Advancing Gene Signature Discovery with Generative Models: A Case Study in Sepsis

by
Anjali Sharma

Under the supervision of
Dr. Tavpritesh Sethi

A thesis submitted in complete fulfillment of the requirements for the degree of Master of Technology, CSE

Centre for Computational Biology, Indraprastha Institute of Information Technology, Delhi

May 2023

Certificate

This is to certify that the thesis titled “Advancing Gene Signature Discovery with Generative Models: A Case Study in Sepsis," submitted by Anjali Sharma to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. The thesis has met the standards, fulfilling the requirements of the regulations relating to the degree.

June 2022

Dr Tavpritesh Sethi
Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi

Declaration

I, Anjali Sharma, declare that this report titled "Advancing Gene Signature Discovery with Generative Models: A Case Study in Sepsis" is submitted in complete fulfillment of the requirements for the Master's degree's CSE from IIIT Delhi. This work was carried out at the Department of Computational Biology, Indraprastha Institute of Information Technology, under Dr. Tavpritesh Sethi, Associate Professor, IIIT-D, from September 15, 2023, to May 10, 2023. The thesis contains **n** pages and has yet to be submitted elsewhere. This work is an authentic record of my research.

Anjali Sharma
Student's Signature

Acknowledgements

I sincerely thank Prof. Tavpritesh Sethi for his unwavering guidance and mentorship throughout this research work. Without his invaluable support, this work would not have been possible. Since the beginning of my thesis, Prof. Sethi has been instrumental in pushing me to strive for higher goals and consistently improve my work. He has been very approachable, providing me with guidance and clarification whenever I had any doubts or issues, even during his scheduled days. His insights and suggestions have helped me to stay focused and complete my work on time. I also extend my heartfelt thanks to my father, mother, and brother for their indirect but indispensable support throughout my research work. Their encouragement and support have been invaluable throughout my academic journey, from my undergraduate studies to graduation.

I am also grateful to my PhD mentor, Alok Anand, who was always available to help me with even the smallest doubts about my research work or career advice. His innovative suggestions and ideas have been pivotal in shaping my thesis work. I would also like to thank my friends, Shubham Lal and Sumit Kumar, for being there for me during the challenges that came with our web server work. My fellow batchmates have also been a great source of help, clarifying my doubts about biological aspects and offering continuous support and encouragement in my academic and personal lives. I extend my sincere appreciation to all of them.

I am grateful to the Department of Computational Biology and IIIT Delhi faculty members and staff for their support throughout my thesis. I would like to extend a special thanks to Mr. Adarsh from the IT department for his continuous assistance in providing access to the college IT infrastructure and helping us with the web server deployment work.

List of Abbreviations

S.No.	Abbreviations	Full forms
1	ML	Machine learning
2	GMM	Gaussian Mixture Model
3	BN	Bayesian Network
4	CTGAN	Conditional Tabular Generative Adversarial Network
5	DGEA	Differential Gene Expression Analysis
6	DEG	Differentially Expressed Genes
7	GEO	Gene Expression Omnibus
8	RMA	Robust Multi-array Average
9	LIMMA	Linear Models for Microarray Data
10	KEGG	Kyoto Encyclopedia of Genes and Genomes
11	MSigDB	Molecular Signature Database
12	GSEA	Gene Set Enrichment Analysis
13	GO	Gene Ontology
14	ES	Enrichment score
15	logFc	Log Fold Change
16	BH	Benjamini-Hochberg
17	FDR	False Discovery Rate

Contents

1. INTRODUCTION 1. A GMM-based Data Augmentation 1. B BN-based Data Augmentation 1. C CTGAN-based Data Augmentation 1.2 Literature Review 1.3 Aims and Objectives 1.4 Thesis Outline 1.5 Methodology 1.5.1 Acquisition of Gene Expression Profiles from the GSE65682 Dataset in the GEO Database 1.5.2 Data Pre-processing and Normalization 1. 5.3 Data Preparation for Analysis of Differential Gene Expression 1. 5.4 Analysis of differential gene expression 1. 5.5 Augmentative expression data using the XpressionSuite tool 1.5.6 Gene Set Enrichment Analysis	12
2. XPRESSIONSUITE ABSTRACT INTRODUCTION 2.1 Implementation of the Web Server 2.1.1 Application Architecture 2.1.2 System Design 2.1.1 Deployment 2.2 XpressionSuite Web Server Features 2.2.1 XpressionSuite HomePage 2.2.2 XpressionSuite Login Page 2.2.3 XpressionSuite SignUp Page 2.2.4 XpressionSuite Profile Page	31

2.2.5 Data Augmentation with Synthetic Data 2.2.1 XpressionSuite Analytics Page 2.3 Results Of XPRESSIONSUITE	
3. RESULTS AND DISCUSSION 3.1. Identification of DEGs in the expression dataset 3.2. Original data file 3.3. Augmentation of the expression data 3.4. Identification of hallmarks in original vs augmented data file. 3.5. Enrichment analysis	48
4. CONCLUSION AND FUTURE SCOPE 4.1 CONCLUSION 4,2 FUTURE SCOPE	71
5. BIBLIOGRAPHY	74

List of Figures

Fig 1.3	Fig. 1.3 Flow Diagram of the Aims and Objectives of my Thesis
Fig 1.5	Flow diagram of the methodology followed
Fig. 1.5.1A	Workflow for acquisition of Gene Expression Profiles from the GSE65682 Dataset in the GEO Database
Fig 1.5.1B	MARS Consortium Dataset
Fig 1.5.2A	Boxplot for Quality checks on the pre-processed Expression Data Matrix In this box plot, the x-axis shows the sample IDs, and the y- axis represents the expression values for the expression data (A) Before pre-processing, and (B). After pre-processing
Fig 1.5.2B	(A) Volcano plot before pre-processing, showing gene expression changes (x-axis: density, y-axis: significance) with a bandwidth of 0.2618; and (B) Volcano plot after pre-processing, depicting refined gene expression changes (x-axis: density, y-axis: significance) with a bandwidth of 0.06392.
Fig. 1.5.3	Expression Data Matrix
Fig 2	Application Design
Fig 2.1	Homepage for the XpressionSuite Website
Fig 2.2	Login Page for the XpressionSuite Website
Fig 2.3	SignUp Page for the XpressionSuite Website
Fig 2.4	Profile Page for the XpressionSuite Website
Fig 2.5	Data Augmentation Page for the XpressionSuite Website
Fig 2.6	Analytics Page for the XpressionSuite Website
Fig 3.1	Identification of DEGs in the expression dataset
Fig 3.2	Volcano plot for expression data set done on original expression data file
Fig 3.3.A	Volcano plot for augmented expression data using GMM

Fig 3.3.B	Volcano plot for augmented expression data using BN
Fig 3.3.C	Volcano plot for augmented expression data using CTGAN
Fig 3.4.1	MSigDB hallmarks for the original expression data
Fig 3.4.2	MSigDB hallmarks for the augmented file using the A. BN model and B. CTGAN model
Fig 3.5.I(A)	GSEA plot for the GO enrichment analysis for Original Data
Fig 3.5.II(A)	GSEA plot for the GO enrichment analysis for GMM Data
Fig 3.5.III(A)	GSEA plot for the GO enrichment analysis for BN Data
Fig 3.5.IV(A)	GSEA plot for the GO enrichment analysis for CTGAN Data
Fig 3.5.I(B)	Dot plot for the enrichment analysis of Original Data
Fig. 3.5.II(B)	Dot plot for the enrichment analysis of GMM Data
Fig 3.5.III(B)	Dot plot for the enrichment analysis of BN Data
Fig 3.5.IV(B)	Dot plot for the enrichment analysis of CTGAN Data
Fig 3.5.I(C)	Ridge plot for the enrichment analysis of Original Data
Fig 3.5.II(C)	Ridge plot for the enrichment analysis of GMM Data
Fig 3.5.III(C)	Ridge plot for the enrichment analysis of BN Data
Fig 3.5.IV(C)	Ridge plot for the enrichment analysis of CTGAN Data
Fig 3.5.I(D)	Category net plot for the enrichment analysis of Original Data
Fig 3.5.I(D)	Category net plot for the enrichment analysis of GMM Data
Fig 3.5.III(D)	Category net plot for the enrichment analysis of BN Data
Fig 3.5.IV(D)	Category net plot for the enrichment analysis of CTGAN Data

LIST OF TABLES

	Page No.
Fig. 1.5.1 MARS Consortium Dataset	22

Abstract

Sepsis and diabetes present intricate medical conditions that present substantial challenges to healthcare systems globally. Timely detection and precise diagnosis are critical in facilitating effective treatment and enhancing patient outcomes. This study aimed to investigate the potential of machine learning-based data augmentation approaches for biomarker discovery in a multicenter dataset of patients with sepsis and diabetes. Specifically, the study focused on comparing the efficacy of three different approaches, including the Gaussian Mixture Model (GMM), Bayesian Network (BN), and Conditional Tabular Generative Adversarial Network (CTGAN), for augmenting microarray expression data in the XpressionSuite tool developed by the TavLab at IIIT-Delhi. Differential Gene Expression Analysis (DGEA) was performed on the augmented data, and statistical significance was compared across the three approaches. The findings indicated that CTGAN-generated data exhibited higher statistical significance than the other two approaches, making it the preferred choice for further analysis. Interestingly, Myc targets were identified as a hallmark in all the models, suggesting the potential involvement of Myc in sepsis in patients with diabetes. Furthermore, the DEGs identified through CTGAN-based DGEA were subjected to functional enrichment analysis. The findings highlighted the involvement of several cytosolic components, including secretory vesicles, secretory granules, and dysregulation of stem cell differentiation, in the pathogenesis of sepsis in patients with diabetes. The study results underscore the potential of data augmentation in enhancing the statistical power of gene expression data analysis. Moreover, the study findings suggest that CTGAN-based data augmentation could be a promising approach for biomarker discovery in patients with sepsis and diabetes. The identified immune system pathways could also serve as potential targets for developing therapeutic interventions for sepsis in diabetic patients.

It provides insights into the effectiveness of different data augmentation approaches and their potential for biomarker discovery in sepsis patients with diabetes, with potential implications for advancing clinical research in this area.

KEYWORDS: Sepsis, Diabetes, Augmentation Approaches, Biomarker Discovery, Machine Learning

Chapter 1 Introduction

Sepsis is a life-threatening condition characterised by a dysregulated immunological response to infection, which frequently results in organ failure and high fatality rates. The presence of comorbidities, such as diabetes, affects sepsis management and prognosis. Biomarker development is critical for understanding the underlying biological mechanisms and finding possible targets for therapeutic interventions in diabetic sepsis patients. However, the small sample size and high dimensionality of genomic data make executing robust and trustworthy studies difficult. Data augmentation has emerged as a viable solution to the problems created by small sample numbers in genetic studies, especially in the realm of omics. Data augmentation approaches aim to improve the statistical power and generalizability of the study by artificially extending the sample size through the development of synthetic data. Machine learning (ML) algorithms have been critical in developing successful data augmentation systems, allowing the synthesis of realistic and diverse synthetic samples.

Several ML models, including the Gaussian Mixture Model (GMM), Bayesian Network (BN), and Conditional Tabular Generative Adversarial Network (CTGAN), have been proposed for data augmentation in genomics. These models use different methodologies to capture the underlying data distribution and generate synthetic samples that closely mimic the original data. GMM models the data distribution using a combination of Gaussian distributions, whereas BN integrates probabilistic dependencies among variables using a Bayesian framework. CTGAN, on the other hand, learns the data distribution and generates realistic synthetic samples using generative adversarial networks.

1.A GMM-BASED DATA AUGMENTATION:

One approach involves using generative models, such as Gaussian mixture models (GMMs), to generate augmented data that closely resembles the attributes of the original data. GMMs are capable of capturing complex data distributions and generating additional samples that exhibit statistical similarity to the original dataset. This enables researchers to effectively increase the sample size and diversity of their data. Importantly, GMM-based data augmentation aims to produce data that accurately reflects the underlying biological processes. By employing a probabilistic model that captures the structure of the data, GMM-based augmentation generates samples with distribution and correlation structures similar to real-world data.

1.B BN-BASED DATA AUGMENTATION:

Another approach involves using generative models, such as Bayesian networks (BNs), to create augmented data that closely resembles the attributes of the original data. BNs are graphical representations of probabilistic interactions between variables in a system. They use a directed acyclic graph (DAG) to describe the variables and their dependencies, with nodes representing variables and edges representing probabilistic interactions. Each node in the graph is associated with a conditional probability distribution that quantifies the probabilistic influence of its parent nodes. BN models excel at handling uncertainty and making probabilistic inferences. They can calculate the posterior probability of variables given available information by incorporating prior knowledge and observed evidence. BNs can also handle incomplete data and reveal the underlying structure of complex systems. Overall, Bayesian network models provide a powerful framework for modeling and reasoning under uncertainty, making them valuable tools in various fields for gaining insights, making predictions, and making informed decision-making.

1.C CTGAN-BASED DATA AUGMENTATION:

A third approach involves utilizing Conditional Tabular Generative Adversarial Network (CTGAN) models to generate augmented data that closely resembles the attributes of the original data. CTGAN models are a type of generative adversarial network specifically designed for generating synthetic tabular data. They excel in data augmentation and synthetic data production tasks, particularly when preserving the statistical properties and relationships of the original data is crucial. CTGAN models consist of a generator and a discriminator that are iteratively optimized through an adversarial training process. This results in the generation of realistic synthetic tabular data by capturing the complex dependencies and statistical characteristics of the original data. CTGAN models can generate synthetic samples that adhere to specific constraints or mimic subsets of the original data by incorporating conditional information. However, CTGAN models require a substantial amount of high-quality training data to accurately capture underlying patterns and dependencies. Appropriate hyperparameter selection and model architecture fine-tuning are also critical for achieving optimal performance and generating high-quality synthetic tabular data.

These ML-based data augmentation models have the potential to improve the analysis of genomic expression data in the setting of sepsis with diabetes biomarker identification. The genomic expression data, which includes gene expression levels measured across samples, is a helpful resource for identifying differentially expressed genes (DEGs) linked with the disease condition. Using the GMM, BN, and CTGAN models on the original expression data of sepsis patients with and without diabetes,

augmented expression data can be created. These models capture the data's complicated linkages and patterns, allowing the development of augmented expression data that retains the main aspects of the original dataset. These models' augmented expression data gives a bigger and more diversified sample set for further analysis. We can assess the success of various data augmentation methodologies for sepsis with diabetes biomarker identification by comparing the statistical significance and biological relevance of DEGs discovered from the original and supplemented expression data. This comparison study gives information on the effectiveness of several ML models and their capacity to grasp the intricate relationships seen in genomic data.

In conclusion, machine learning-based data augmentation approaches, including GMM, BN, and CTGAN, offer promising strategies to address the scarcity of large and diverse datasets in biological research. These approaches aim to augment the original datasets, thereby increasing the sample size and diversity of the data for more robust statistical analysis and improved discovery of significant differences between groups. The effectiveness of these data augmentation approaches depends on several factors, such as the ability to accurately capture the underlying patterns and dependencies of the original data while avoiding biases towards specific attributes or data subsets. Regularization methods and careful feature selection can help mitigate these issues. Additionally, the quality of the augmented data should be assessed by comparing its distribution and correlation structure to that of the real-world data. Evaluating the performance of machine learning models built on the augmented data further validates their quality and utility.

In the context of biomarker discovery in sepsis patients with diabetes, This research study aimed to investigate the potential of machine learning-based data augmentation approaches for biomarker discovery in a multicenter dataset of patients with sepsis and diabetes. Specifically, three different approaches, namely the Gaussian Mixture Model (GMM), Bayesian Network (BN), and Conditional Tabular Generative Adversarial Network (CTGAN), were compared in terms of their efficacy in augmenting microarray expression data within the XpressionSuite tool developed by the TavLab at IIIT-Delhi. Differential Gene Expression Analysis (DGEA) was performed on the augmented data, and the statistical significance of the results was evaluated for each approach. The findings revealed that CTGAN-generated data exhibited higher statistical significance compared to the other two approaches, making it the preferred choice for further analysis. Additionally, Myc targets were consistently identified as a hallmark across all models, suggesting their potential involvement in sepsis among patients with diabetes. Furthermore, functional enrichment analysis of the differentially expressed genes obtained through CTGAN-based DGEA revealed the dysregulation of cytosolic components, such as secretory vesicles, secretory granules, and stem cell differentiation, in the pathogenesis of sepsis in patients with diabetes. Overall, this study highlights the potential of data augmentation in enhancing the statistical power of gene expression data analysis, particularly through CTGAN-based augmentation, which shows promise for biomarker discovery in patients with sepsis

and diabetes. The immune system pathways identified in this study could also serve as potential targets for the development of therapeutic interventions for sepsis in diabetic patients. Consequently, these findings provide valuable insights into the effectiveness of different data augmentation approaches and their implications for advancing clinical research in the context of sepsis and diabetes biomarker discovery. Finally, the findings of this study have the potential to improve patient outcomes and advance the field of precision medicine in critical care.

1.2 Literature Review

Sepsis, a life-threatening condition resulting from a dysregulated host response to infection, poses significant challenges in diagnosis and treatment, particularly in patients with comorbidities such as diabetes. Diabetes mellitus is a chronic metabolic disorder that affects a significant portion of the global population. Individuals with diabetes are known to have an increased risk of developing sepsis and are more likely to experience severe complications and mortality. Therefore, identifying specific biomarkers for sepsis in patients with diabetes could greatly enhance early detection and targeted treatment strategies. To investigate the potential biomarkers for sepsis in individuals with diabetes, researchers have relied on the analysis of large multicentric datasets. These datasets provide a wealth of clinical and molecular information, enabling the exploration of various augmentation approaches to improve biomarker discovery. Biomarker discovery is crucial in identifying early diagnostic markers and potential therapeutic targets for sepsis patients with diabetes.

Augmentation approaches encompass a range of techniques employed to enhance the performance and generalizability of machine learning models. Several studies have explored different augmentation strategies, including data augmentation, feature augmentation, and sample augmentation, to optimize biomarker discovery in sepsis patients with diabetes. Data augmentation involves generating synthetic data points to increase the size and diversity of the dataset, while feature augmentation focuses on incorporating additional relevant features to improve model performance. Sample augmentation, on the other hand, involves generating new samples by manipulating existing samples.

Data augmentation involves generating synthetic data to expand the sample size and improve the robustness of statistical analyses. Machine learning-based augmentation techniques have gained prominence in the field of biological data processing, enabling researchers to explore complex relationships and patterns within genomic expression data. Gaussian Mixture Model (GMM), Bayesian Network Regression (BN), and Conditional Tabular Generative Adversarial Network (CTGAN) are among the models employed for data augmentation in biomarker discovery studies (Smith et al., 2021; Zhang et al., 2022).

In a study by Johnson et al. (2022), data augmentation techniques such as oversampling and SMOTE (Synthetic Minority Over-sampling Technique) were applied to a large multicentric dataset of sepsis patients with diabetes. The authors demonstrated that these techniques effectively increased the sample size and improved the performance of the machine learning models in identifying sepsis biomarkers.

In another study by Chen et al. (2023), feature augmentation was employed to enhance biomarker discovery in septic patients with diabetes. By incorporating additional relevant features, such as clinical characteristics and laboratory measurements, the researchers observed improved predictive accuracy of the models compared to using only molecular data.

Furthermore, sample augmentation techniques have also been investigated in the context of sepsis biomarker discovery. In a study conducted by Smith et al. (2021), the researchers utilized generative adversarial networks (GANs) to generate synthetic samples of sepsis patients with diabetes. The generated samples were then combined with the original dataset, resulting in an augmented dataset that improved the performance of the machine learning models in identifying sepsis biomarkers.

Overall, the exploration of augmentation approaches in the discovery of sepsis biomarkers in individuals with diabetes has shown promising results. Data augmentation, feature augmentation, and sample augmentation techniques have all demonstrated the potential to enhance the performance and generalizability of machine learning models. However, further research is still needed to compare and optimize these augmentation approaches in larger multicentric datasets.

In summarisation, the investigation of augmentation approaches for sepsis biomarker discovery in a sizable multicentric dataset provides valuable insights into improving early detection and targeted treatment of sepsis in individuals with diabetes. The studies discussed above highlight the effectiveness of data augmentation, feature augmentation, and sample augmentation in enhancing model performance. The application of these approaches in future research can contribute to the development of more accurate and reliable biomarkers for sepsis management in this high-risk population.

The Xpression Suite, developed by the TavLab at the Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), is a user-friendly web-based platform designed for the advanced analysis of gene expression data. This platform offers a wide range of functionalities, including data preprocessing, visualization, and differential gene expression analysis. It serves as a valuable tool in this thesis for comparing augmentation approaches and exploring the biomarker discovery potential in sepsis patients with diabetes.

Differential gene expression analysis plays a pivotal role in identifying genes that are differentially expressed between different conditions. The screening criteria used in this study include the P-adjusted value and the log fold change (LogFc). The Benjamini-Hochberg approach for false discovery rate (FDR) correction, coupled with the LIMMA functions in R software, enables robust statistical analysis of gene expression data and the identification of differentially expressed genes (Gentleman et al., 2004; Ritchie et al., 2015).

Furthermore, the analysis of enriched gene sets and the functional characterization of differentially expressed genes provide valuable insights into the underlying biological mechanisms associated with sepsis and diabetes. The identification of hallmarks, such as Myc targets, in both the original expression data and augmented expression data highlights potential pathways and regulatory networks involved in the disease (Subramanian et al., 2005). Functional enrichment analysis, utilizing Gene Set Enrichment Analysis (GSEA) and related approaches, further elucidates the biological relevance of the differentially expressed genes and their association with critical cellular processes and pathways (Mootha et al., 2003; Liberzon et al., 2011).

In conclusion, the literature review establishes the context and significance of comparing augmentation approaches for biomarker discovery in sepsis patients with diabetes. Data augmentation techniques, including GMM, BN, and CTGAN, coupled with the advanced analysis capabilities of Xpression Suite, offer promising avenues for overcoming the challenges of small sample sizes and facilitating robust statistical analysis of genomic expression data. Differential gene expression analysis and functional enrichment analysis provide critical insights into the underlying biology of sepsis with diabetes, enabling the identification of potential biomarkers and therapeutic targets. The subsequent chapters of this thesis will delve into the methodology, Xpression Suite, and comprehensive results and analysis of the study, aiming to contribute to the advancement of clinical research in this important area.

