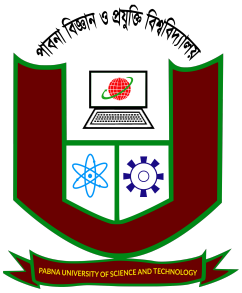
A Hybrid Methodology for Recognizing Bangla Sign Language Using A Deep Transfer Learning Model in Combination with A Machine Learning Classifier



Department of Computer Science and Engineering
Pabna University of Science and Technology, Pabna-6600

Course Title: Thesis
Course Code: CSE 4100 and CSE 4200

A thesis has been submitted to the Department of Computer Science and Engineering for the partial fulfillment of the requirement of B.Sc in Engineering in Computer Science and Engineering

Submitted By:
The Examinee of B.Sc Engineering Final Examination-2022
Mst. Shaima Ashlam Chaity
Roll Number: 200121
Registration Number: 101832
Session: 2019-20

Supervised By:
S. M. Hasan Sazzad Iqbal
Associate Professor
Department of Computer Science and Engineering
Pabna University of Science and Technology

Laboratory
Advanced Computer Lab, Department of Computer Science and Engineering
Pabna University of Science and Technology

August, 2024

DECLARATION

I, **Shaima Aslam Chaity**, hereby declare that the work presented in this thesis is entitled **"Genetic and Molecular Perspective on Endocrine and Cardiovascular Complications in Beta-Thalassemia Patients"**, is the outcome of my own research and effort carried out under the supervision of **S. M. Hasan Sazzad Iqbal**, Department of Computer Science and Engineering, Pabna University of Science and Technology (PUST), Pabna.

I also declare that this work is the best of my knowledge and does not contain any material that previously published or written by another person, nor it has been submitted in whole or in part, for the award of any degree or diploma at any other university or institution. All materials or content taken from other sources has been appropriately referenced in this thesis

This thesis reflects my own findings, opinions and conclusions and do not necessarily represent the views of Pabna University of Science and Technology or any other institution.

Signature of the Examinee

CERTIFICATION

I am happy to certify that Mst. Shaima Ashlam Chaity, Roll Number: 190116, Registration Number: 101832, Session: 2018-2019, has completed a thesis work that enabled **"A Hybrid Methodology for Recognizing Bangla Sign Language Using A Deep Transfer Learning Model in Combination with A Machine Learning Classifier"** under my supervision to fulfill the requirements of the thesis course. This thesis was completed for a year at the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh.

According to my knowledge, this thesis paper has not been submitted elsewhere or replicated by another thesis paper before being submitted to the department.

S. M. Hasan Sazzad Iqbal

Associate Professor,

Department of Computer Science and Engineering

Pabna University of Science and Technology, Pabna, Bangladesh

ACKNOWLEDGEMENT

My deepest gratitude to the Allah for granting me the strength and guidance to complete this thesis successfully.

I am sincerely thankful to my supervisor, **S. M. Hasan Sazzad Iqbal**, Associate Professor in Department of Computer Science and Engineering at Pabna University of Science and Technology (PUST), for his continuous support, insightful advice and valuable feedback throughout the completion of this research. His insightful suggestions, encouragement, and guidance have been instrumental in shaping this work from its beginning to completion.

I would also like to extend my appreciation to all the respected teachers of the Department of Computer Science and Engineering, PUST, whose teachings and academic guidance have laid the foundation for my research knowledge and skills. My heartfelt thanks go to my family, especially my parents as well as my friends and well-wishers for their support, motivation, and cooperation during this work.

I express my genuine thanks to everyone who contributed by any means to the successful ending of this thesis.

