.1016/s1885 5857(09)72655-6, indexed in Pubmed: 19706246.
4. Jeong H, Mason S, Barabasi A, et al. Lethality and centrality in protein networks. *Nature*. 2000; 411: 1.
5. Kittleson MM, Ye SQ, Irizarry RA, et al. Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy. *Circulation*. 2004; 110(22): 3444–3451, doi: 10.1161/01.CIR.0000148178.19465.11, indexed in Pubmed: 15557369.
6. Li-Jun Wang1,, Bai-Quan Qiu1,, Ming-Ming Yuan2,* , Hua-Xi Zou1, Cheng-Wu Gong1, Huang Huang2, Song-Qing Lai3, Ji-Chun Liu1, "Identification and Validation of Dilated Cardiomyopathy-Related Genes via Bioinformatics Analysis," *International Journal of General Medicine*, vol. 15, pp. 3663-3676, 2022.
7. Liu Wm, Mei R, Di X, et al. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*. 2002; 18(12): 1593–1599, indexed in Pubmed: 12490443.
8. Press W, Teukolsky S, Vetterling W, et al. Hierarchical clustering by phylogenetic trees. *Section*, 2007: 164.
9. Ritchie ME, Phipson B, Wu Di, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015; 43(7): e47, doi: 10.1093/nar/ gkv007, indexed in Pubmed: 25605792.
10. Rodkey S, Ratliff N, Young J. Cardiomyopathy and myocardial failure.

Comprehensive Cardiovascular Med. 1998: 2.

11. Sangeetha, T., & Manikandan, K. (2023). An extensive study on cardiomyopathy classification technique using microarray data. *Journal of Data Acquisition and Processing*, 38(1), 3835. DOI: 10.5281/zenodo.7765934
12. Suma H, Anyanwu AC. Current status of surgical ventricular restoration for ischemic cardiomyopathy. *Semin Thorac Cardiovasc Surg.* 2012; 24(4): 294–301, doi: 10.1053/j.semtcvs.2013.01.002, indexed in Pubmed: 23465678.
13. Szekely G, Rizzo M. Hierarchical Clustering via Joint Between Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification.* 2005; 22(2):151–183, doi: 10.1007/ s00357-005-0012-9.
14. Wang L, Cao C, Ma Q, et al. RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications. *BMC PlantBiol.* 2014; 14: 169, doi: 10.1186/1471-2229-14-169, indexed in Pubmed: 24939556.

Yang Y, Yang W, Huo W,
Huo P, and H. Yang, "Identification of biomarkers for ischemic cardiomyopathy based on microarray data analysis," *Cardiology Journal*, vol. 24, no. 3, pp. 305-313, 2017. DOI: 10.5603/CJ.a2017.0005