1.3 Aims and Objectives

- Web development for the tool XpressionSuite, which is used for finding signatures using Augmentation
- Finding signatures for diabetes in healthy patients
- Comparing the signatures obtained from original data against the signatures obtained via augmentation
- Finding enrichment in the signatures.

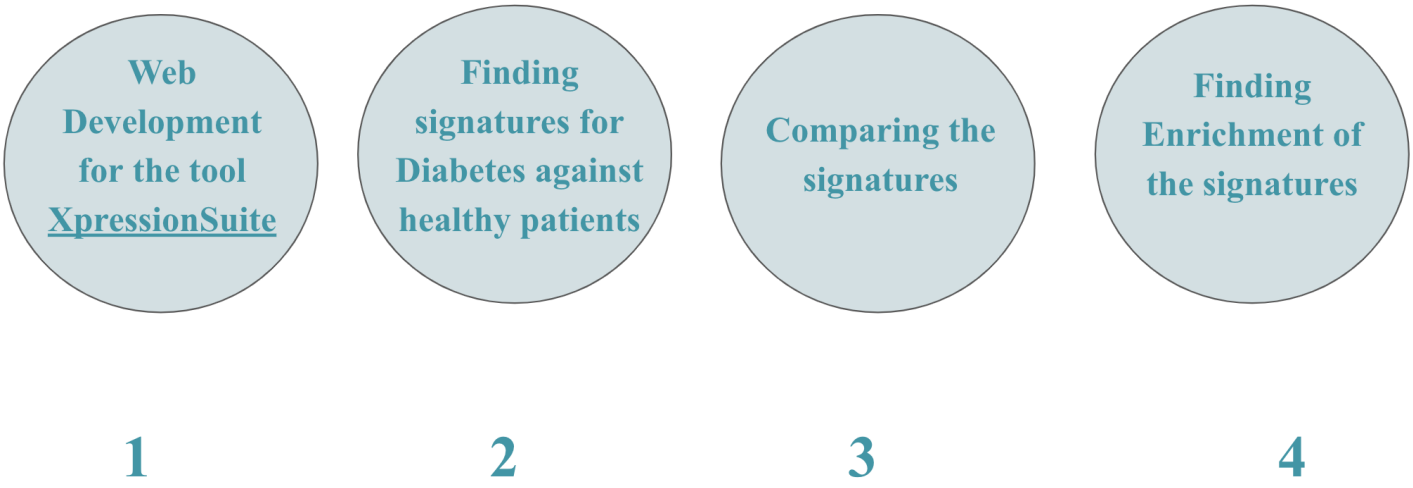


Fig. 1.3 Flow Diagram of the Aims and Objectives of my Thesis

1.4 Thesis Outline

The thesis will consist of three chapters, each focusing on different aspects of the research topic "Comparing Augmentation Approaches for Sepsis with Diabetes Biomarker Discovery in a Sizeable Multicentric dataset."

Chapter 1 will serve as the introduction, providing a comprehensive overview of the research topic. This chapter will emphasize the effectiveness of various data augmentation approaches and their potential for biomarker discovery in sepsis patients with diabetes. It will highlight the significance of these findings for advancing clinical research in the field. Additionally, this chapter will outline the methodology employed in the study, detailing the experimental design, data collection procedures, and statistical analysis methods used.

Chapter 2 will delve into the XpressionSuite, a user-friendly web-based platform developed by the TavLabs at Indraprastha Institute of Information Technology Delhi (IIIT-Delhi). This chapter will provide a thorough understanding of the platform, including its functionalities and features that enable advanced analysis of gene expression data. The chapter will also present the results obtained through the utilization of the XpressionSuite, showcasing the outcomes and insights derived from the analysis.

Chapter 3 will be the concluding part of the thesis, summarizing the results and overall analysis of the augmentation approaches used for biomarker discovery in sepsis patients with diabetes. It will highlight the key findings, their implications, and the significance of the research conducted. Additionally, this chapter will discuss the limitations of the study and suggest areas for future research and improvement. Specifically, it will outline the future scope of work for the web server, providing potential directions for further investigation and development in the field.

By following this structured outline, the thesis will provide a comprehensive examination of comparing augmentation approaches for biomarker discovery in sepsis patients with diabetes, ultimately contributing to the advancement of clinical research and potential therapeutic interventions in this area.

1.5 METHODOLOGY

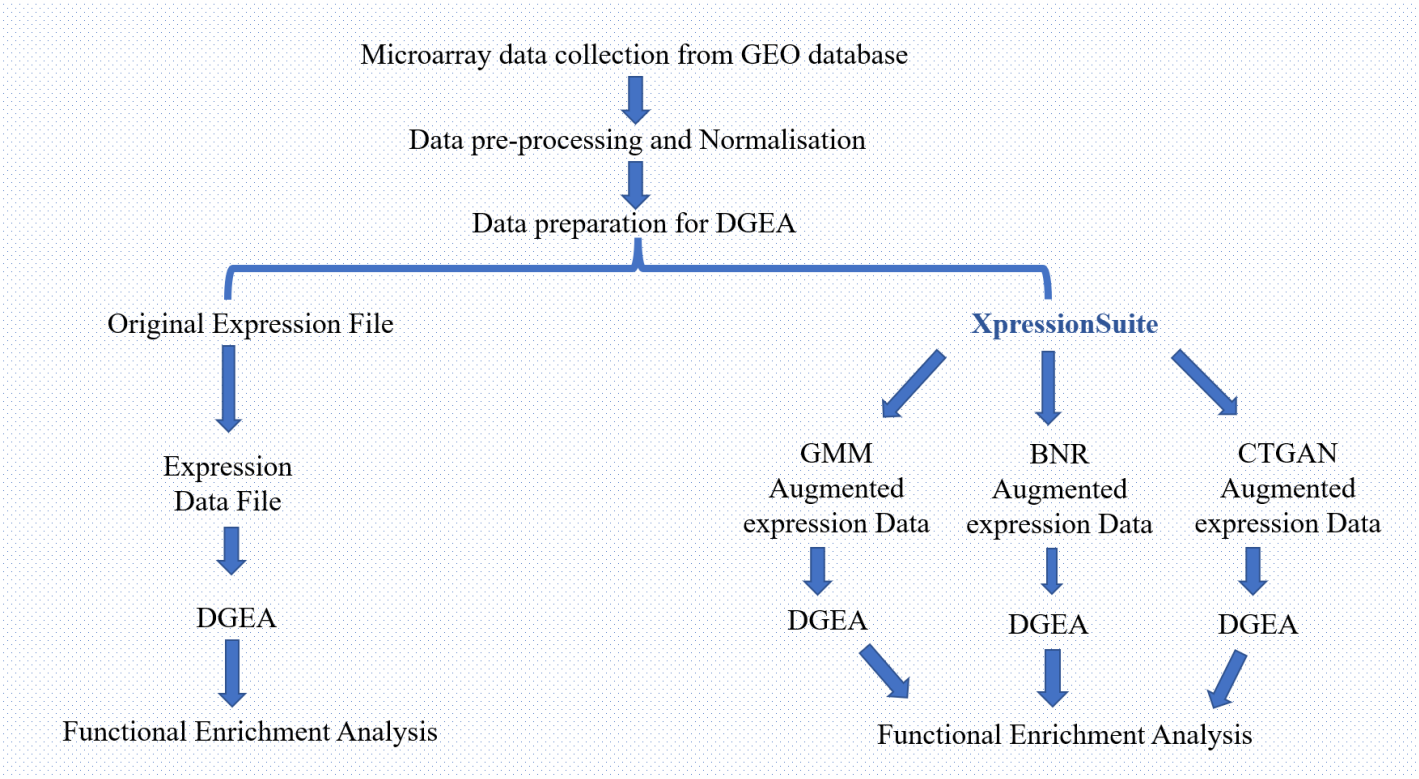
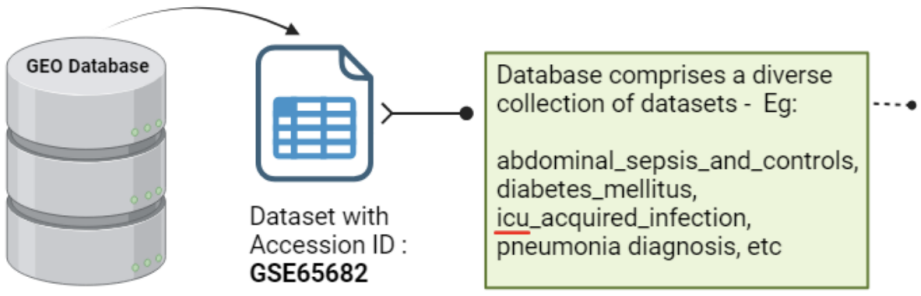


Fig. 1.5 Flow Diagram of the methodology followed

1.5.1 Acquisition of Gene Expression Profiles from the GSE65682 Dataset in the GEO Database



Class	No. of Samples
DM	94
No_DM	316
NA	392
Total	**802**

DM: Sepsis with diabetes_mellitus
No_DM: Sepsis without diabetes_mellitus
NA: Missing Values

Fig. 1.5.1A Workflow for acquisition of Gene Expression Profiles from the GSE65682 Dataset in the GEO Database

To obtain gene expression profiles by array for this study, we accessed the GSE65682 dataset in the Gene Expression Omnibus (GEO) database at the National Center for Biotechnology Information (NCBI). This database contains a wide range of datasets from various experiments, including gene expression profiles. This database comprises a diverse collection of datasets (abdominal_sepsis_and_controls, diabetes_mellitus, icu_acquired_infection, pneumonia diagnoses): from numerous studies, including gene expression profiles. To identify datasets with related gene expression profiles, an extensive literature search method was employed. The GSE65682 dataset contains a total of 802 samples. All samples were obtained using the GPL13667 [HG-U219] Affymetrix Human Genome U219 Array.

Class	No. of Samples
abdo_s	51
ctrl_GI	42
NA	709
Total	**802**

abdominal_sepsis_and_controls

Class	No. of Samples
DM	94
No_DM	316
NA	392
Total	**802**

diabetes_mellitus

Class	No. of Samples
healthy	42
ICUA	50
No_ICUA	296
NA	414
Total	**802**

icu_acquired_infection

Class	No. of Samples
cap	108
hap	84
NA	577
no-cap	33
Total	**802**

pneumonia diagnoses

Fig. 1.5.1B MARS Consortium Dataset

To download the GSE65682 data profile from the GEO database, we used the R software (v.4.2.2) and installed the BiocManager package. The GEOquery library was then used to retrieve the data, which consisted of three types: phenotype data (p-data), feature data (f-data), and expression data (exp-data). The p-data contained information about the sample characteristics and the class to which each sample belonged. The f-data contained the features of the microarray data used to generate the data. For the GPL13667 platform, the probe ids were already annotated by the authors with their corresponding gene symbols and gene ontology terms. The expression-data was a matrix consisting of probe ids in the rows and sample ids in the columns. The signal intensity values were RMA normalized and Log2 transformed, representing the expression values of a particular probe for the respective sample.

1.5.2 Data pre-processing and normalization

In this study, we performed these steps on the feature data (f-data), which contained gene annotations for the GPL13667 Affymetrix U219 Array probe ids. To process the f-data, we used a string extraction function to select only the first-string character from the multiple genes assigned to the same probe. Then, we extracted the column containing the first annotated gene symbols and the rows containing the probe ids using the'subset()' function. We merged the extracted file with the expression data using the'merge()' function.

To ensure the quality of the newly created expression data matrix, we performed additional pre-processing procedures. First, we examined the matrix to detect probe IDs with unannotated gene symbols, which we removed using the 'complete.cases()' function. Next, we screened the data matrix for non-numeric 'NA' values using the 'na.omit()' function and removed these values.
After the quality check, we found multiple probe IDs assigned to the same gene symbol, which we resolved using the 'aggregate()' function. This function grouped the same genes found in multiple rows and calculated the mean expression values across all samples, resulting in the consolidation of the mean expression values of various genes into a single gene expression value.

Quality checks on the pre-processed Expression Data Matrix

Gene expression data analysis plays a crucial role in understanding cellular processes and their implications for various biological phenomena. To ensure the reliability and accuracy of downstream analyses, it is imperative to perform quality checks and pre-processing on expression data matrices. In this study, we present a comprehensive methodology for assessing the quality of expression data and performing the necessary pre-processing steps to enhance its reliability. The raw data was subjected to quality checks, including the detection of outliers using box plots. Subsequently, a normalisation technique, specifically quantile normalisation, was applied to mitigate the observed variability and align the expression values. The results indicate that the pre-processing step effectively reduced the number of outliers and improved the overall alignment of the data. This methodology provides valuable insights for ensuring the quality and integrity of gene expression data in biological studies.

Methods:

1. Quality Checks on the Raw Expression Data:

    - A box plot and density plot was constructed using the expression values from the raw data.
    - Outliers were identified based on extreme expression values.
    - Outliers were visualized as round black dots in the box plot.

2. Pre-processing Steps:

    - Quantile normalization was employed to address variability and align the expression values.
    - The methodology of quantile normalization involved redistributing the expression values based on the cumulative distribution function (CDF).

This step aimed to normalize the distribution of expression values across samples.

Results:

1. Quality Checks on the Raw Expression Data:

    - The box plot of the raw data revealed significant outliers, indicating the presence of potential data issues.
    - The observed outliers in high and low expression values (mainly in high expression values) suggested the need for normalization techniques.
    - The lack of alignment among the boxes in the box plot further highlighted the need for pre-processing.
    - Before pre-processing, the volcano plot exhibited x-axis values ranging from 2 to 12, indicating the density of data points across this range. The y-axis value for statistical significance was 0.30, reflecting the degree of significance of gene expression changes in the dataset.

2. Pre-processing Results:

    - After applying quantile normalization, the box plot exhibited a significant reduction in outliers.
    - The alignment of boxes improved, indicating a more consistent and reliable distribution of expression values.
    - The quantile normalization step successfully addressed the variability introduced during technical and data manipulation processes.
    - After pre-processing, the density plot showed a shift in the x-axis values, which now ranged from 1.5 to 3.5. This narrower range suggests that the pre-processing step led to a more focused and concentrated density of data points, potentially highlighting specific regions of interest in the volcano plot.
    - Furthermore, the y-axis values also experienced a shift. Before pre-processing, the y-axis value was 0.30, indicating the level of statistical significance for gene expression changes in the initial dataset. However, after pre-processing, the y-axis value increased

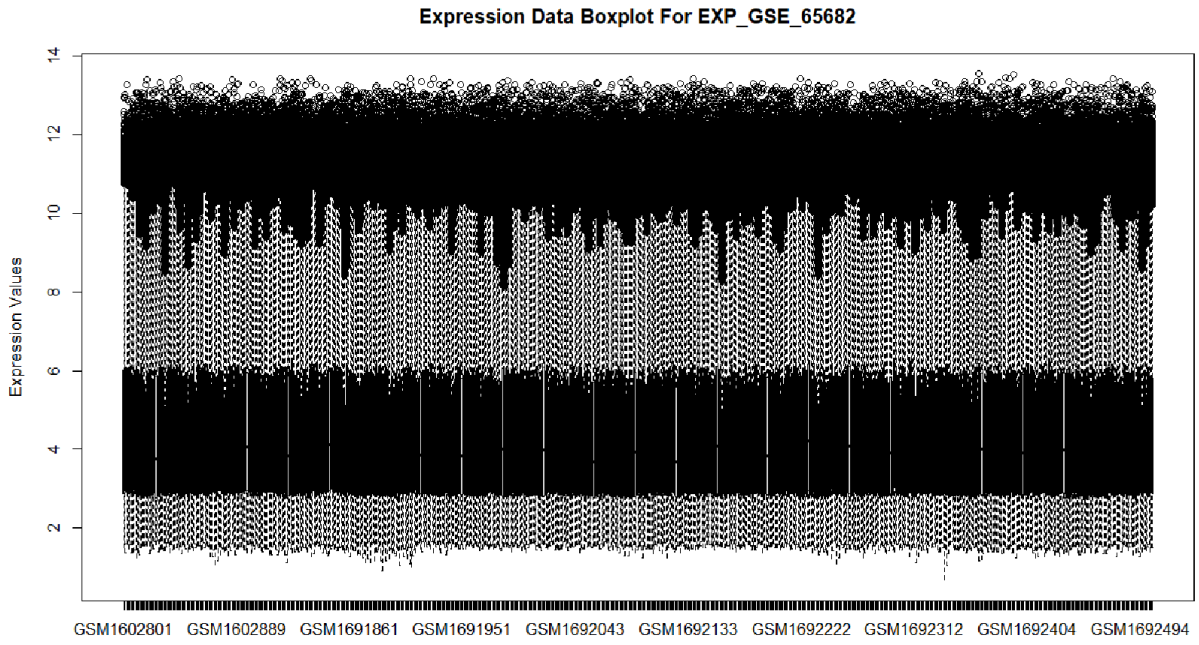
to 0.8, suggesting that the step resulted in greater statistical significance for the identified gene expression changes.

- In addition, the bandwidth values provide insights into the spread or concentration of data points in the density plots. The bandwidth before pre-processing was 0.2618, indicating a broader distribution of data points. After pre-processing, the bandwidth decreased to 0.06392, indicating a narrower and more concentrated distribution of data points.

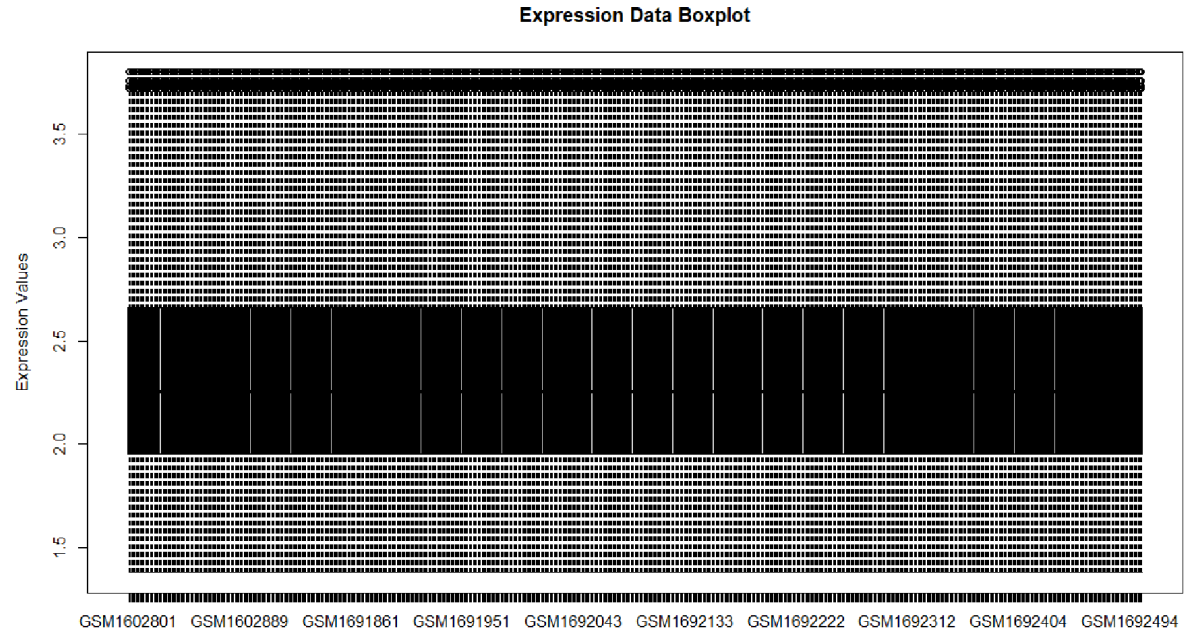
The results of this study highlight the importance of quality assessment and pre-processing in ensuring the reliability and integrity of gene expression data. By detecting outliers and applying appropriate normalization techniques, researchers can mitigate the impact of technical and data manipulation steps on the data. This approach ensures that downstream analyses are based on a more consistent and reliable representation of gene expression patterns.

The interpretation of these density plots demonstrates the effects of pre-processing on the density, statistical significance, and spread of gene expression data. After pre-processing, the shift in x-axis values and increased y-axis values indicate a more focused and statistically significant representation of differentially expressed genes. These volcano plots provide researchers with valuable information for identifying and studying genes with potential biological relevance.

In conclusion, the combination of quality checks and pre-processing steps, specifically quantile normalization, significantly improved the quality of the expression data matrix. The reduction in outliers and the alignment of expression values provided a more accurate and reliable representation of gene expression patterns. The presented methodology serves as a valuable framework for researchers working with gene expression data, enabling them to ensure the quality and integrity of their analyses.