August, 2025

Author

ABSTRACT

Beta-thalassemia (BT), a genetic blood anomaly brought on by abnormalities in the HBB gene that occurs when beta-globin chain production is reduced or absent in blood. It is a major worldwide health issue because of its complicated clinical management and lifetime reliance on blood transfusions. Some recent evidence highlights that beta-thalassemia is associated with several comorbidities, including endocrine diseases involving polycystic ovarian syndrome (PCOS), hypothyroidism, hypogonadism and type 2 diabetes (T2D) and cardiovascular diseases involving arrhythmogenic cardiomyopathy (ACM) and arrhythmia. These comorbidities may share common molecular mechanisms with beta-thalassemia that potentially exacerbate disease severity and treatment complications. In this study, a computational approach was applied to evaluate the genetic relationships between beta-thalassemia and its associated comorbidities using microarray and mRNA datasets that are publicly available in NCBI. Genetic profiling was constructed to identify the common matching genes and built disease-gene networks (DGNs) of matching genes. It also visualized a heatmap to present patterns of gene expressions. We also explored multiple bioinformatics analysis including pathways, gene ontology, protein-protein interaction (PPI) and protein-drug interaction (PDI) that strongly indicate their correlation. A validation network was created to verify our selected comorbidity and then phylogenetic analysis was performed for all diseases to determine their evolutionary relationships. This study found that beta-thalassemia shares 13, 165, 13, 14, 11 and 44 significantly expressed genes with hypothyroidism, hypogonadism, PCOS, T2D, ACM and arrhythmia respectively. The outcomes of this study may help in integrative medical approaches and enhance a significant understanding of genetic and molecular structure of comorbidities in beta-thalassemia by providing valuable insights.

Keywords: Beta-thalassemia, comorbidities, genetic profiling, endocrine diseases, cardiac diseases, arrhythmogenic cardiomyopathy, pathway analysis, ontology, phylogenetic analysis.

TABLE OF CONTENTS

ABBREVIATIONS AND ACRONYMS

The following abbreviations and acronyms are used in this thesis:

- **KU-BdSL**: Khulna University Bengali Sign Language dataset
- **BdSL**: Bangla Sign Language
- **WHO**: World Health Organization
- **SHAP**: SHapley Additive exPlanation
- **SE**: Squeeze Excitation
- **ML**: Machine Learning
- **DL**: Deep Learning
- **CNN**: Convolution Neural Network
- **ResNet**: Residual Neural Network
- **RGB**: Red, Green, and Blue
- **MSLD**: Multi-scale Sign Language Dataset
- **USLD**: Uni-scale Sign Language Dataset
- **AMSLD**: Annotated Multi-scale Sign Language Dataset
- **ROC**: Receiver Operating Characteristics

LIST OF TABLES

TABLE **Page**

LIST OF FIGURES

FIGURE **Page**

CHAPTER 1

INTRODUCTION

This chapter presents an overview of our research thesis, organized into seven sub-chapters for clarity. In section 1.1, we introduce the topic, focusing on the interplay between diabetes and cancers (breast, ovarian, cervical, and gastric). Section 1.2 outlines the motivation, emphasizing the need to understand shared biological pathways. Section 1.3 defines the problem statement, addressing the increased cancer risk in diabetic patients. Section 1.4 specifies the objectives, including identifying common differentially expressed genes (DEGs) and analyzing molecular interactions. Section 1.5 highlights the research outcomes and their potential impact on early diagnosis and treatment. Section 1.6 details the thesis structure, and section 1.7 provides a discussion on the significance and relevance of our study.

1.1 Introduction

Beta-thalassemia is a complex hematological disorder inherited from parents to offspring. It is brought by mutations in the HBB gene and caused by the body,Äôs inability to produce enough beta-globin protein in red blood cells, which results in chronic anemia, iron overload and several medical complications. Beta-thalassemia patients require lifetime medication and often need regular blood transfusions to manage chronic anemia [1].
Some compelling research evidence shows it contributes remarkably to the development of cardiovascular diseases like ACM and arrhythmia, as well as endocrine diseases like T2D, PCOS, hypothyroidism and hypogonadism. Through gene expression analysis, this study aims to explore the link between beta-thalassemia and its associated comorbidities. Many studies show that iron overload from blood transfusions in beta-thalassemia interrupts the function of endocrine glands and reduces hormone synthesis and secretion [1]. Also, chronic anemia and iron overload from regular blood transfusion cause cardiac abnormalities [2].
To explore genetic links more deeply between beta-thalassemia and comorbidities, this study identified differentially expressed genes (DEGs) from RNA-sequence and microarray datasets. This study constructed disease-gene association networks (DGNs) by mapping the genes of up-regulated and down-regulated and then visualizing those matching genes using a heatmap [3]. Pathway analysis is conducted with graphical representation and gene ontology (GO) analysis to show molecular functions [3]. This work built the PPI and PDI networks to investigate the functional connectivity between matching genes. A phylogenetic analysis is performed to observe the interrelation of all seven diseases. Finally, this thesis validated the results using standard biomedical libraries like OMIM (Online Mendelian Inheritance in Man) and databases like dbGaP. Our proposed work has been undertaken to achieve a more comprehensive genetic understanding of beta-thalassemia with its comorbidities, which will improve the clinical approaches for its diagnosis, management and treatment.