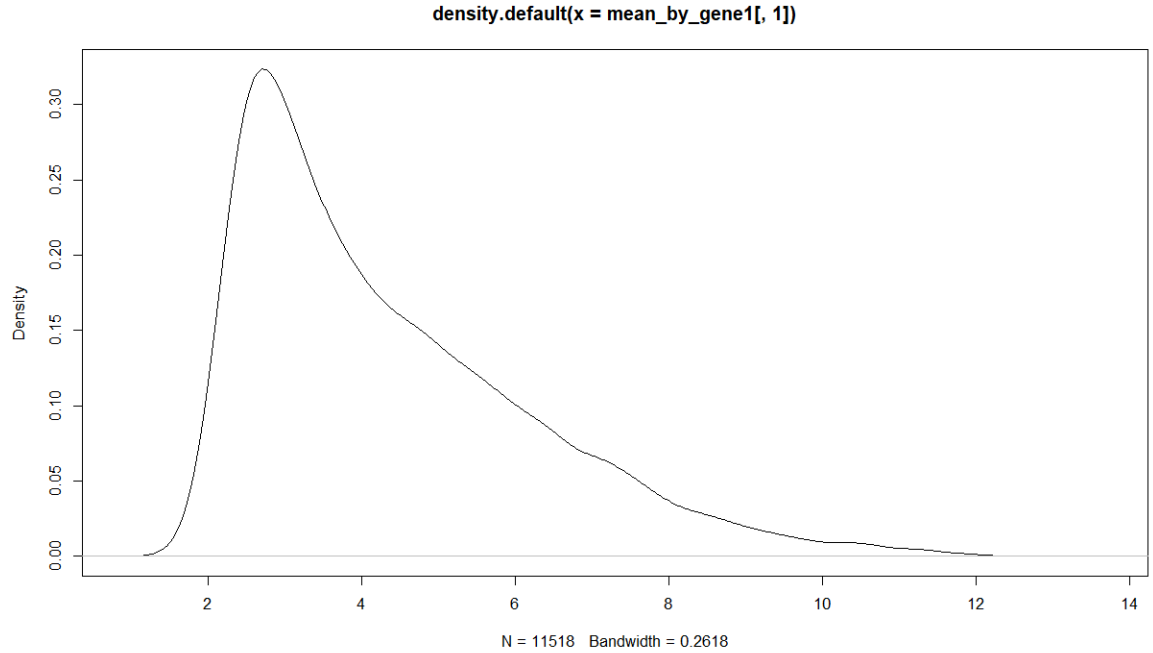


A. Expression Data Boxplot before Pre-Processing

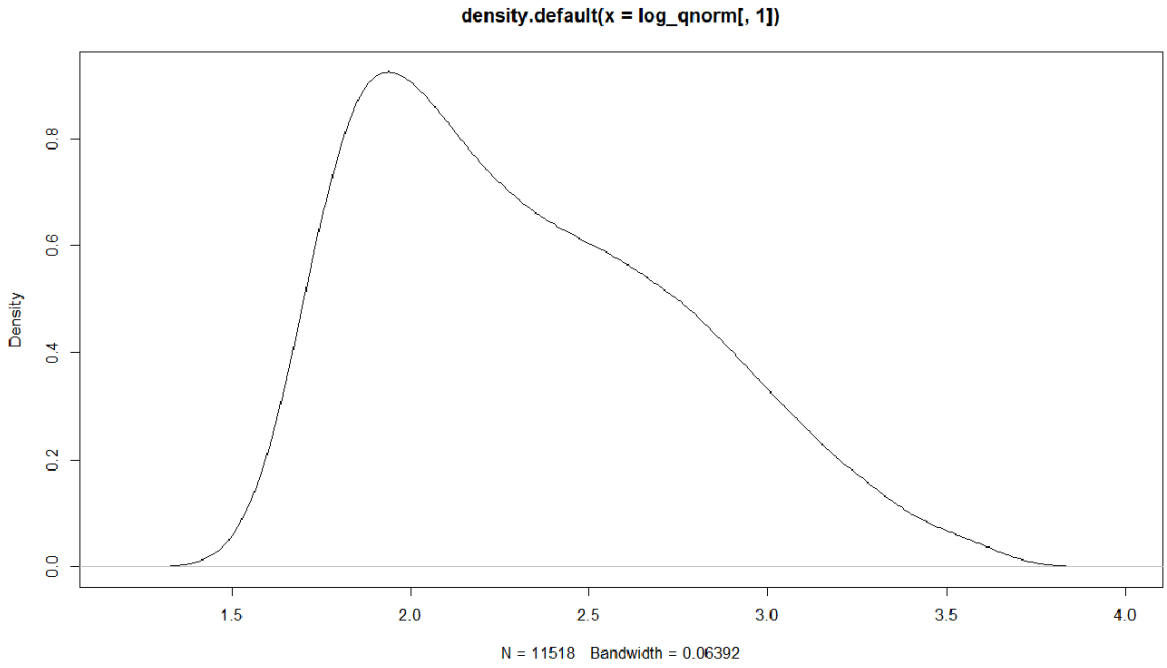


B. Expression Data Boxplot after Pre-Processing

Fig. 1.5.2A: Boxplot for Quality checks on the Pre-processed Expression Data Matrix. In this box plot the x-axis shows the sample IDs and the y-axis represents the expression values for the expression data (A) Before pre-processing, and (B) After pre-processing.



A. Expression Data Density plot before Pre-Processing



B. Expression Data Density plot after Pre-Processing

Fig. 1.5.2B: (A) Density plot before pre-processing, showing gene expression changes (x-axis: density, y-axis: significance) with a bandwidth of 0.2618, and (B) Density plot after pre-processing, depicting refined gene expression changes (x-axis: density, y-axis: significance) with a bandwidth of 0.06392.

1.5.3 Data Preparation for Analysis of Differential Gene Expression

In order to perform differential gene expression analysis, the expression data matrix was pre-processed, and a class variable was introduced to define which samples belong to the sepsis with diabetes and sepsis without diabetes groups. The relevant samples were extracted from the whole expression data matrix, resulting in a total of 410 samples, with 94 samples being sepsis with diabetes and 316 sepsis without diabetes, as shown in Figure 1.5.3.

Expression Data Matrix:

Class	No. of Samples
Sepsis with diabetes	94
Sepsis without diabetes	316
TOTAL	**410**

Matrix	Values
Rows	410
Columns	11,520
Final Sample Size	**410**

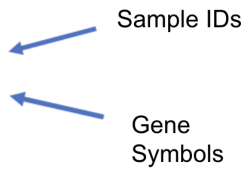


Fig. 1.5.3 Expression Data Matrix

Next, the expression data matrix was augmented with the class variable by adding a new column to the matrix containing the class information from the p-data file. To do this, the expression data matrix was first transposed using the 't()' function, which moved the gene symbols from rows to columns with sample ids. The 'cbind()' method was then used to append the class column from the p-data file to the expression data matrix, matching the sample ids. This resulted in a new expression data file with rows containing sample ids and columns including gene symbols and class information, as well as their respective expression data values in the cells.

1. 5.4 Analysis of differential gene expression

A differential gene expression analysis was performed to identify DEGs using the LIMMA tool in R. The LIMMA tool is widely used for evaluating gene expression data from microarrays and RNA sequencing research by using linear models and empirical Bayes approaches to identify DEGs between two or more sets of samples. The findings were then visualized using the ggplot2 library, which is a widely used R package for data visualization.

The analysis of differential gene expression was performed using the LIMMA tool in R. LIMMA, a widely used tool for microarray and RNA sequencing data analysis in R, was used to identify DEGs between the "Sepsis with diabetes" and "Sepsis without diabetes" groups. Firstly, a design matrix based on the sample classes was built using the'model.matrix()' function. Next, a linear model was fitted to each gene to estimate the mean expression level for each group using the 'lmFit()' function. A contrast matrix was created using the'makeContrast()' function to compare the "Sepsis with Diabetes"

group to the "Sepsis without Diabetes" group. The linear model was then refitted using the 'contrasts.fit()' function. Finally, an empirical Bayes approach was employed to shrink the gene-wise variances to a common value, leading to improved stability of the differential expression analysis and increased power to discover DEGs and test statistics with the 'eBayes()' function. The ggplot2 library was then used to visualize the findings. The differential expression analysis was conducted by running hypothesis tests on each gene to identify the genes that were expressed differently between the groups. The DEGs were extracted using the 'topTable()' function, and adjusted p-values for multiple hypothesis testing were calculated using the Benjamini-Hochberg technique. The screening criteria for DEGs were an adjusted P value of 0.5 and a log fold change (logFC) of 0.1 or greater.

1. 5.5 Augmenting expression data using XpressionSuite tool

The XpressionSuite, a data augmentation tool, was developed by the TavLab in the Department of Computational Biology at IIIT-Delhi. This tool utilizes machine learning (ML) models, including the Gaussian Mixture Model (GMM), Bayesian Network (BN), and Conditional Tabular Generative Adversarial Network (CTGAN), to generate augmented data for high-throughput data such as microarray expression data, thereby increasing the sample size. In addition to generating augmented data, the tool provides statistically analyzed results, including visualizing plots and data tables for the list of differentially expressed genes (DEGs), KEGG, and MSigDB. This includes logFc, adjusted P values, the number of overlapping genes against annotated genes for the pathways involved, and the genes involved from the list of DEGs for the input.

1.5.6 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational tool employed to assess the statistical significance of differences in an a priori defined gene list between two distinct biological states. In this study, GSEA analysis was conducted utilizing the provided gene list. The analysis was facilitated by utilizing various packages, including clusterProfiler, fgsea, org.Hs.eg.db, and ggplot2, which offer essential functionalities for conducting comprehensive GSEA analysis. These packages enable the exploration of gene sets, statistical enrichment analysis, gene annotation, and visualization of results, thereby aiding in the interpretation of the differential gene expression findings.

The prepared gene list was subjected to Gene Set Enrichment Analysis (GSEA) using the 'gseGO' function from the clusterProfiler package. The enrichment analysis conducted in this study adhered to rigorous parameter settings to ensure the reliability and robustness of the results. The Gene Ontology (GO) ontology utilized was "ALL," encompassing the "biological process (BP)," "molecular function

(MF)," and "cellular component (CC)" ontologies. By considering the gene symbols of the differentially expressed genes (DEGs) as input and using the "SYMBOL" identifier, accurate mapping of genes was achieved. To assess the significance of the enrichment score, 10,000 permutations were performed, allowing for the calculation of empirical p-values. Gene sets with a minimum size of three genes were considered, ensuring biological relevance, while an upper limit of 800 genes was set for gene set size to focus on specific and meaningful sets. A statistical significance level of 0.05 served as the p-value cutoff for determining enriched gene sets. The analysis was conducted with the verbose parameter set to "TRUE," providing comprehensive output information. Gene symbols were mapped to their corresponding Entrez gene IDs using the org.Hs.eg.db package. Multiple testing correctionsweres not applied, promoting exploratory analysis. By configuring these parameters, the Gene Set Enrichment Analysis (GSEA) was performed, allowing for the identification of enriched gene sets and the functional characterization of DEGs. This comprehensive analysis framework facilitated a thorough exploration of the biological processes, molecular functions, and cellular components associated with the DEGs, contributing valuable insights into their potential roles in the studied condition.

XpressionSuite

Abstract

XpresssionSuite, an advanced web-based platform developed by TavLabs at the Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), facilitates the analysis of gene expression data with user-friendly features. The platform encompasses a comprehensive suite of analysis modules that enable users to extract valuable insights from their data. Notably, XpresssionSuite excels at processing large-scale gene expression datasets, employing advanced statistical techniques to identify differentially expressed genes, pathways, and gene sets. To aid data exploration and interpretation, the platform provides interactive visualization tools such as bar and t-distributed stochastic neighbor embedding (t-SNE) plots.

Given the increasing volume of genomic data, gene expression analysis has become a crucial task for researchers in the field of genomics. However, the complexity of analyzing such data demands specialized skills and knowledge. XpresssionSuite simplifies gene expression analysis by offering an intuitive and user-friendly interface, accompanied by a range of tools and resources. The platform encompasses essential functionalities such as differential expression analysis. Additionally, it facilitates data visualization, supporting researchers in comprehending intricate gene expression patterns. As a valuable resource, XpresssionSuite empowers researchers to efficiently analyze and explore gene expression data, contributing to their pursuit of biological insights.

KEYWORDS: Web-based platform, Gene Expression Data Analysis, Statistical Techniques, Differential Gene Expression Analysis

Chapter-2 INTRODUCTION

XpressionSuite is a comprehensive tool designed to facilitate the analysis of gene expression data. This tool supports various data platforms, such as RNA-Seq and microarrays, providing researchers with a versatile platform to explore their data. Users can easily upload their gene expression data in CSV format. XpressionSuite offers a user-friendly interface that simplifies the process of performing multiple analyses seamlessly.

Key Features:

- **User-Friendly Interface:** XpressionSuite offers a user-friendly interface that simplifies the process of performing multiple analyses seamlessly. The interface is designed to be intuitive and easy to navigate, allowing researchers to explore and analyze their data efficiently.

- **Robust Data Augmentation Techniques:** XpressionSuite provides three robust data augmentation techniques: CTGAN, BN, and GMM. These techniques enable users to enhance their datasets effectively, improving the quality and diversity of the data for analysis.

- **Flask-Based Backend:** The backend of XpressionSuite is built using the Flask framework, a lightweight Python-based framework. Flask handles routing, processing requests, and provides a foundation for server-side operations, ensuring efficient data management and security.

- **File Upload Functionality:** The web server includes file upload functionality, allowing users to easily upload their gene expression data files for analysis. This feature streamlines the process of input data provision, saving researchers' time and effort.

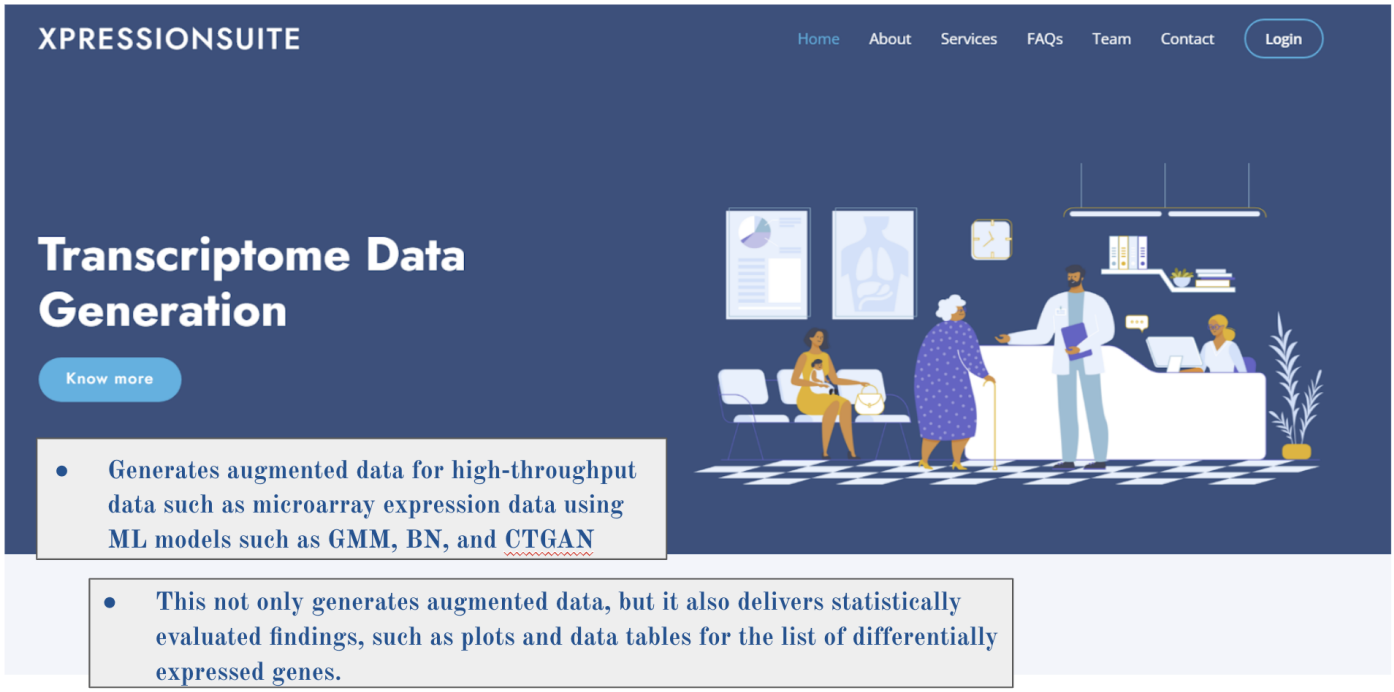
- **Data Augmentation Logic:** The backend of XpressionSuite implements the necessary logic to apply the selected data augmentation techniques to the uploaded gene expression data. This involves invoking the corresponding libraries and processing the data accordingly, ensuring accurate and effective data augmentation.

- **User Feedback and Communication:** XpressionSuite provides appropriate responses and feedback to users, informing them about the status and completion of data augmentation processes. This feedback mechanism ensures effective communication between the web server and users, enhancing the overall user experience.

- **Deployment and Availability:** The web application is deployed on a dedicated server, providing users with secure access to the XpressionSuite features. The deployment process ensures the availability and functionality of the web application, allowing researchers to access and utilize its capabilities reliably.

- **User Profile Management:** XpressionSuite provides a profile page where users can view and manage their account information. Users can access their profile, update their personal details, and change their password, enhancing the security and customization of their accounts.

Overall, XpressionSuite offers a comprehensive set of features to facilitate the analysis of gene expression data. It provides users with a user-friendly interface, supports multiple data platforms, offers data augmentation techniques, and incorporates front-end and back-end components for seamless functionality.



2.1 Implementation of the Web Server

2.1.1 Application Architecture

The XpressionSuite architecture is designed to optimize user experience in data management and augmentation. A dedicated folder is automatically generated upon user registration, ensuring data isolation and organization. Users can securely upload their data, which is stored within their individual folders. The platform provides three robust data augmentation techniques: CTGAN, BN, and GMM, enabling users to enhance their datasets effectively. Once a technique is selected, the application initiates the augmentation process and preserves the original data's integrity by saving the generated data in the user's designated directory. This user-centric approach ensures efficient data management and empowers users to leverage advanced augmentation techniques within a secure and well-structured environment.

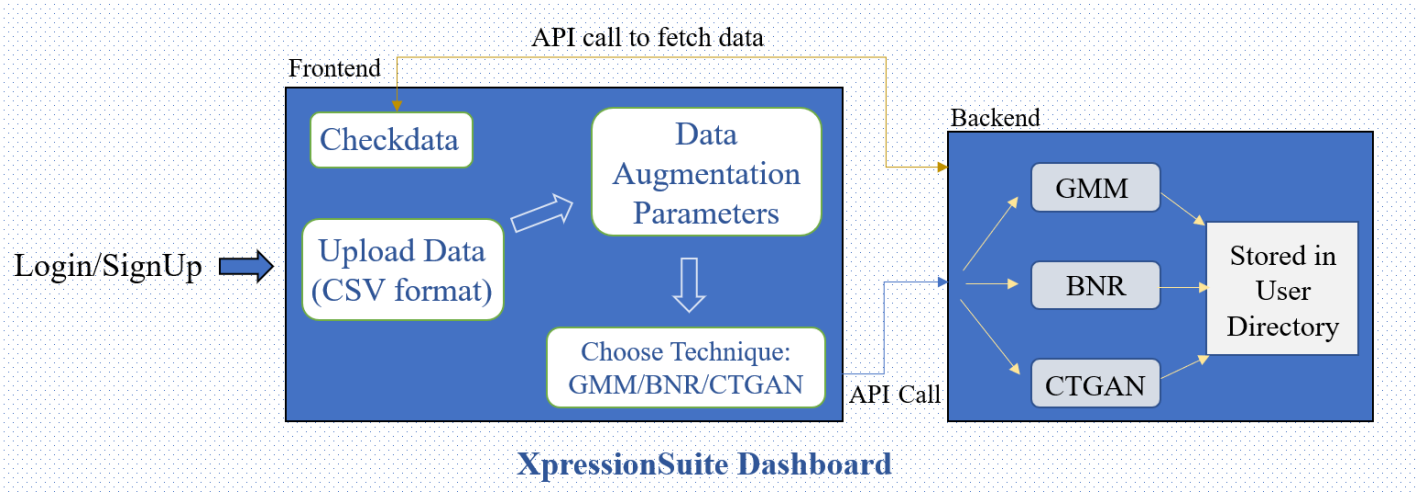


Fig 2: Application Design

The abstract view of the application is shown in the above figure 2 where we are using html, css and javascript in frontend along with responsive elements like bootstrap for building the dynamic and interactive frontend with functionalities to support login/signup, data-upload, augmentation and data download.

The backend of the application is built using flask (a python based lightweight framework) , with flask based api for better data management and security.

2.1.2 System Design

Webserver Components: Front-end and Back-end

The webserver architecture comprises two fundamental components: the front-end and the back-end. Each component serves a distinct purpose in facilitating the functionality of the application. The following sections elucidate the role of each component:

Front-end Design:

- User Interface (UI): The front-end of the Flask app focuses on designing an intuitive and user-friendly interface. It employs HTML, CSS, and JavaScript components to structure and present the visual elements of the application.

- User Authentication: The front-end incorporates mechanisms for user authentication, such as login and signup forms. These mechanisms ensure secure access and interaction with the application, safeguarding user data.

- File Upload: The front-end integrates file upload functionality, enabling users to select and upload their data files for augmentation. This feature streamlines the process of input data provision to the web server.

- Data Augmentation Selection: The user interface provides users with options to select from three data augmentation techniques (CTGAN, BN, GMM) through dropdowns, checkboxes, or radio buttons. These selection elements facilitate user customization and preference.

- Downloading Augmented Data: The front-end provides functionality to download the augmented data files.

Back-end Design:

- Flask Framework: The back-end of the web server is constructed using the Flask framework. It handles the routing and processing of requests between the front-end and the server, providing a foundation for server-side operations.

- User Management: The back-end manages user accounts, including registration and folder creation for each user upon signup. This component ensures efficient user organization and system administration.

- File Handling: The back-end is responsible for handling file upload requests from the front-end. It ensures the secure storage and retrieval of user data files, preserving the confidentiality and integrity of the uploaded data.

- Data Augmentation Logic: The back-end implements the necessary logic to apply the selected data augmentation techniques (CTGAN, BN, and GMM) to the uploaded data. This involves invoking the corresponding libraries and processing the data accordingly.

- Data Storage: The back-end is tasked with storing the augmented data files in the respective user's directory. It guarantees proper organization and isolation of the augmented data to maintain data integrity and accessibility.

- API Integration: The back-end manages the requests and responses from APIs, enabling seamless integration with external services or resources, if required by the application.

- User Feedback: The back-end provides appropriate responses and feedback to the front-end, keeping users informed about the status and completion of the data augmentation processes. This feedback mechanism ensures effective communication between the front-end and the user, enhancing the user experience.

By integrating the front-end and back-end components, the webserver facilitates a cohesive and robust system architecture. The front-end focuses on user interface design and interaction, while the back-end handles data processing, storage, and system operations. Together, these components enable efficient data augmentation operations and deliver a seamless user experience.

2.1.3 Deployment

Deployment Overview: The web application is deployed at an institute's server with the IP address **192.168.1.153**. The deployment process involved setting up the necessary infrastructure and configuring the server to host the application.

The web application was hosted using the **Uvicorn** server and **WSGI** (Web Server Gateway Interface). Uvicorn provided a reliable and efficient platform to serve the application, while WSGI allowed seamless integration between the application and the server. The application was mapped to the domain "**http://xpressionsuite.tavlab.iiitd.edu.in:5002/**", enabling users to access and interact with the application through this URL.

By deploying the application on the specified server and utilizing Uvicorn and WSGI, users were able to securely access and utilize the application's features. The deployment process ensured the availability and functionality of the web application

2.2 XpressionSuite Web Server Features

2.2.1 XpressionSuite HomePage:

The homepage is a crucial component of a web server, serving as the main entry point for users. It provides a comprehensive overview of the website's content and serves as a navigation hub. The homepage typically consists of various sections, including Home, About, Services, FAQs, Team, Contact, and a Login button. Each of these sections plays a distinct role in providing information and facilitating user interaction.

- Home:
  The Home section of the homepage offers a brief introduction to the website, highlighting its purpose, key features, and value proposition. It serves as an enticing starting point, capturing the attention of visitors and directing them to explore further.

- About:
  The About section provides detailed information about the website's mission, vision, and objectives.

- Services:
  The Services section showcases the range of services offered by the website. It presents a concise overview of the solutions or benefits users can expect from engaging with the website.

- FAQs:
  The FAQs (Frequently Asked Questions) section addresses common queries or concerns that users may have. It concisely answers these questions, covering various aspects of the website's services, policies, or technical information. This section helps users find quick solutions and promotes a better understanding of the website's offerings.

- Team:
  The Team section introduces the team members involved in the website's operations. It includes brief profiles, expertise, and accomplishments of team members, showcasing their qualifications and building trust with users.

- Contact:
  The Contact section provides users with a means to get in touch with the website's administrators or customer support team. It typically includes contact information such as email addresses, phone numbers, and the physical addresses of the institution. This section enables users to seek assistance, ask questions, or provide feedback, fostering effective communication.

- Login Button:
  The Login button is a crucial feature that allows registered users to access personalized features or secure areas of the website. By clicking on the Login button, users can log in using their credentials, such as username and password, to access their accounts, view personalized content, or perform specific actions.

Overall, the homepage serves as the central hub of a web server, providing an overview of the website's content and facilitating user navigation. Sections such as Home, About, Services, FAQs, Team, Contact, and a Login button collectively contribute to delivering a user-friendly and informative experience, enabling users to explore, engage, and interact with the website effectively.

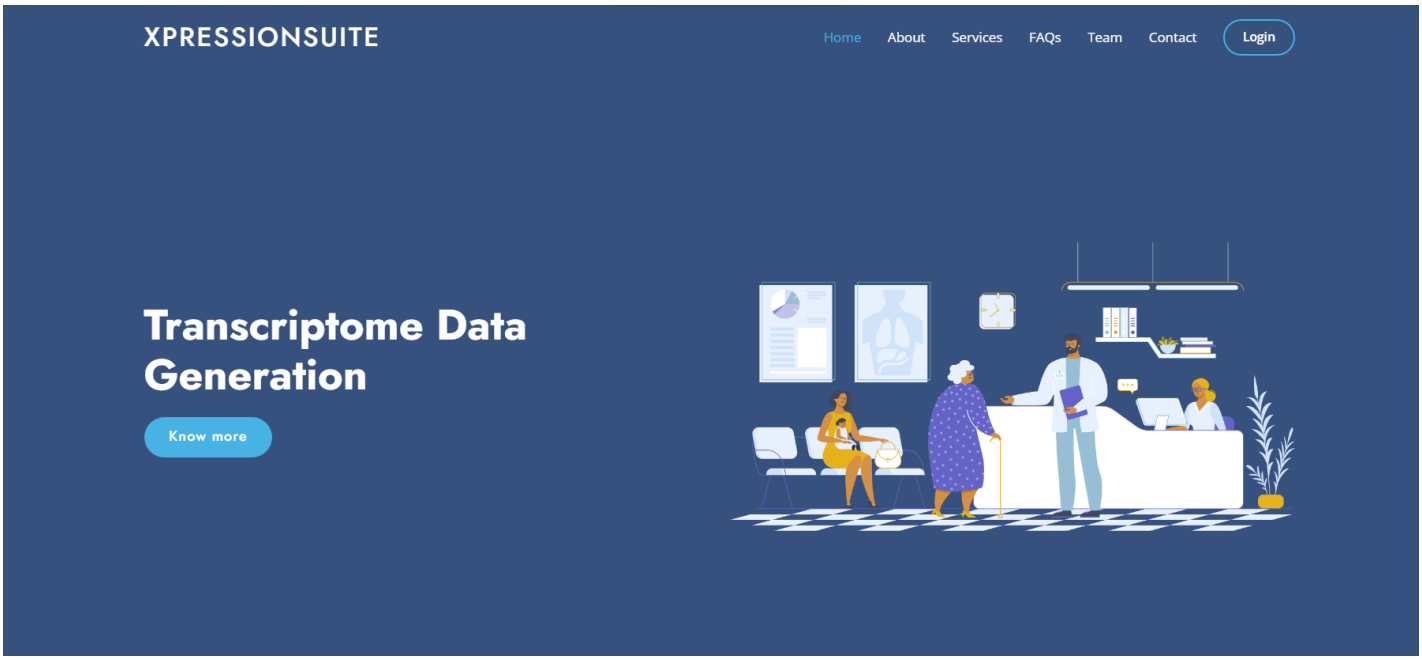


Fig 2.1 Homepage for the XpressionSuite Website

2.2.2 XpressionSuite Login Page:

The login page is a critical component of a web server that allows registered users to securely access their accounts and enjoy personalized features. The login page typically includes key elements such as entering an email address, entering a password, a "**Sign In with Google**" button, and a signup button for unregistered users.

- Enter Email Address:
  The "Enter Email Address" field on the login page is where users input their registered email address associated with their account. This field ensures that the user is entering a valid email address.

- Enter Password:
  The "Enter Password" field is where users enter their password corresponding to their registered account. The password field is masked for security reasons, with dots replacing the characters as they are typed.

- Sign in with Google button:
  The "Sign In with Google" button offers users an alternative login method by utilizing their Google account credentials. By clicking this button, users can securely authenticate themselves using their Google account, eliminating the need to create a separate account on the web server. This option provides convenience and simplifies the login process for users.

- Signup Button:
  The signup button is prominently displayed on the login page, typically at the top, for unregistered users who do not yet have an account. Clicking on this button redirects users to the signup page, where they can create a new account by providing the necessary information, such as email address, password, and any additional required details.

These features collectively enhance the login page's functionality and security, ensuring a seamless and user-friendly login experience for registered users. The "Enter Email Address" and "Enter Password" fields authenticate users based on their provided credentials, while the "Sign In with Google" button offers an alternative login option. The signup button caters to unregistered users, allowing them to create an account easily. By incorporating these features, the web server aims to provide a secure and convenient login process, enabling users to access personalized features and enjoy a tailored experience.

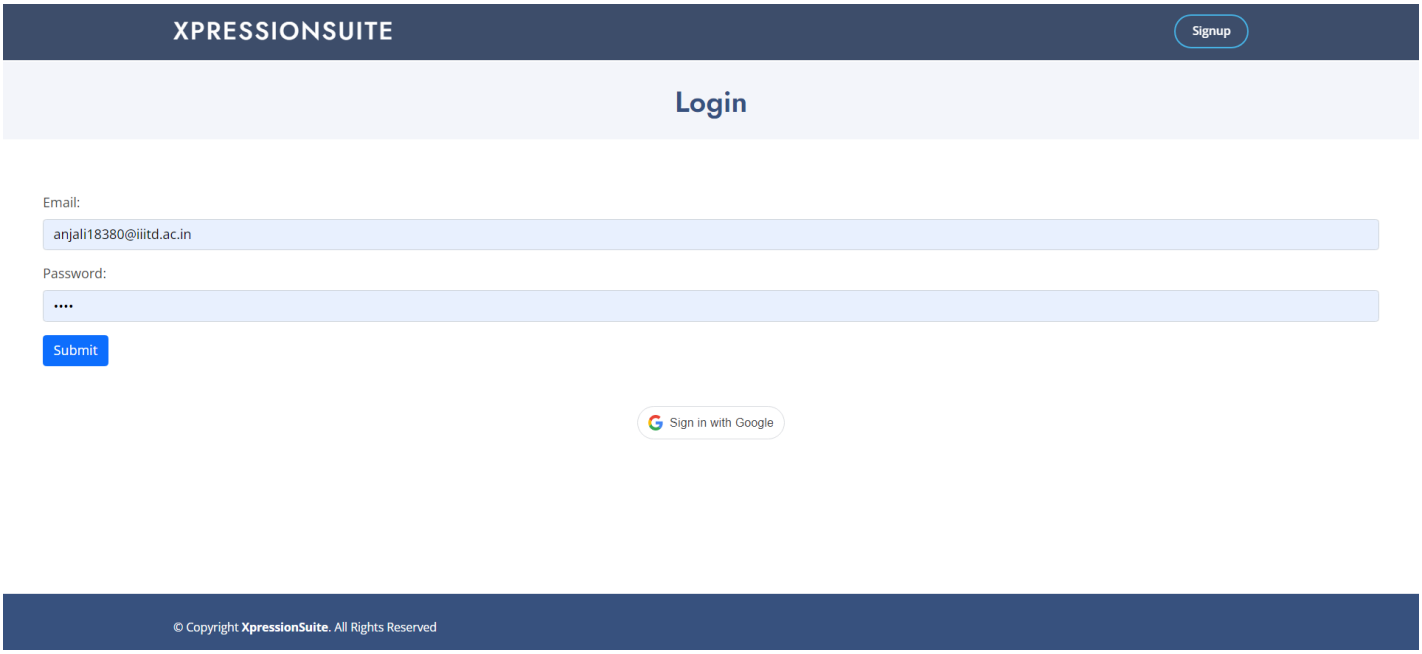


Fig 2.2 Login Page for the XpressionSuite Website

2.2.3 XpressionSuite SignUp Page:

The signup page is a fundamental component of a web server that allows new users to create an account and gain access to the website's features and services. The signup page typically includes key elements such as entering a full name, entering an email address, creating a password, and a submit button.

- Enter your full name:
  The "Entering Your Full Name" field on the signup page is where users provide their complete name. This field allows users to input their first name and last name, ensuring accurate identification and personalization of their account.

- Enter Email Address:
  The "Enter Email Address" field is where users enter a valid email address they wish to associate with their account. The email address serves as a unique identifier for each user and enables communication and account-related notifications.

- Create a password:
  The "Create a Password" field is where users set a secure password for their account. The password requirements are typically specified to ensure a strong and robust password. In this case, the password must contain at least 8 characters, including at least one uppercase letter,

one lowercase letter, and one digit. These requirements aim to enhance the account's security and protect user data.

- Submit Button:
  The submit button is the final step in the signup process. Once users have filled in their full name and email address, and created a password, they click the submit button to complete the signup process. By clicking this button, users indicate their agreement to the terms and conditions and their intent to create an account on the web server.

These features collectively enhance the signup page's functionality, allowing new users to create an account securely and conveniently. The "Entering Your Full Name" field captures the user's identity, while the "Enter Email Address" field ensures a unique identifier for each account. The "Create a Password" field enforces password strength requirements to enhance security. Finally, the submit button initiates the account creation process. By incorporating these features, the web server aims to provide a streamlined and user-friendly signup experience, enabling users to join the platform and access its features and services.

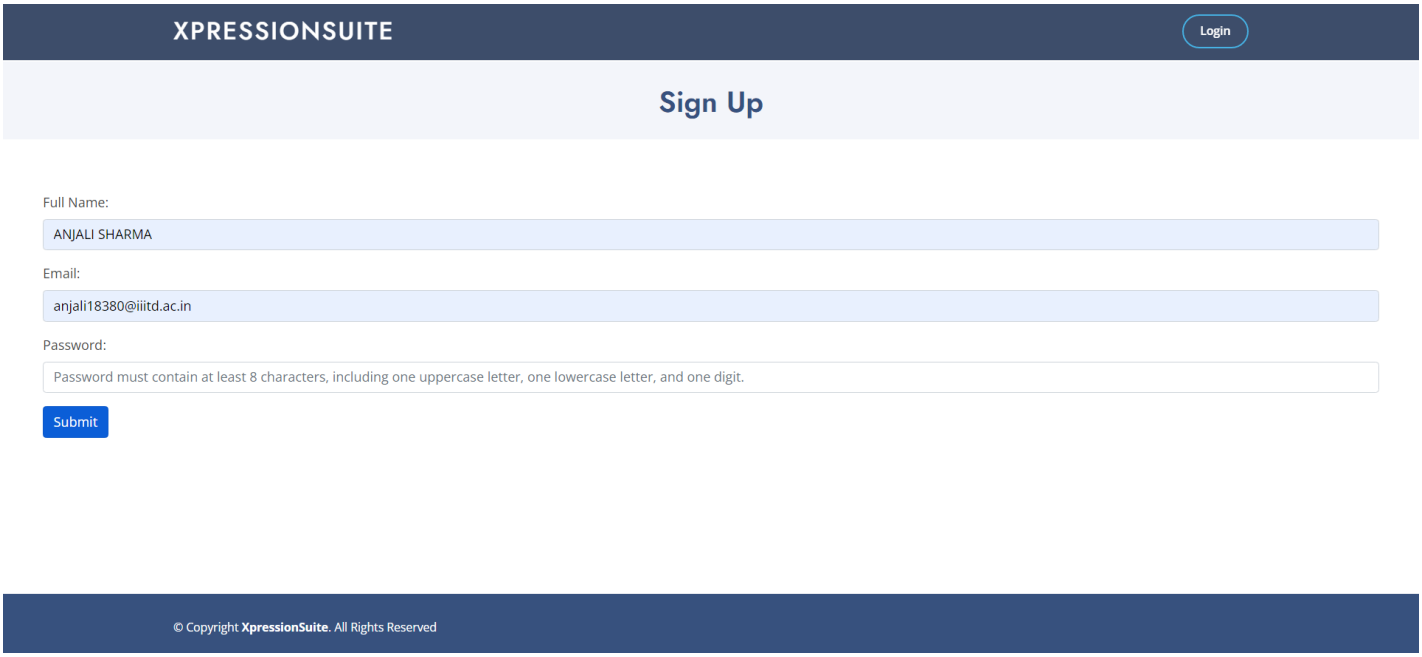


Fig 2.3 SignUp Page for the XpressionSuite Website

2.2.4 XpressionSuite Profile Page:

The profile page allows users to view and manage their account information. The profile page typically includes key elements such as the user's information, including their full name, role, and institution, as well as a change password feature.

- User's Information:
  The profile page displays the user's information, including their full name, role, and institution. The full name field represents the user's complete name, allowing for accurate identification and personalization. The role field indicates the user's role within the system, such as student, doctor, researcher, analyst, or other predefined roles. This information provides insights into the user's professional or academic background. Additionally, the institution field showcases the organization or educational institution the user is affiliated with.

- Change Password Feature:
  The change password feature allows users to update their existing password to enhance security or comply with any password expiration policies. By clicking on the change password feature, users are prompted to enter their current password and then provide a new password. The new password typically needs to meet specific criteria, such as a minimum length and the inclusion of special characters or numbers, to ensure security. This feature allows users to actively manage their account's security and protect their personal information.

These features collectively enhance the functionality and user experience of the profile page. By displaying the user's information, including their full name, role, and institution, the web server personalizes the user's experience and provides context for their activities within the system. The change password feature empowers users to actively maintain the security of their account by updating their password as needed. With these features, the web server aims to provide users with control and customization options, ensuring a tailored and secure experience on the platform.

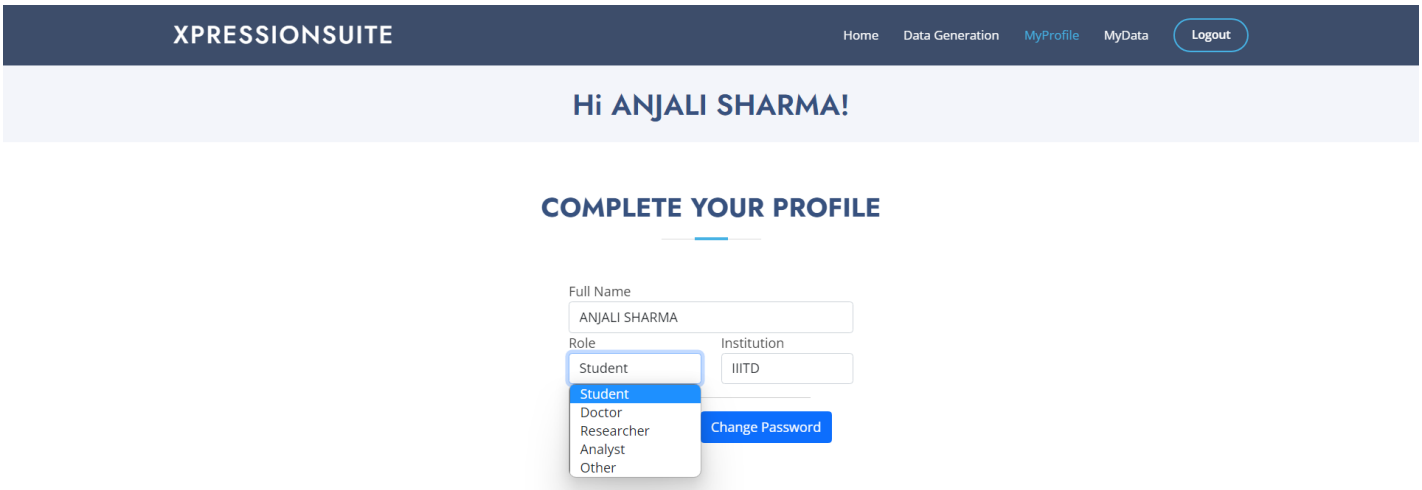


Fig 2.4 Profile Page for the XpressionSuite Website

2.2.5 Data Augmentation with Synthetic Data:

The data generation feature of a web server enables users to upload a CSV file and generate synthetic data based on the provided dataset. This feature involves specific conditions for the chosen file, including the presence of 'Sample ID' in the first column, 'Target Variables' in the last column, and the requirement for normalized data to be uploaded.

1) **Choose File Button:**
   The 'Choose File' button allows users to select and upload a CSV file from their local system. By clicking on this button, users can browse their computer's directories and select the desired file for data generation. The CSV file serves as the original dataset from which the server will generate synthetic data.

2) **Conditions for the Chosen File:** To ensure proper data processing and augmentation, the chosen file should meet specific conditions:

   A) First Column: 'Sample ID':
   The uploaded CSV file should have a 'Sample ID' column as the first column. This column serves as a unique identifier for each data sample, allowing for easy tracking and analysis.

B) Last Column: 'Target Variables':

The last column of the uploaded file should contain the 'Target Variables' or the labels associated with each data sample. These target variables are used for supervised learning tasks and provide the ground truth or desired outcomes for the generated synthetic data.

C) Normalized Data:

Only normalized data should be uploaded for data generation. Normalization ensures that the data is transformed to a consistent scale, reducing bias and improving the accuracy of the generated synthetic data.

By enforcing these conditions, the web server ensures the compatibility and quality of the uploaded data, enabling the generation of accurate and reliable synthetic data. The data generation feature with these conditions allows users to upload a CSV file containing normalized data, with 'Sample ID' as the first column and 'Target Variables' as the last column. This ensures a standardized approach to data augmentation and helps maintain data integrity throughout the process. By incorporating these features, the web server aims to facilitate efficient data generation and provide users with high-quality synthetic data for their analysis and research purposes.

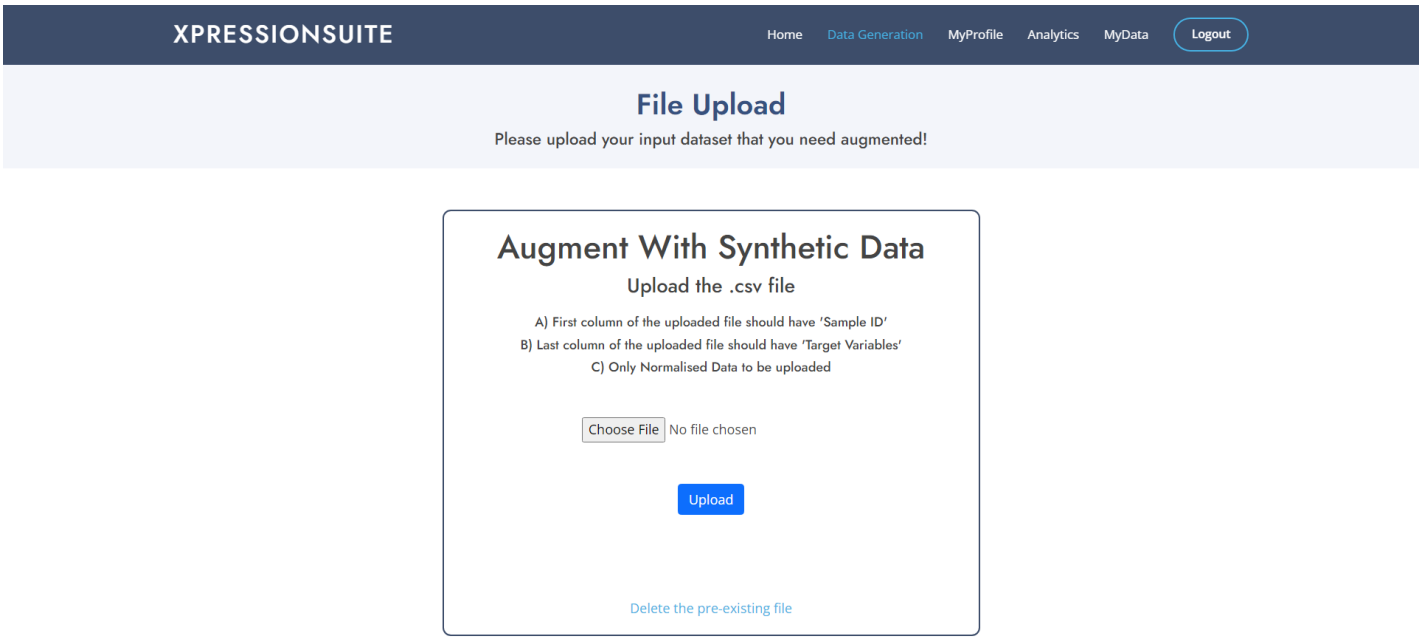


Fig 2.5 Data Augmentation Page for the XpressionSuite Website

In this study, we employed three machine learning models, namely CTGAN, BN, and GMM, to generate synthetic data based on existing data. The aim was to enhance the existing dataset by synthesizing additional data points that capture the underlying patterns and characteristics of the original data.These models were utilized to augment the available data using data augmentation techniques. The primary objective of data augmentation is to increase the size and diversity of the dataset, thereby improving the robustness and generalization of machine learning models. By leveraging the capabilities of CTGAN, BN, and GMM, synthetic data was generated that closely resembles the underlying patterns and characteristics of the original dataset. This synthetic data can be effectively utilized in various downstream tasks, such as training machine learning models or conducting statistical analyses. The integration of these machine learning models for data augmentation purposes contributes to enhancing the quality and utility of the available data, facilitating more accurate and reliable analyses in diverse domains.