1.2 General Background

Beta-thalassemia is an inherited blood disorder transferred from parents in a received patterns that is characterized by absence of beta-globin production, which resulting in anemia, excessive iron accumulation and various complications affecting multiple organ systems. Over time, excessive iron deposition in endocrine glands contributes to a spectrum of complications collectively known as thalassemic endocrine disease (TED). These include hypogonadism, hypothyroidism, diabetes, adrenal dysfunction,

and reduced bone mineral density [11]. It also resulting in various cardiovascular issues such as vasculopathies, cardiomyopathy, hypertension and arrhythmias. These complications significantly contribute to the risk of sudden cardiac death [12].
Although from the current existing research much is known about beta-thalassemia and its complications, the genetic and molecular links between BT and its endocrine and cardiovascular comorbidities remain unclear. Understanding of the shared genes, pathways, PPI and PDI in still limited.
This study aims to fill these gaps by analyzing gene expression data to uncover the genetic connections between BT and its related complications. This findings could improve the knowledge of disease mechanisms and help to create better diagnostic tools and therapies.

1.3 Motivation

As BT is a chronic inherited disorder that often leads to serious endocrine and cardiovascular complications. Although clinical management has improved day by day but the genetic and molecular perspectives leading these comorbidities are not still clearly understandable. The lack of knowing about internal genetic and molecular relationships limits the ability to design appropriate diagnostic ways and proper treatments.
By analyzing deeply gene expression datasets and using advanced bioinformatics methods, this work provides an opportunity to investigate the shared genetic pathways, GO analysis and gene interaction networks BT and its comorbidities. This study aims to contribute some opinions to the development of more effective diagnostic approaches and treatment strategies for BT patients, ultimately improving their quality of life and reducing the burden of associated complications.
By identifying key genes and molecular mechanisms, this study aims to contribute to the development of more effective diagnostic biomarkers, therapeutic goals and treatment strategies that obviously improve outcomes and quality of health for patients.

1.4 Problem Statement

Mutations in the HBB gene of blood causing the hereditary disorder named Beta-thalassemia (BT) because of the reduced production of hemoglobin. Hemoglobin is one type of protein found in red blood cells that contains iron and transfers oxygen to tissues across the body. Patients with BT have decreased hemoglobin levels that leads to a reduced supply of oxygen throughout the body. As a result, patients

have anemia which requires regular blood transfusions for life and causes weakness, skin problems, fatigue and sometimes serious health issues. Thalassemia major and thalassemia intermediate are the two types of BT where BT major patient,Äôs dependent on regular blood transfusions and BT minor patient,Äôs need not to blood transfusion. Although the blood-related symptoms of BT are well established, increasing evidence shows that the disease is linked to various serious comorbid conditions. These include endocrine diseases such as T2D, PCOS hypothyroidism and hypogonadism as well as cardiovascular complications such as ACM and arrhythmia. Such comorbidities significantly worsen patient prognosis, reduce quality of life, and increase mortality rates. These complications produced negative prognosis, life with health issues and increase the risk of death.