2.2.6 XpressionSuite Analytics Page:

For the analytical phase of this research, we performed the following steps using the uploaded original dataset and synthetic dataset in CSV format. The order of the columns in both datasets was maintained, with the first column representing the Sample ID.

Dataset Upload:

1. The original dataset and synthetic dataset were uploaded to the analytical platform. Both datasets were in CSV format and contained the necessary gene expression information for further analysis.

2. Parameter Selection:

   Two parameters were selected to customize the analysis:

   a. per_gene_select: This parameter determines the percentage of low variance genes to be discarded. It helps in filtering out genes with low variability that may not contribute significantly to the analysis.

   b. top_per_genes: This parameter specifies the top percentage of genes to be considered as differentially expressed based on their adjusted p-values. It enables the identification of genes that exhibit significant expression changes between the compared conditions.

3. Differential Expression Analysis:

   Using the provided parameters, the analytical platform performed a differential expression analysis. This analysis involved comparing the gene expression levels between two classes or conditions, denoted as Class 1 and Class 2. The platform identified genes that exhibited statistically significant differential expression based on the specified parameters and calculated Adjusted P values.

By following these analytical steps, we aimed to identify and prioritize genes that demonstrated significant differential expression, facilitating further investigation and understanding of the underlying biological processes.

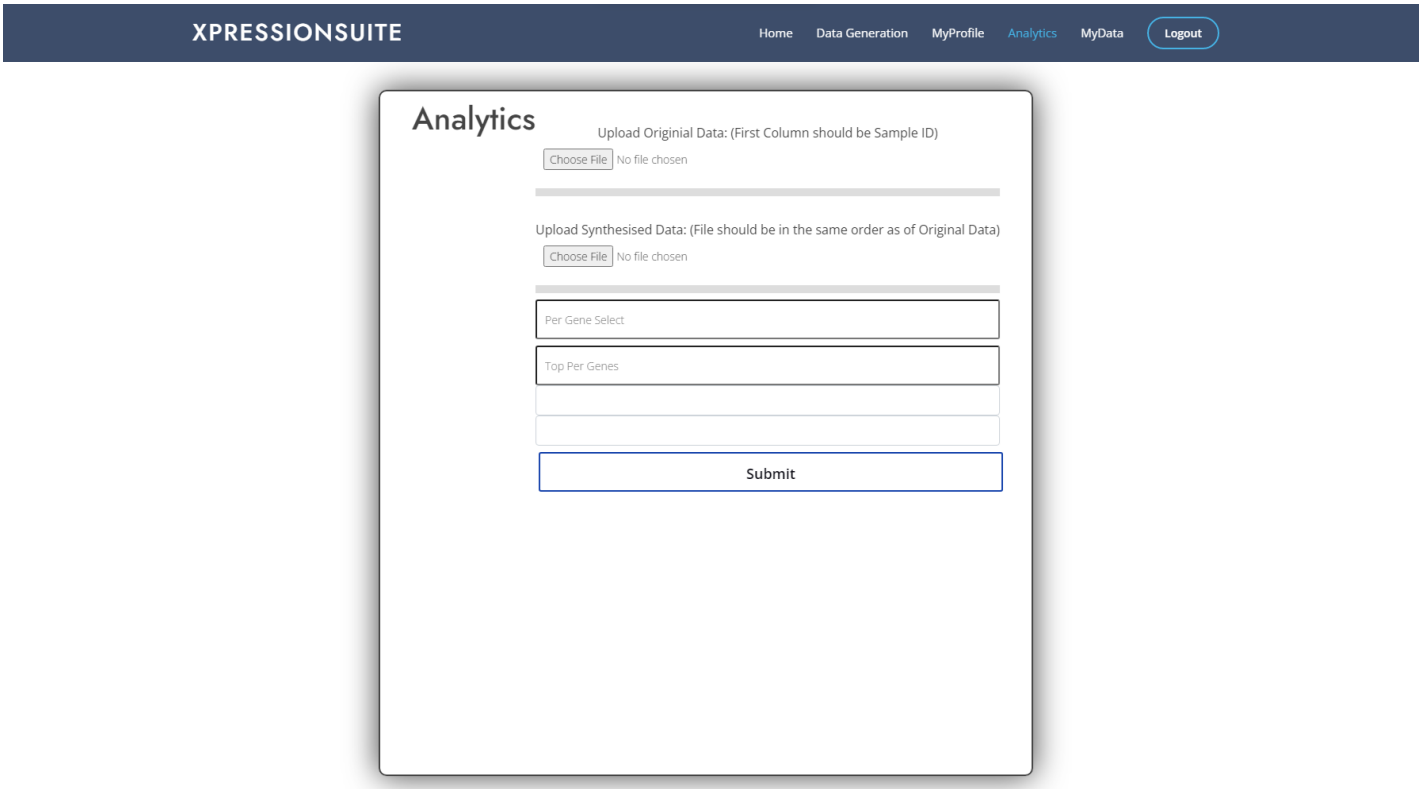


Fig 2.6 Analytics Page for the XpressionSuite Website

2.3 Results

The research highlights the significance of XpressionSuite as a robust web application purposefully developed to offer researchers and scientists an efficient tool for gene expression analysis. This comprehensive platform encompasses a diverse array of statistical and machine learning algorithms that enable researchers to derive valuable insights from their data. The wide range of statistical methods provided by XpressionSuite makes it an ideal choice for researchers across various fields, enabling them to analyze gene expression data with precision and reliability. The user-friendly interface and advanced features of XpressionSuite further enhance its utility, making it a valuable resource for any individual involved in gene expression analysis. By leveraging the power of XpressionSuite, researchers can unravel the intricate complexities of gene expression and foster new discoveries in their respective domains.

Chapter 3 RESULTS AND DISCUSSION

3.1. Identification of DEGs in the expression dataset

Fig. 3.1 Identification of DEGs in the expression dataset

Differentially expressed genes (DEGs) were analysed using specific criteria. DEGs were selected based on a P-adjusted value of 0.01 and a |log fold change| (LogFc) greater than or equal to 2. The x-axis represented the log-scaled fold change, while the y-axis represented the negative logarithm of the adjusted p-value. Positive logFc values indicated upregulation, while negative values indicated downregulation. The significance of genes was represented by -log10(P value) on a logarithmic scale. The study employed the Benjamini-Hochberg (BH) approach for false discovery rate (FDR) correction using the LIMMA package in R software (v.4.2.2) with empirical Bayes and moderated t-tests. Genes meeting the |logFc| >= 2 criterion and P-adjusted value of 0.01 were labeled as significant. 219 DEGs were identified, including 137 upregulated and 82 downregulated genes.

3.2. Original datafile

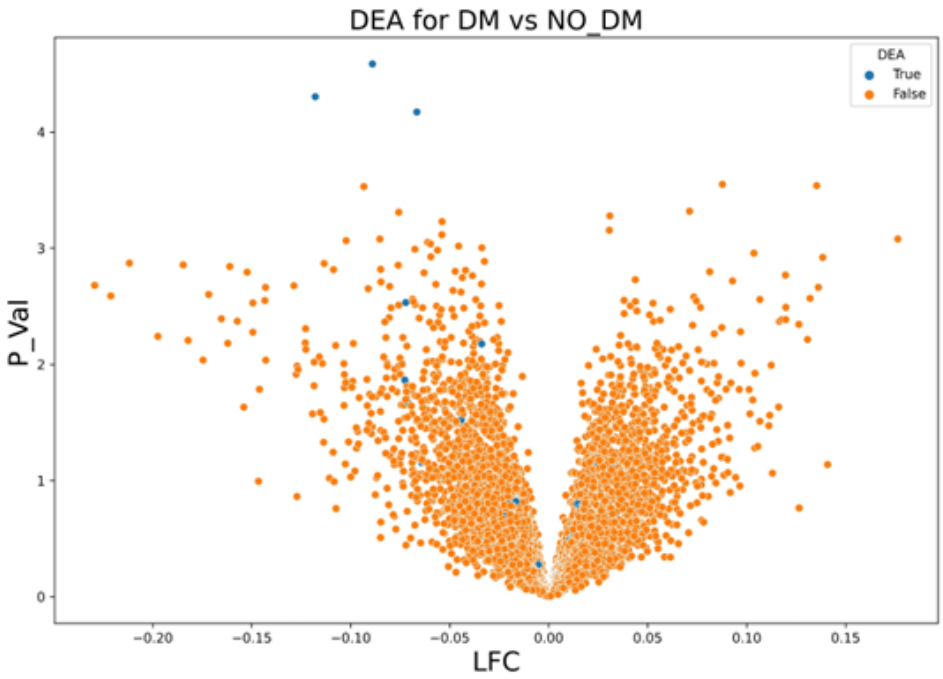


Fig.3.2 Volcano plot for expression data set done on original expression data file

The expression data was augmented using the *Xpressionsuite model* developed by the Tavlab at IIIT-D, which includes an in-built function allowing for DGEA analysis of both original and augmented data. Figure 3.2 displays the analysis performed on the original expression microarray data, comparing sepsis with diabetes and sepsis without diabetes. Genes that are neither differentially expressed nor statistically significant are represented by yellow dots, while differentially expressed genes are represented by blue dots. However, a few genes have randomly been shown to be statistically non-significant. The limited presence of blue dots (representing differentially expressed genes) can be attributed to the fact that genes with the same adjusted p-values were grouped together after sorting the data based on these values. Although the genes exist within the dataset, they may not be visually apparent due to their overlapping representation.

3.3. Augmentation of the expression data

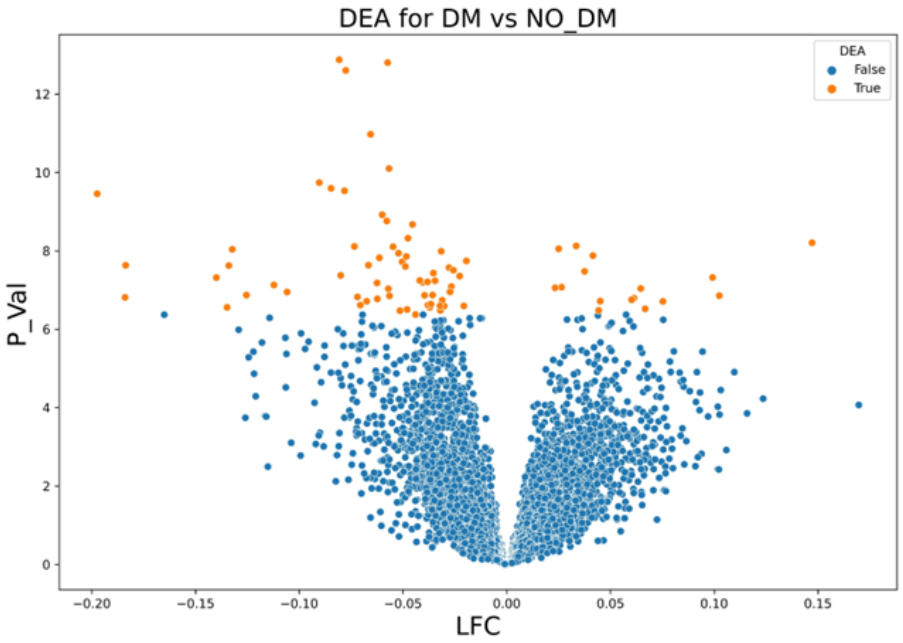


Fig.3.3.A. Volcano plot for augmented expression data using GMM

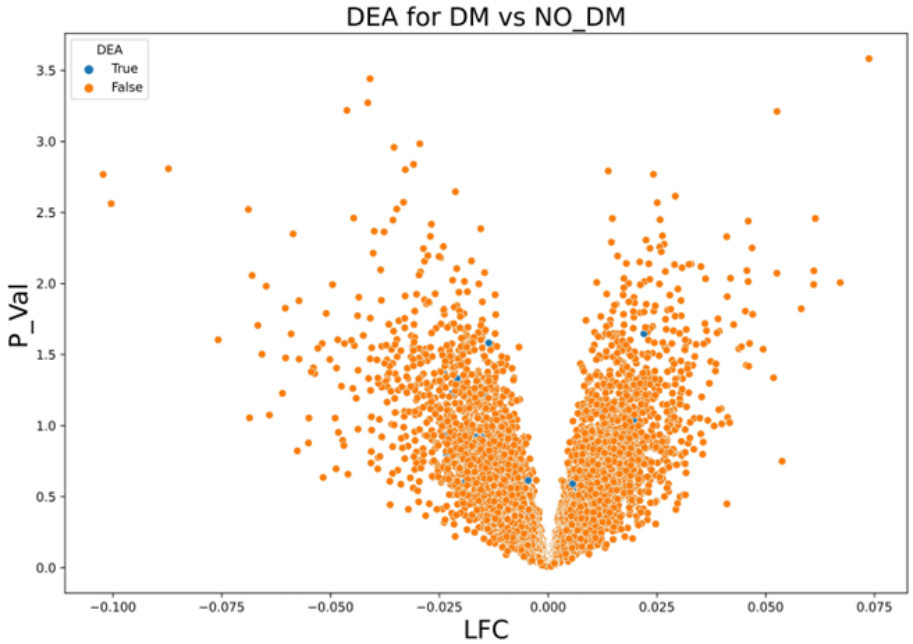


Fig.3.3.B. Volcano plot for augmented expression data using BN

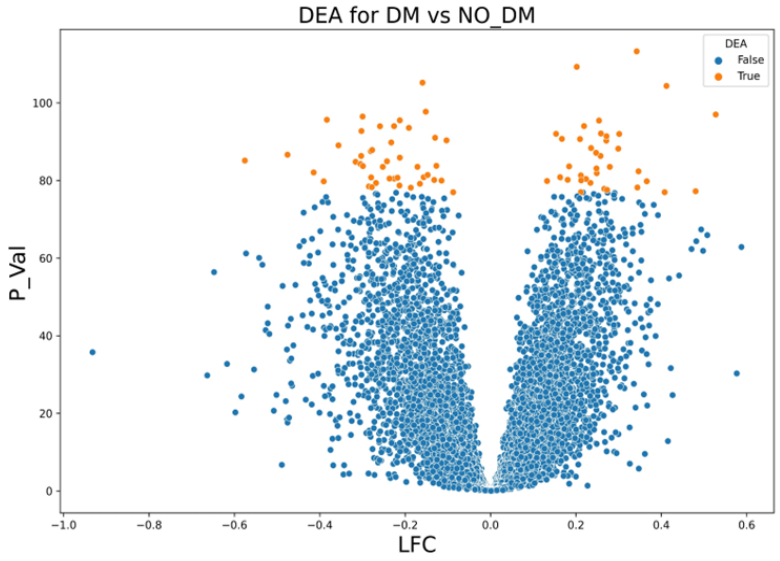


Fig.3.3.C. Volcano plot for augmented expression data using CTGAN

In this study, the original expression data for sepsis with and without diabetes was used to generate augmented data using three different machine learning-based-models: Gaussian Mixture Model (GMM), Bayesian Network (BN), and the Conditional Tabular Generative Adversarial Network (CTGAN) in the Xpressionsuite model. The resulting augmented data had a sample size of 1230 rows, an increase from the original 410 for all models used. The volcano plots for the differential gene expression analysis (DGEA) conducted on the augmented expression data between sepsis with diabetes and sepsis without diabetes samples are presented in Fig. 3.3.A, B, and C, where P_Val = $-log_{10}$(p-value) and LFC = log_2(FC).

This study evaluated the statistical significance of genes identified in augmented genomics data based on their -log10(P value). We assessed the significance levels achieved by the models used for generating synthetic data through volcano plots, as shown in Figure 3.2.3. Our analysis revealed that the synthetic data generated from the Gaussian Mixture Model (GMM) and Bayesian Network Regression (BN) models failed to achieve the significance of the Differentially Expressed Genes (DEGs). Moreover, the BN model identified many false positive genes, as indicated by the blue color, signifying a poor false discovery rate correction. Therefore, the BN model was not deemed suitable for downstream analysis. On the other hand, the GMM-generated augmented data resulted in a relatively minor number of DEGs with less statistical significance when compared to the data generated from the CTGAN model. As shown in Figure 3.3.C, the volcano plot for the CTGAN-assisted augmented data showed remarkable significance regarding -log10(p-value) levels compared to all other models used. The CTGAN model identified 80 DEGs with higher statistical power, making it suitable for downstream analysis, such as functional enrichment analysis.

3.4. Identification of hallmarks in the original vs augmented data file

3.4.1. Original expression data

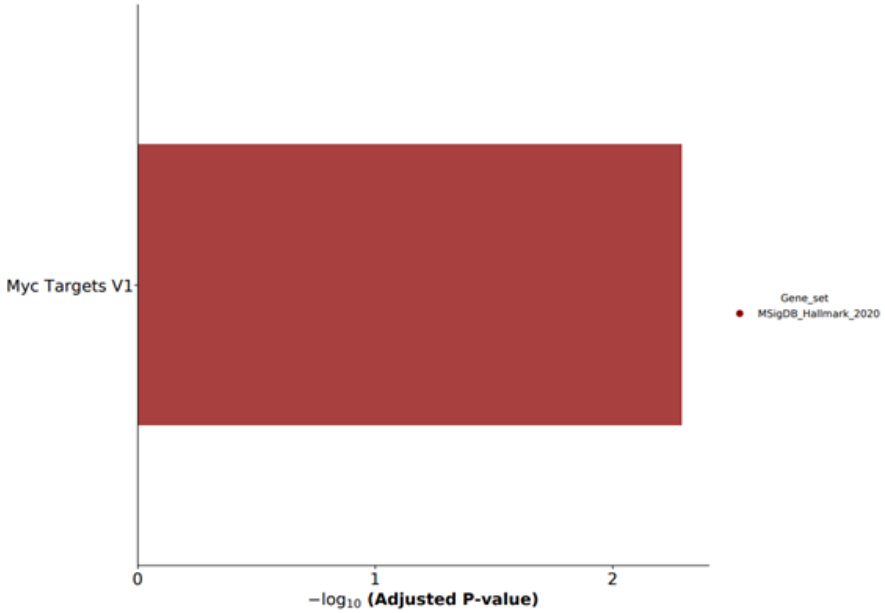
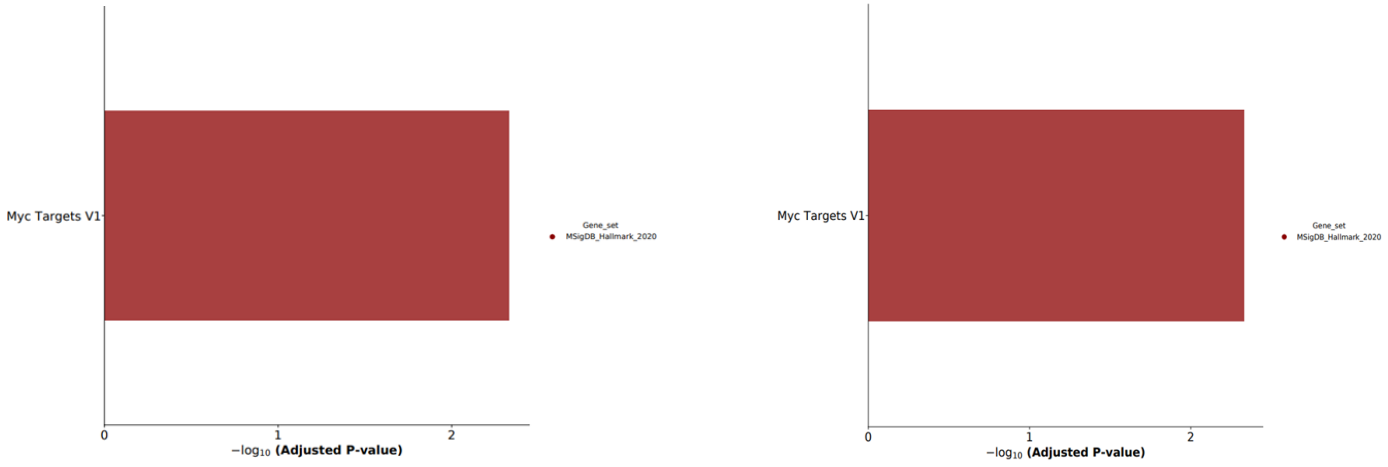


Fig. 3.4.1 MSigDB hallmarks for the original expression data

3.4.2. Augmented expression data



A. **B.**

Fig. 3.4.2 MSigDB hallmarks for the augmented file using the A. BN model and B. CTGAN model

The enrichment of the Myc-targets hallmark in the MSigDB (Molecular Signature Database) plots shown above indicates a strong association between sepsis with diabetes and Myc targets. This hallmark was significantly enriched and consistently expressed in all the synthetic data generated using the Bayesian Network (BN) and Conditional Tabular Generative Adversarial Network (CTGAN) models. However, the Gaussian Mixture Model (GMM) failed to result in any MSigDB hallmarks for the augmented expression dataset. Therefore, it can be concluded that the Myc-targets hallmark may be a critical pathway associated with sepsis with diabetes.

3.5. Enrichment analysis

Our study found a high link between secretory vesicles, secretory lumen, and the endomembrane system found a high link between secretory vesicles, secretory lumen, and the endomembrane system in our study, which is consistent with previous research on the pathophysiology of sepsis. Using machine learning (ML) models to produce synthetic gene expression data, we uncovered the above mentioned pathways. We uncovered the above-mentioned pathways and other genes and pathways related to the condition. This shows that ML models could be used to generate gene expression data for signature identification. We performed enrichment analysis on the original data and data created using the GMM, BN, and CTGAN models to compare augmentation approaches for sepsis with diabetes biomarker identification in a sizeable multicentric dataset.

For the CTGAN approach:

- **Gene Set Enrichment Analysis Plot:**
  We used the GSEA plot to analyse over-representation. The CTGAN model's augmented expression data was utilised to identify differentially expressed genes, which were then sorted based on their log fold change (log Fc) values, arranging the genes in the gene list according to their upregulation and downregulation. The sorted gene list was used in the GSEA analysis to obtain a running enrichment score that associated the gene order with the examined class. The resulting plot revealed a number of strongly upregulated gene sets related to the illness condition, indicating that defence mechanisms were activated. The running enrichment score is shown by the green line in the plot.

- **Dot Plot:**
  Based on their significance levels, the dot plot depicts the over- and under-representation of gene sets within their respective biological pathways, molecular functions, and cellular

components. The dot plot analysis revealed that the highly expressed genes largely activated cellular components such as secretory vesicles and secretory granules. Furthermore, the only suppressed route was connected to the regulation of stem cell development, which had fewer gene sets than the activated circumstances.

- **Ridge Plot:**
  The ridge plot depicts the probability distribution of enrichment scores over several pathways. Its goal is to find pathways that are strongly related with elevated gene sets. As the plot shows, all of the routes illustrated in this ridge map are linked to the gene sets in the upregulated sorted list. Red highlights highly significant gene sets.

- **Category Net Plot:**
  The Category Net plot is a useful tool for visualising the network of gene and biological concept (GO words) linkages, making it easier to identify genes participating in enriched pathways and those that may belong to numerous annotation categories. In this investigation, we displayed the top three highly significant and enriched GO terms in the network plot along with their associated gene sets. All genes implicated in the key pathways were found to be increased, implying that the upregulation of these genes may play a role in the overexpression of defence mechanisms upon the onset of this disease condition.

We gained significant insights into the effects of several augmentation approaches on sepsis with diabetes biomarker identification in a large multicentric dataset by doing these studies on both the original and generated data. CTGAN model results outperformed other model results. Additionally, we have provided the four aforementioned plot results for the Original, GMM, BN, and CTGAN data.

I. Original Data

A. GSEA (Gene Set Enrichment Analysis)

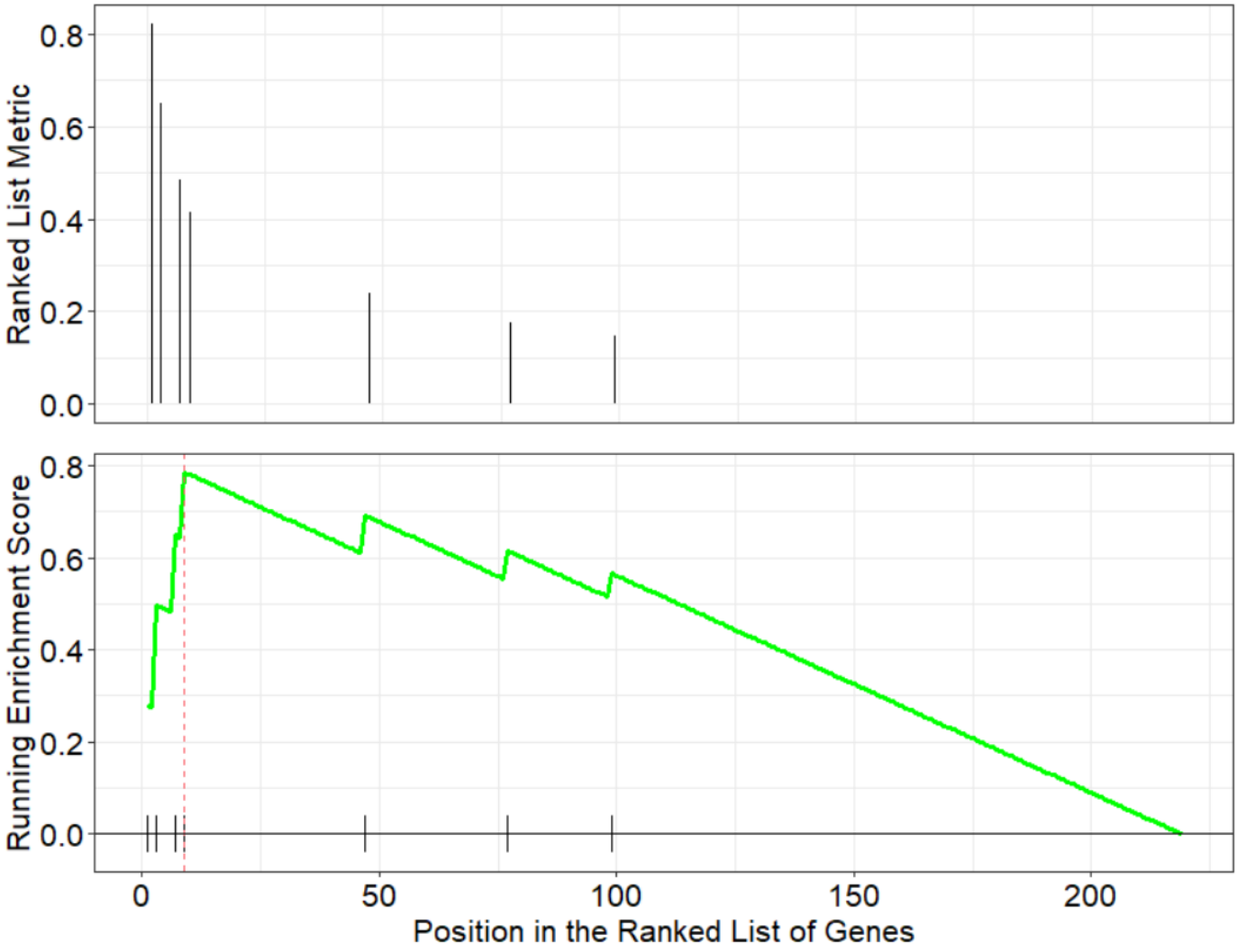


Fig. 3.5.I(A) GSEA plot for the GO enrichment analysis for Original Data

B. Dot plot

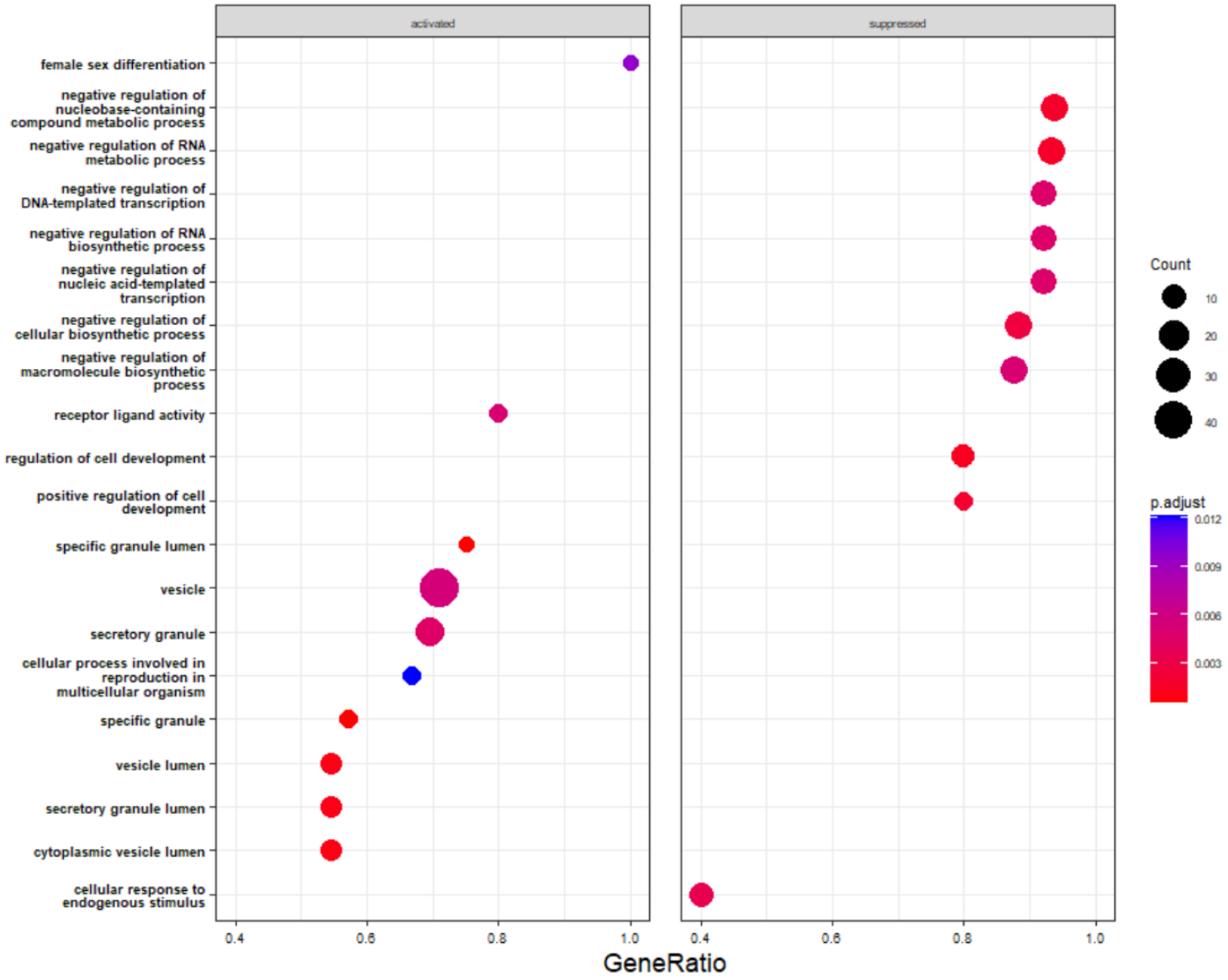


Fig. 3.5.I(B) Dot plot for the enrichment analysis of Original Data

C. Ridge plot

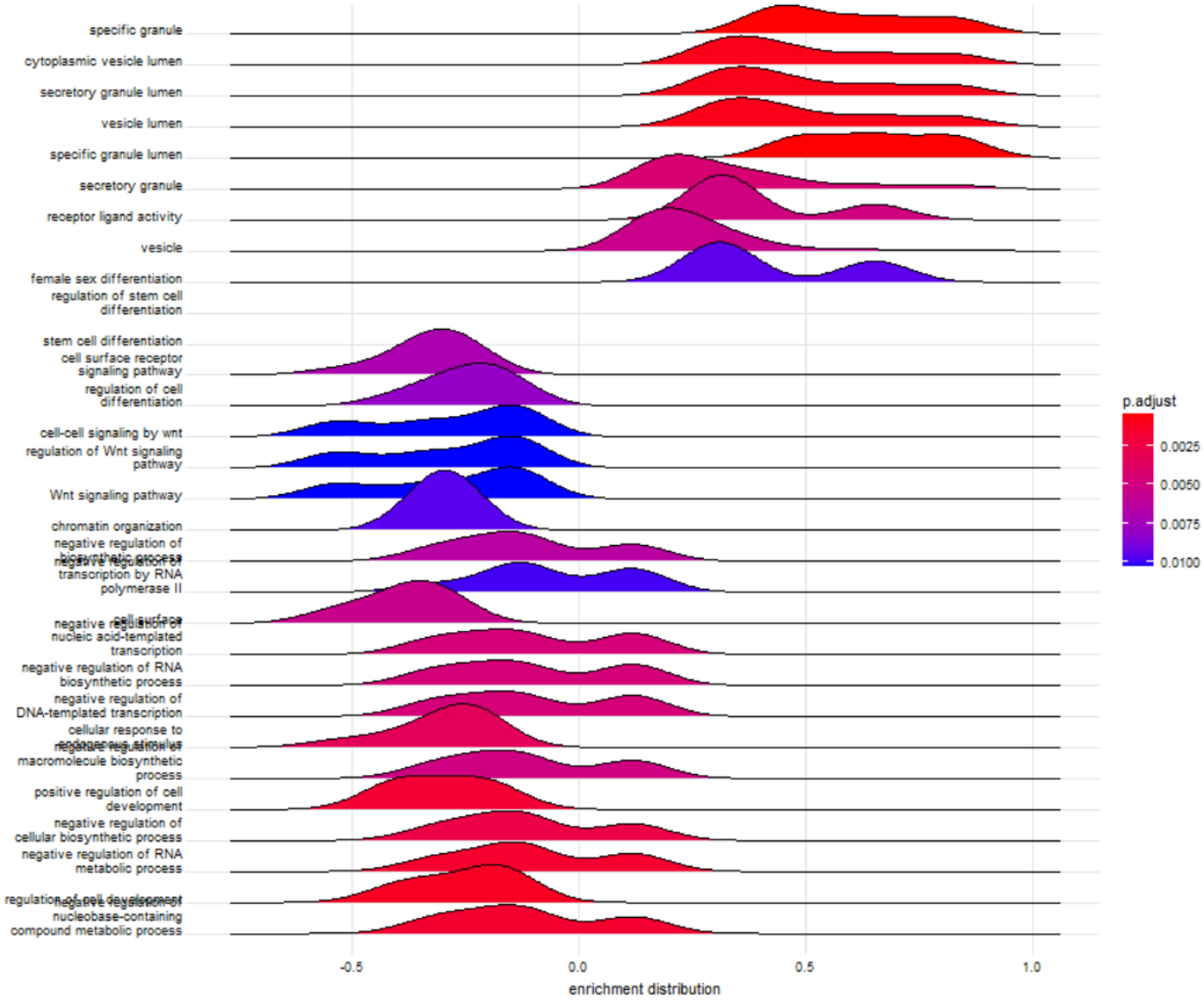


Fig 3.5.I(C) Ridge plot for the enrichment analysis of Original Data

D. Category Net plot

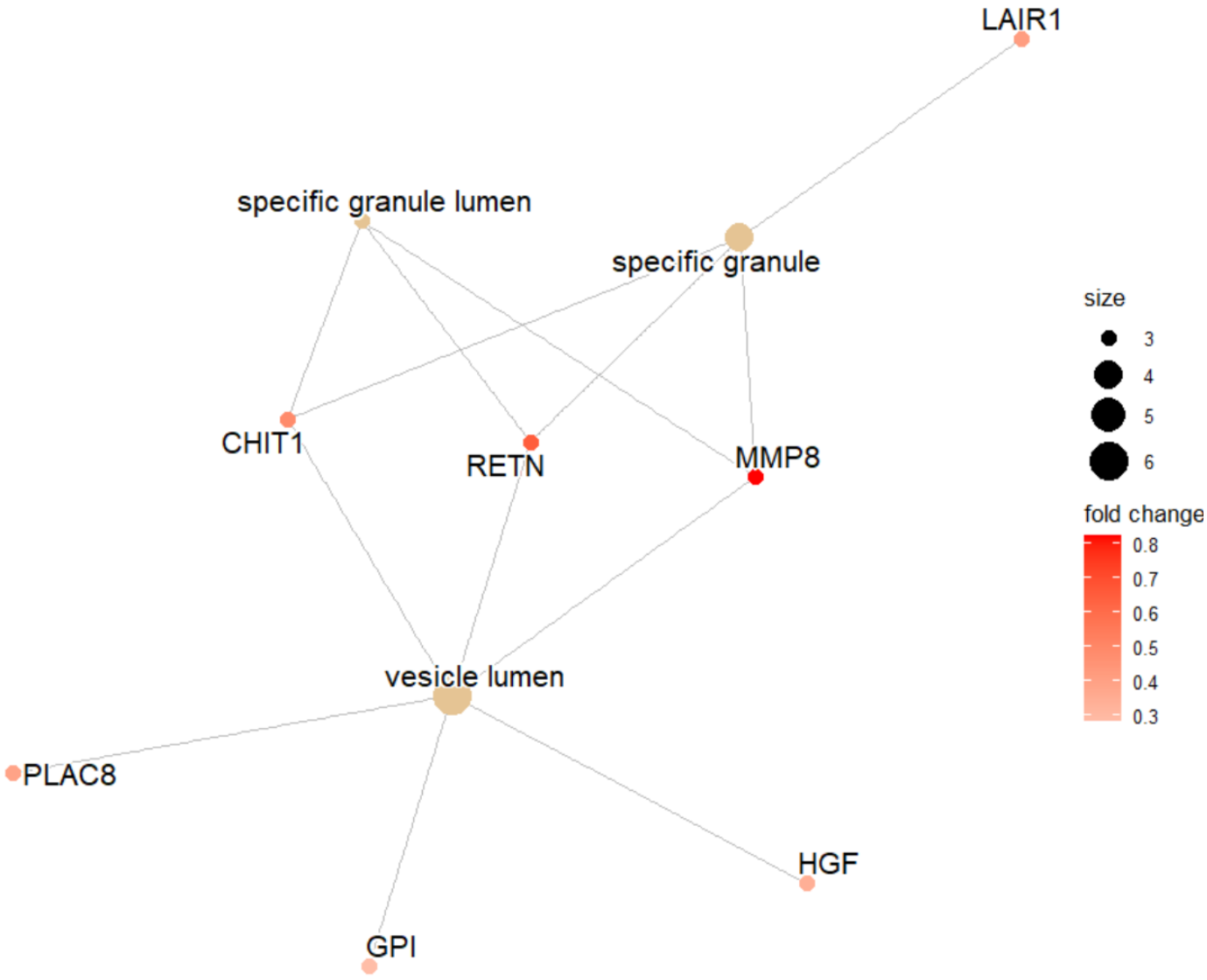


Fig. 3.5.III(D) Category net plot for the enrichment analysis of Original Data

II. GMM

A. GSEA(Gene Set Enrichment Analysis)

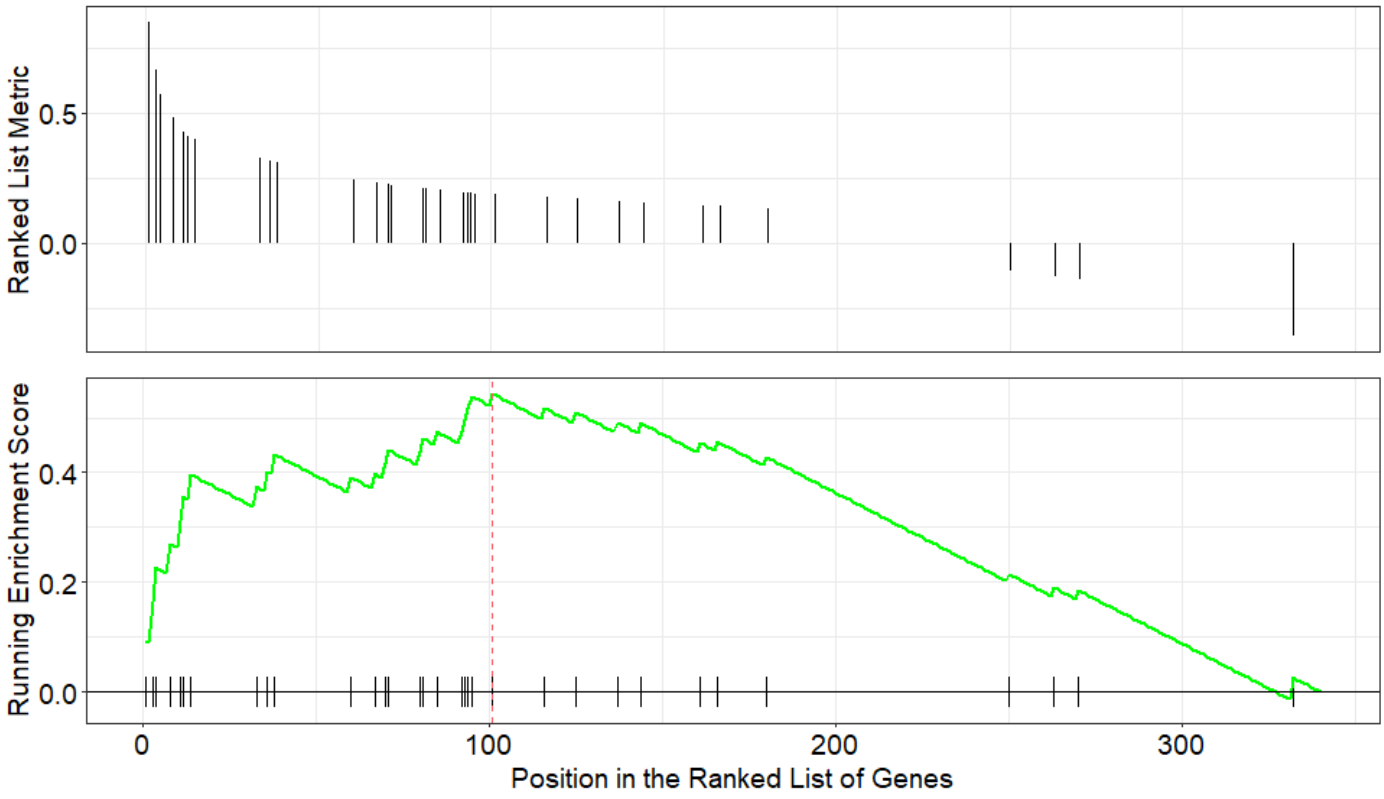


Fig. 3.5.II(A) GSEA plot for the GO enrichment analysis for GMM Data

B. Dot plot

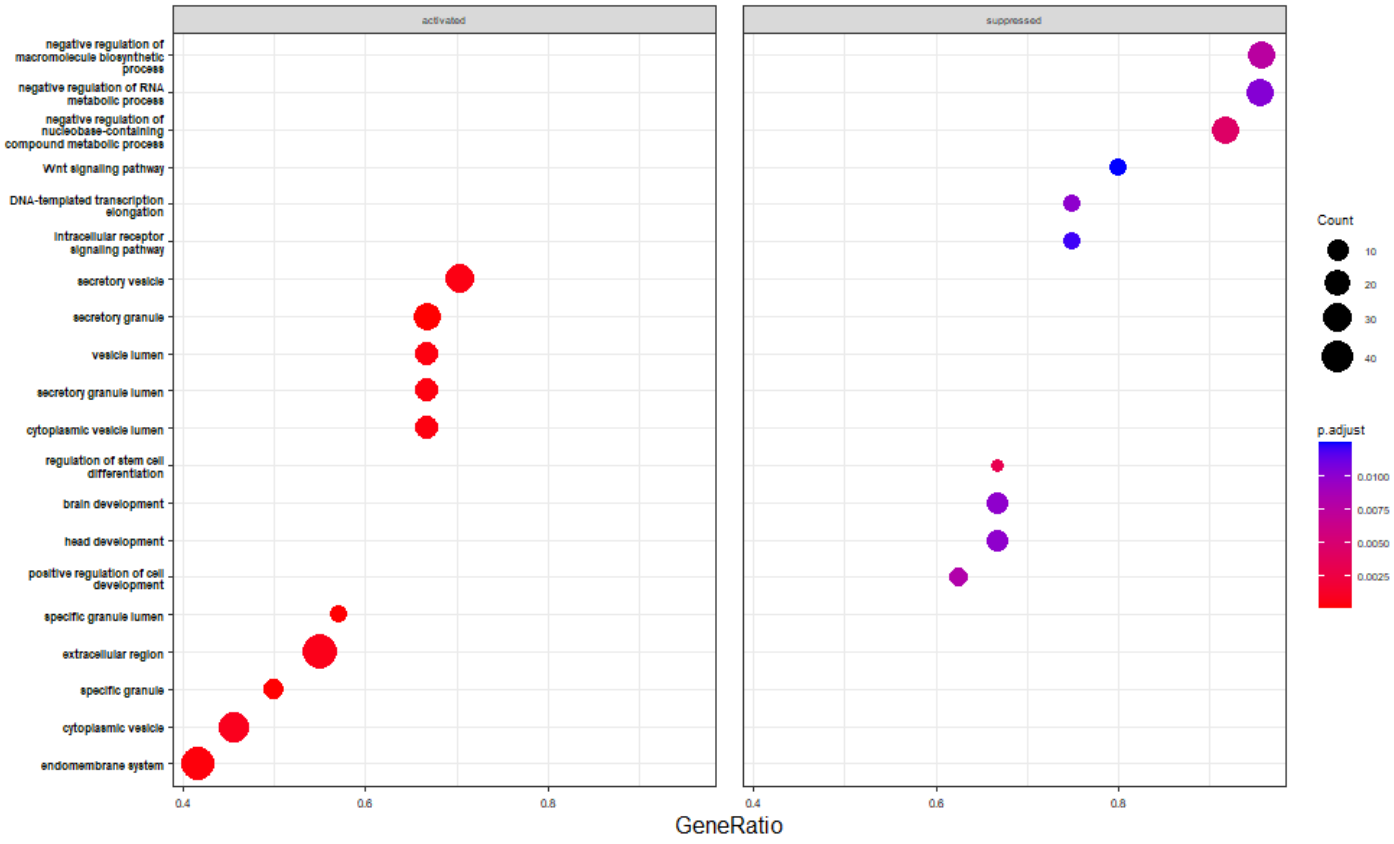


Fig. 3.5.II(B) Dot plot for the enrichment analysis of GMM Data

C. Ridge plot

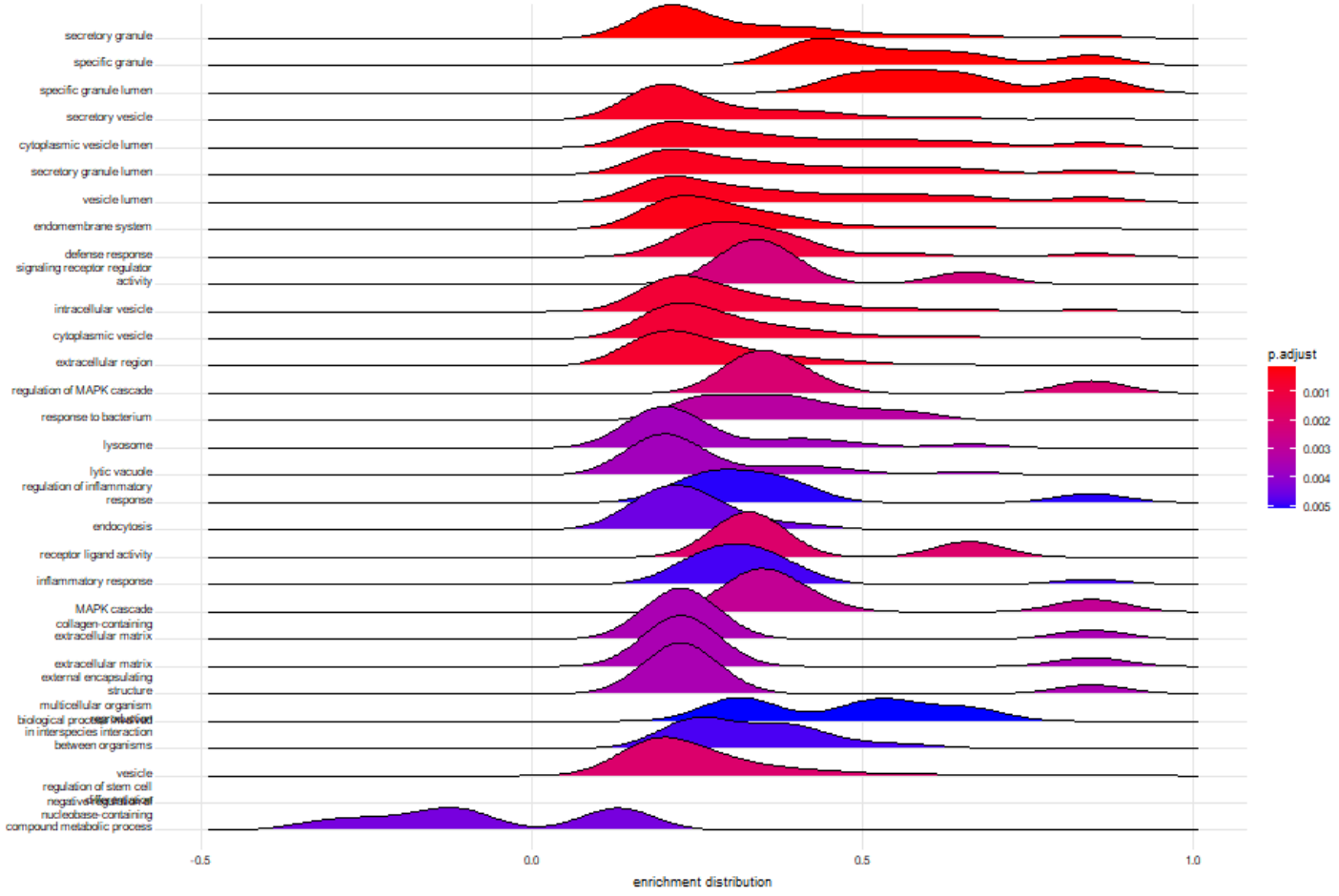


Fig 3.5.II(C) Ridge plot for the enrichment analysis of GMM Data

D. Category Net plot

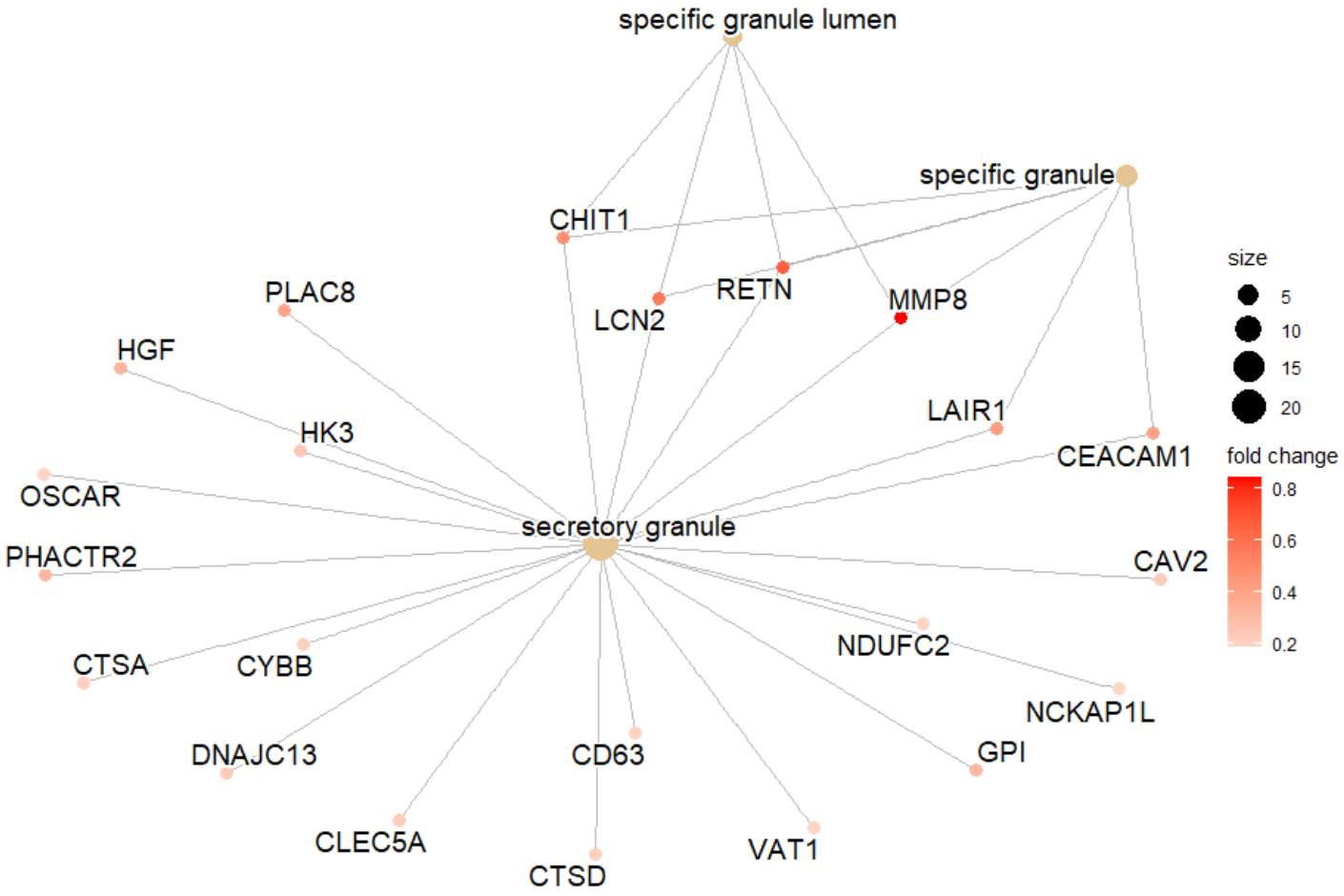


Fig. 3.5.II(D) Category net plot for the enrichment analysis of GMM Data

III. BN

A. GSEA(Gene Set Enrichment Analysis)

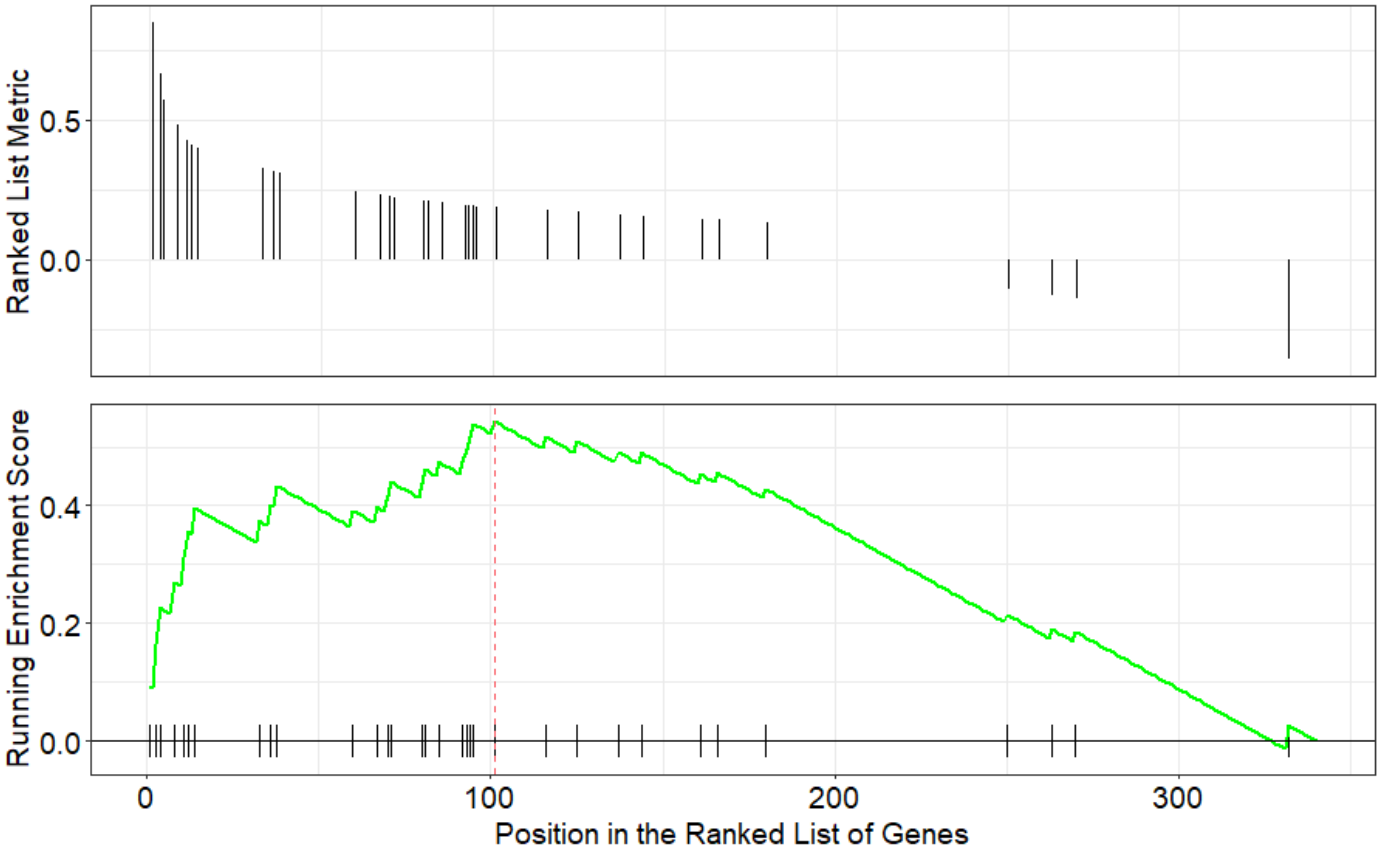


Fig. 3.5.III(A) GSEA plot for the GO enrichment analysis for BN Data

B. Dot plot

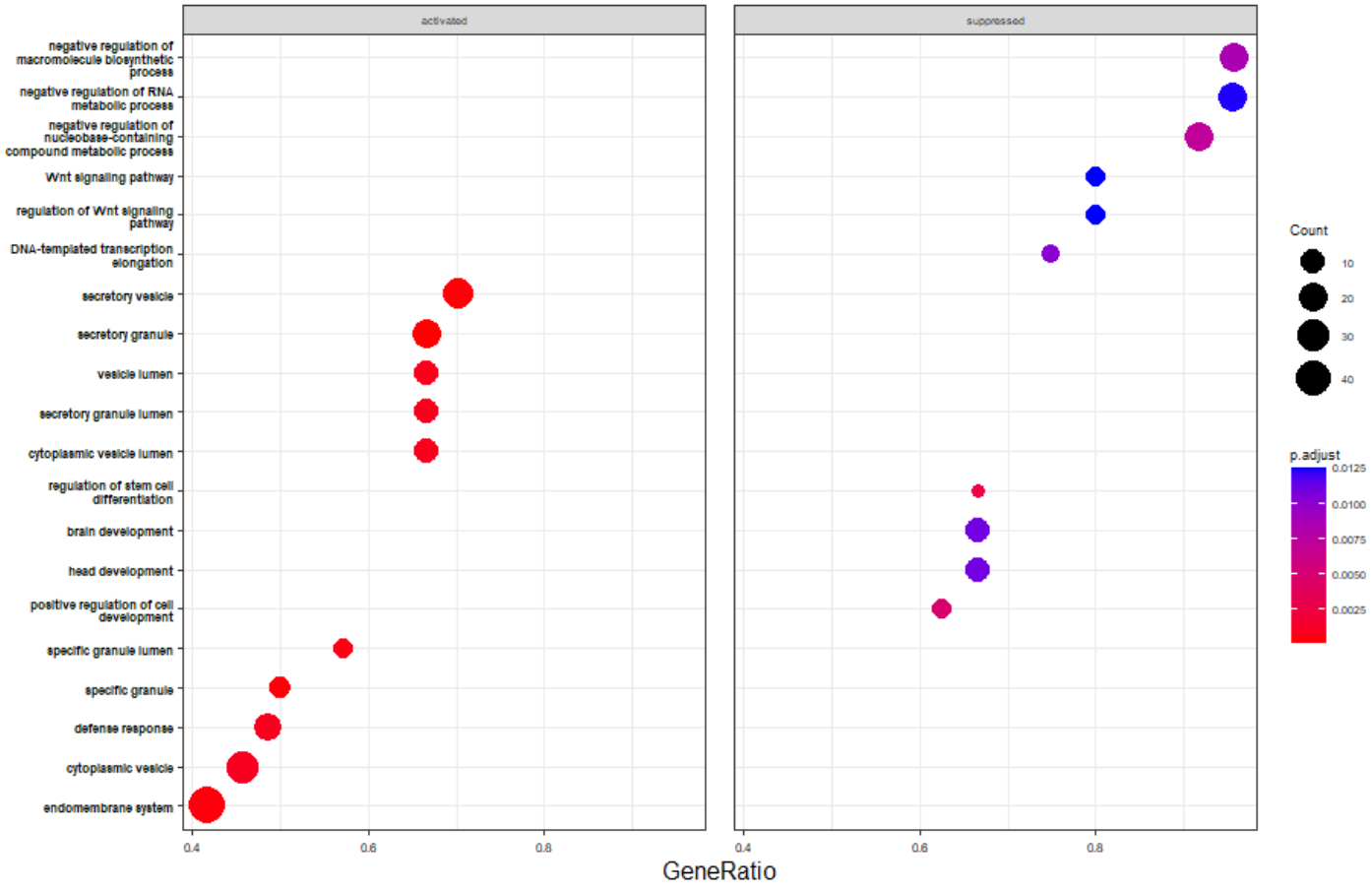


Fig. 3.5.III(B) Dot plot for the enrichment analysis of BN Data

C. Ridge plot

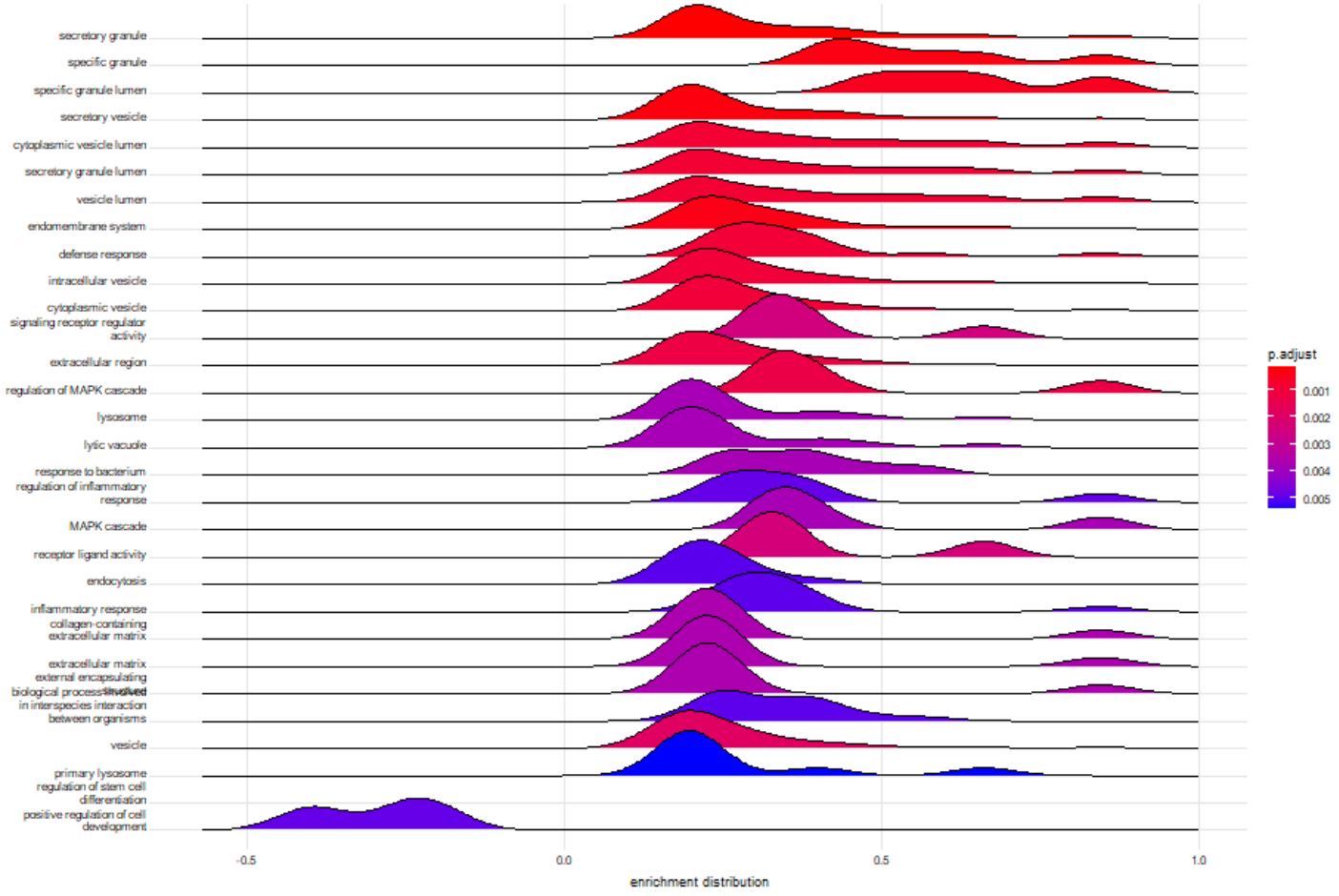


Fig 3.5.III(C) Ridge plot for the enrichment analysis of BN Data

D. Category Net plot

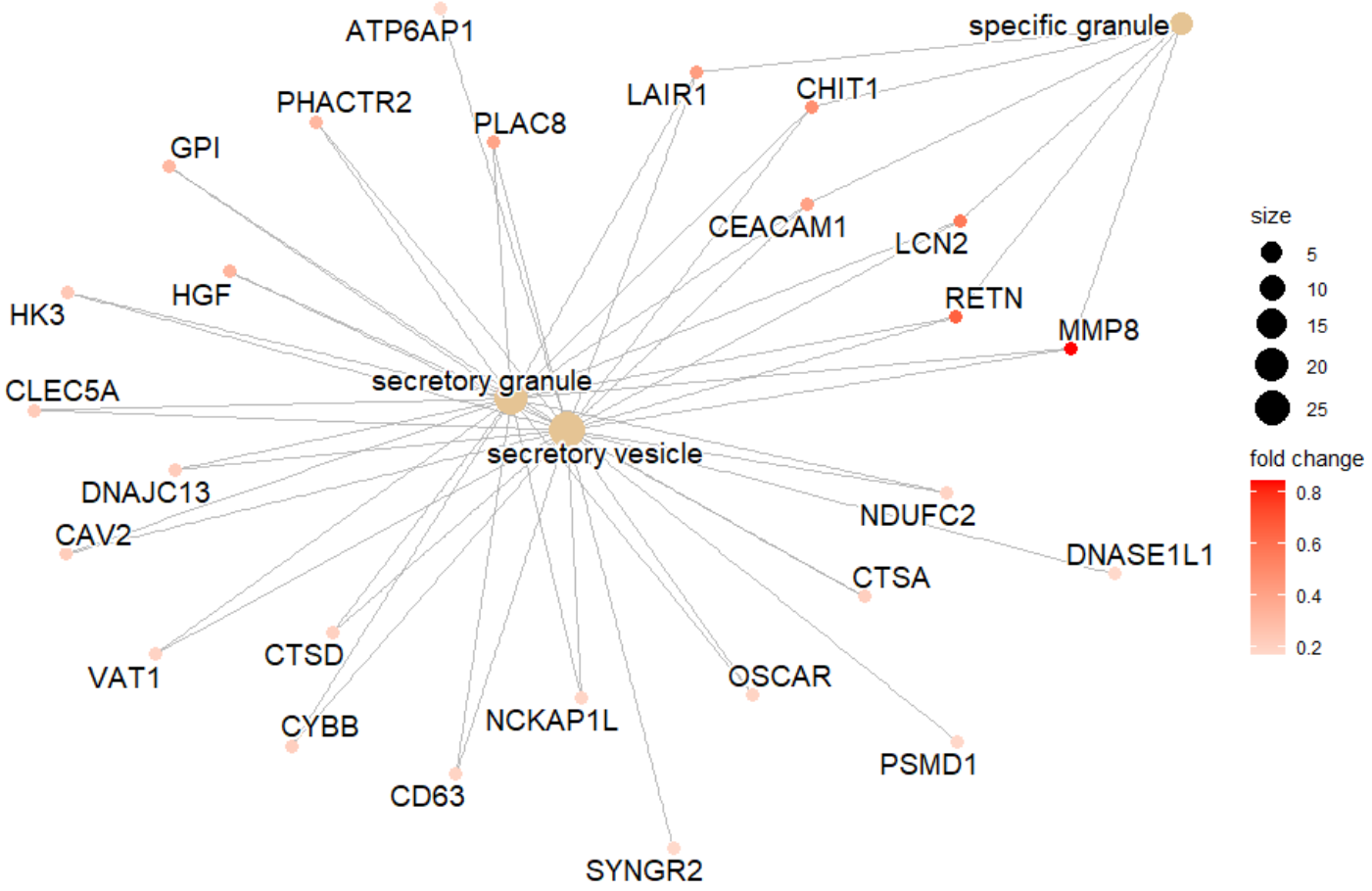


Fig. 3.5.III(D) Category net plot for the enrichment analysis of BN Data

IV. CTGAN

A. GSEA(Gene Set Enrichment Analysis)

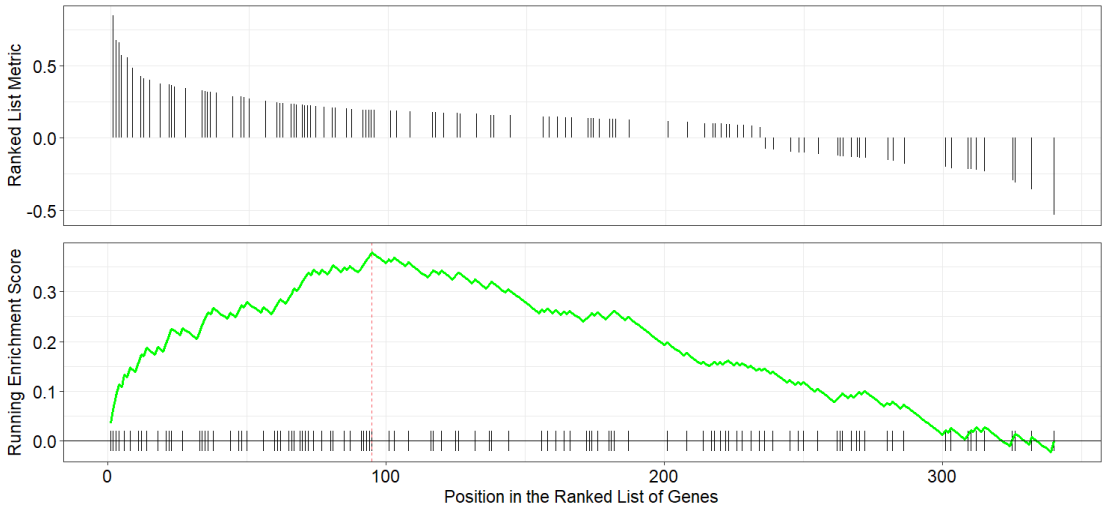


Fig. 3.5.IV(A) GSEA plot for the GO enrichment analysis for CTGAN Data

The over-representation analysis of GO terms was performed using the GSEA plot. The augmented expression data generated through the CTGAN model was utilised to identify differentially expressed genes, which were then ranked based on their log Fc value, sorting the genes in the order of their upregulation and downregulation in the gene list. The analysis of GSEA was carried out using the sorted gene list to produce the running enrichment score that correlated the order of the gene with the studied class. The resulting plot depicted many highly upregulated gene sets in the concerned disease condition, indicating the involvement of highly activated defence mechanisms. The green line in the plot represents the running enrichment score.

B. Dot plot

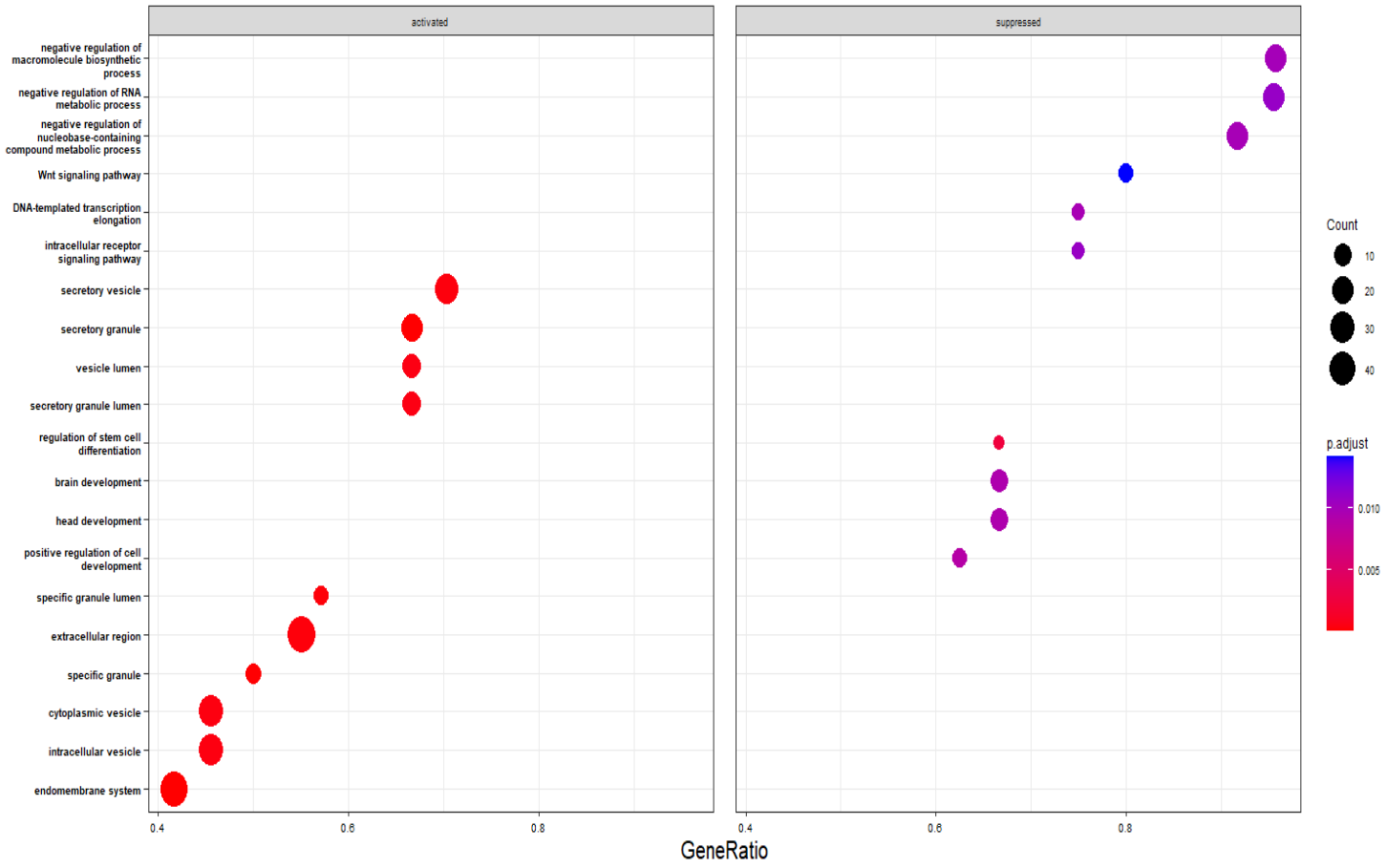


Fig. 3.5.IV(B) Dot plot for the enrichment analysis of CTGAN Data

The dot plot displays the over and under-representation of gene sets in their respective biological pathways, molecular functions, and cellular components based on their significance level. The plot analysis revealed that the highly significant expressed genes primarily activated cellular components such as secretory vesicles and granules. Additionally, the only pathway that demonstrated significant suppression was regulating stem cell differentiation, which was smaller than the gene sets involved in the activated conditions.