Despite extensive clinical research, the proper molecular and genetic mechanisms that link the BT to these comorbidities remain poorly understood. Many Current studies focus on individual complications separately without exploring the potential shared genetic pathways or overlapping disease mechanisms. These approaches limit the identification of common genes and molecular perspectives that limit utilization of early diagnosis, prediction of future risk and specific treatment. Furthermore, most available research is clinical or descriptive lacking of integrative bioinformatics approaches. Where this work can analyze large-scale genomic data to uncover underlying relationships between BT and its associated disorders.

Facing several key challenges because of the absence of such integrative approaches. Firstly, it is difficult to develop diagnostic tools that can detect comorbidities at an early stage without identifying the shared genes and molecular relationships. Secondly, the lack of molecular insights affects the development of targeted therapies that address simultaneously both BT and its comorbidities rather than treating them as separated. Finally, the opportunities for drug recycling or the development of therapeutic approaches will be missing without mapping DGNs, PDI and PPI networks.

Addressing these gaps requires a systematic and computational approach that can integrate gene expression datasets from multiple diseases to identify overlapping genes, pathways, and protein interactions. In this study, publicly available microarray and mRNA datasets from NCBI are analyzed to perform genetic profiling of BT with its major comorbidities. The analysis involves constructing DGNs, heatmap visualization of gene patterns, pathway and ontology, PPI and PDI mapping as well as phylogenetic analysis to investigate the evolutionary connections between the diseases.

This research aims to overcome these knowledge gaps, enabling a integrative medical strategies that can improve the ways of diagnosis, treatment design and ultimately enhance the outcomes of patient situation by revealing the shared genetic relationships

between BT and its comorbidities.

1.5 Thesis Objectives

The primary objectives of this research is to investigate the genetic relationships between BT and its selected endocrine and cardiovascular comorbidities. This work aims to identifying the mechanisms of shared molecular that may help in integrative diagnostic and therapeutic approaches.
The specific objectives are:

1) Identify the common differentially expressed genes (DEGs) between BT and its associated comorbidities including hypothyroidism, hypogonadism, PCOS, T2D, ACM and arrhythmia by using microarray and mRNA datasets from publicly available NCBI tools.

2) Construct and visualize disease-gene networks (DGNs) to show the genetic overlaps between BT and each comorbidities.

3) Visualize heatmap analysis to reveal the gene expression patterns across BT and the selected comorbidities.

4) To explore the biological processes, molecular functions and cellular components of the shared genes pathway and gene ontology enrichment analysis is conducted.

5) Build protein-protein interaction (PPI) networks to identify the key hub proteins from molecular interaction.

6) Map protein-drug interactions (PDI) to highlight potential therapeutic treatments and opportunities for drug recreations.

7) Create a validation network to confirm that the selected comorbidities are valid for the beta-thalassemia based on genetic perspectives.

8) To determine the evolutionary relationships among BT and its comorbidities phylogenetic analysis is performed based on the shared genes.

This study aims to bridge the gap between genetic profiling and molecular interactions that helps in clinical application, supporting early detection of health risk, personalized treatment and improved patient outcomes affected by BT and its related comorbidities.

1.6 Thesis Contribution

The proposed computational framework provides an integrated bioinformatics approach to exploring the genetic connections of beta-thalassemia and its associated comorbidities. By analyzing microarray and mRNA datasets from National Center for Biotechnology Information NCBI, this study identifies shared genes, molecular pathways and uncovers potential therapeutic targets.
The main contributions of this research are:

1) A total of 15,008 differentially expressed raw genes were analyzed from NCBI platforms that identifying significantly expressed genes associated with beta-thalassemia and its comorbidities.

2) A total of six comorbidities have been selected to validate the connections with BT using gold benchmark.

3) This study discovered 13, 165, 13, 14, 11, and 44 common DEGs for hypothyroidism, hypogonadism, PCOS, T2D, ACM and arrhythmia respectively that highlighting the potential molecular links between beta-thalassemia and these conditions.

4) Built and analyzed DGN networks to visualize genetic interconnections and identify key hub genes that playing a central role in disease progression.

5) Revealed relevant signaling pathways and Gene Ontology terms strongly associated with the shared DEGs.