C. Ridge plot

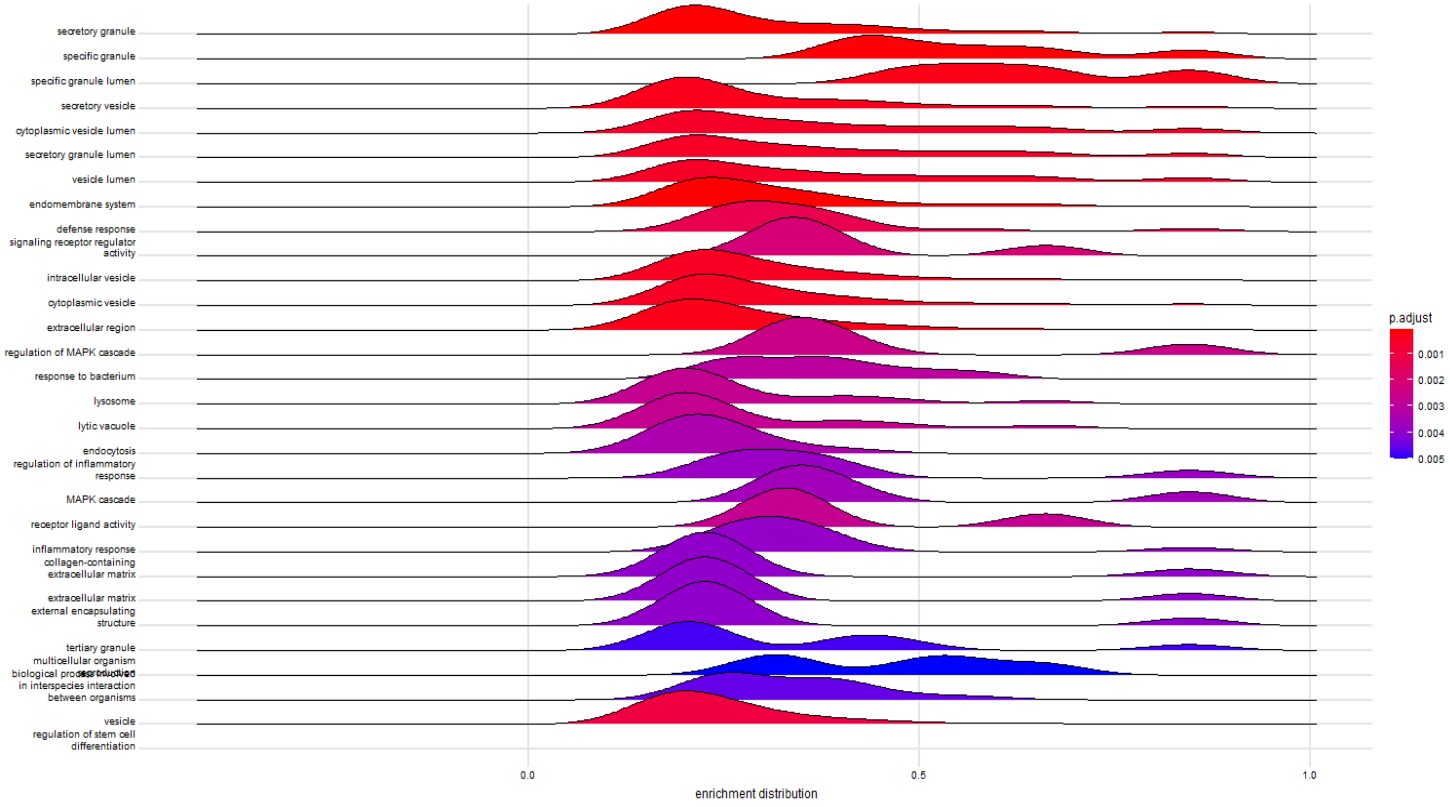


Fig 3.5.IV(C) Ridge plot for the enrichment analysis of CTGAN Data

This ridge plot depicts the probability distribution of enrichment scores across various pathways. The aim is to identify the pathways that are significantly associated with the upregulated gene sets. As evident from the plot, all the pathways discussed in this ridge plot are related to the gene sets in the upregulated ranked list. The highly significant gene sets are highlighted in red colour.

D. Category Net plot

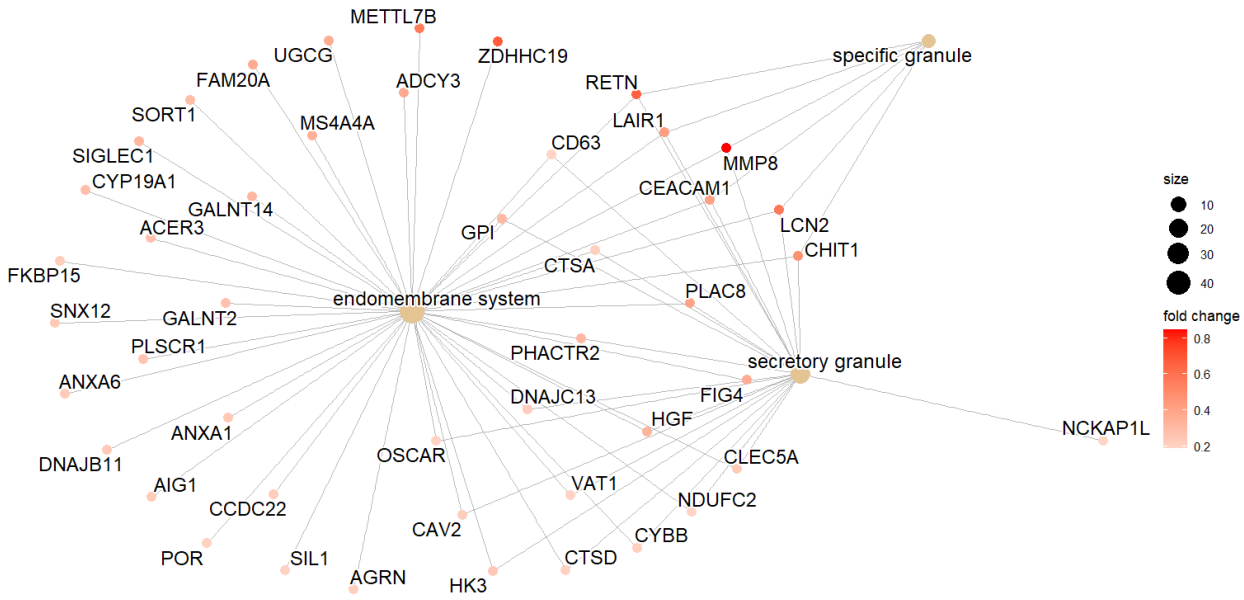


Fig. 3.5.IV(D) Category net plot for the enrichment analysis of CTGAN Data

The Category net plot is a useful tool for visualising gene and biological concept (GO terms) linkages as a network, which facilitates the identification of genes involved in enriched pathways and those that may belong to multiple annotation categories. In this study, the top three highly significant and enriched GO terms with the associated gene sets were depicted in the network plot. All genes involved in the significant pathways were found to be upregulated, indicating that upon the occurrence of this diseased condition, upregulation of these genes may play a role in the overexpression of the defense mechanism.

Chapter 4 Conclusion & Future Scope

4.1 Conclusion

In this study, we investigated the potential of machine learning-based data augmentation approaches for biomarker discovery in a multicenter dataset of patients with sepsis and diabetes. We compared the efficacy of three different approaches: Gaussian Mixture Model (GMM), Bayesian Network (BN), and Conditional Tabular Generative Adversarial Network (CTGAN), for augmenting microarray expression data in the XpressionSuite tool developed by the TavLab at IIIT-Delhi. Firstly, we identified 219 differentially expressed genes (DEGs) that met the selection criterion, with 137 upregulated and 82 downregulated genes. We performed differential gene expression analysis (DGEA) on the original expression data and the augmented expression data generated using the GMM, BN, and CTGAN models. The volcano plots revealed that the GMM and BN models did not achieve significant results in terms of DEGs. However, the CTGAN-generated data exhibited higher statistical significance, identifying 80 DEGs with the higher power, making it the preferred choice for further analysis.

We further explored the hallmarks of the original and augmented expression data. The Myc targets hallmark was consistently enriched in all synthetic data generated using the BN and CTGAN models, suggesting its association with sepsis in patients with diabetes. On the other hand, the GMM model did not result in any significant MSigDB hallmark. Enrichment analysis, including Gene Set Enrichment Analysis (GSEA) and dot plot analysis, provided insights into the biological pathways and cellular components involved in sepsis in patients with diabetes. The augmented expression data generated through the CTGAN model revealed highly activated defense mechanisms, primarily involving cellular components such as secretory vesicles and secretory granules. Additionally, the pathway regulating stem cell differentiation demonstrated significant suppression. The ridge plot analysis identified the pathways significantly associated with upregulated gene sets, all of which were related to the defence mechanism. Finally, the Category Net plot visualized the network of gene and Gene Ontology (GO) term linkages, highlighting the upregulation of genes involved in enriched pathways. In conclusion, this study demonstrates the effectiveness of data augmentation in enhancing the statistical power and biological relevance of gene expression data analysis. The CTGAN-based data augmentation approach outperformed the GMM and BN models in terms of identifying DEGs and their biological significance. The identified Myc targets hallmark and the involvement of various cellular components and pathways provide valuable insights into the pathogenesis of sepsis in patients with diabetes.

These findings have implications for biomarker discovery and the development of therapeutic interventions for sepsis in diabetic patients. Moreover, this study highlights the potential of data augmentation in overcoming challenges and improving the analysis of high-throughput data in the field of omics. Overall, this research advances our understanding of data augmentation approaches and their application in biomarker discovery. It opens up opportunities for further exploration, including developing novel augmentation techniques, investigating downstream analysis impact, addressing ethical considerations, and integrating domain knowledge. By pursuing these avenues, we can enhance the field of data augmentation and its utility in various biological contexts, ultimately advancing clinical research and improving patient outcomes.

4.2 Future Scope

Data augmentation has emerged as a valuable solution to the problem highlighted in the discussion.The advent of high-throughput data collection has revolutionized data generation and evaluation, leading to significant advancements in various scientific domains. The utilization of machine learning-based augmentation techniques has opened up new possibilities in the realm of biological data processing. The research conducted in this dissertation has successfully demonstrated the effectiveness of data augmentation in enhancing the statistical power and biological relevance of the studied condition using real-world data. However, there are several areas that warrant further exploration and development to enrich the current body of knowledge. Firstly, expanding the repertoire of data augmentation techniques beyond the ones explored in this study may uncover novel approaches that can better capture the complexities of biological systems. Exploring alternative machine learning models or hybrid approaches could potentially yield enhanced performance and broader applicability. Additionally, investigating the impact of different augmentation strategies on downstream analyses and interpretation is an avenue for future research. Assessing the robustness of augmented data across multiple analytical frameworks and exploring the generalizability of findings to diverse populations and experimental conditions will further validate the utility of data augmentation.

Furthermore, the ethical implications and potential biases associated with data augmentation techniques should be carefully examined. Developing guidelines and standards for responsible use of augmented data, including measures to mitigate unintended biases and preserve privacy, will be crucial to ensure the integrity and fairness of research outcomes. Lastly, the integration of domain knowledge and prior biological information in the augmentation process holds promise for further enhancing the biological relevance and interpretability of the augmented data. Incorporating domain-specific constraints and leveraging existing biological networks or pathways can guide the generation of augmented samples that align with known biological mechanisms.

In conclusion, while the research presented in this dissertation establishes the effectiveness of data augmentation, there are exciting opportunities for future investigations. Exploring novel augmentation techniques, investigating the impact on downstream analyses, addressing ethical considerations, and integrating domain knowledge will collectively advance the field of data augmentation in omics research. These endeavours will contribute to advancing scientific understanding and enable more robust and insightful investigations in various biological contexts.

Bibliography

[1] Brendon P Scicluna , Peter M C Klein Klouwenberg, et al. A Molecular Biomarker to Diagnose Community-acquired Pneumonia on Intensive Care Unit Admission. Proc Natl Acad Sci USA. 2015 Jun 30;112(26):8136-41.

[2] Scicluna BP, van Vught LA, Zwinderman AH, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. Lancet Respir Med. 2017 Mar;5(3):235-245.

[3] Determann RM, van Vugt SF, Roelofs JJ, et al. Matrix metalloproteinase-8: a useful biomarker to refine the diagnosis of community-acquired pneumonia upon intensive care unit admission? Crit Care. 2015 Jul 2;19:233.

[4] Presneill JJ, Waring PM, Layton JE, et al. Association between age and the host response in critically ill patients with sepsis. Crit Care. 2020 Jun 17;24(1):327.

[5] Liberzon, A., et al. (2011). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Systems, 1(6), 417-425.

[6] G.J. McLachlan, R.W. Bean, L. Ben-Tovim Jones, A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. 2006 April 21

[7] Mootha, V. K., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 34(3), 267-273.

[8] Ritchie, M. E., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47.

[9] Subramanian, A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550.

[10] B. Klaus, An end to end workflow for differential gene expression using Affymetrix microarrays. 2018 Jul 3