6) Conducted PPI and PDI analyses to identify the central proteins and the potential drug molecules that may target both beta-thalassemia and its comorbidities.

7) Developed a validation network to validate the selection of comorbidities and then performed phylogenetic analysis to determine the evolutionary relationships among the diseases.

8) Provides an established view of how genetic, molecular and evolutionary factors interplay in beta-thalassemia and its associated complications to providing the foundation of more targeted diagnostics and personalized treatment strategies.

1.7 Significance

This study has the substantial importance in advancing the understanding of beta-thalassemia and its associated comorbidities through an integrative genetic and

bioinformatics approach. By identifying key genetic mutations, enriched pathways, functional ontologies and regulatory mRNAs, this study aims to pave the way for identifying potential biomarkers that could improve early diagnosis and help predict disease outcomes. Exploring the networks of protein-protein and protein-drug interaction will help in the identification of novel therapeutic targets and support the opportunities of drug repurposing.
Furthermore, phylogenetic analysis of disease-associated mutations will provide evolutionary insights into their origin and prevalence that enhancing the global understanding of epidemiological. Overall, the findings of this study are expected to support more personalized treatment strategies, guide future molecular research and increase the awareness about the complex health risks faced by beta-thalassemia patients.

1.8 Thesis Layout

The rest of the thesis is organized as follows.

CHAPTER 1: This chapter introduces the background of beta-thalassemia, its clinical significance and the motivation for this study. It also outlines the problem statement, objectives, scope, goal and expected contributions of the research.

CHAPTER 2: This chapter presents a comprehensive literature review, summarizing previous studies from reputable journals and conferences. It highlights the current understanding of shared molecular mechanisms and identifies the research gaps.

CHAPTER 3: This chapter describes the methodology and datasets used in the study. It details the sources of microarray and mRNA data obtained from NCBI, the preprocessing steps, and the computational approaches applied such as the Benjamini Hochberg for identifying shared genes. Another method is used for diseasome network construction, pathway and ontology analysis, protein-protein and protein-drug interaction mapping and phylogenetic analysis.

CHAPTER 4: The most significant chapter that presents the results of the analyses, including the identification of common differentially expressed genes between beta-thalassemia and each comorbidity, visualization of gene expression patterns and functional enrichment findings. It also reports the construction of interaction networks, key hub proteins, drug association analysis, and evolutionary relationships.

CHAPTER 5: This chapter concludes the thesis by summarizing the major findings and their implications for clinical research and treatment strategies. It also outlines the limitations of the study and proposes directions for future work in exploring

genetic and molecular connections between beta-thalassemia and other diseases.

1.9 Conclusion

This introductory chapter describes the main research questions, goals, objectives and the motivation of driving this study. The subsequent chapters will present a detailed review of existing literature, describe the methodology for data collection, data analysis, and discuss the results in relation to the current knowledge. The next chapters will go over each of them in depth.

CHAPTER 2

LITERATURE REVIEW

This chapter highlights the literature-related work carried out in recent years, along with a discussion of the methodologies adopted in those studies and their identified limitations. Many related works have been conducted using a variety of bioinformatics algorithms, computational tools, and gene expression datasets to explore the genetic and molecular links between diseases, but not for beta-thalassemia and its comorbidities, which is highlighted in this work through a literature review. The main goal of this chapter is to review and discuss the latest literature, methodologies, and findings in this domain, to identify key areas where further research can be undertaken to bridge existing gaps.

2.1 Thesis Literature

Some literature that triggered this study is discussed here:
In [3] they explored the genetic links between Type 2 diabetes (T2D) and its comorbidities, including kidney failure, liver cancer, myocardial infarction, endometrial cancer, embolic stroke, xanthoma, and xerostomia. They used multiple Gene Expression Omnibus (GEO) microarray datasets for each disease, constructed gene-disease networks (GDNs), performed pathway analysis on KEGG, conducted GO analysis, and built PPI networks. They identified several shared differentially expressed genes (DEGs) between T2D and each comorbidity, suggesting significant molecular associations. However, the study did not include PDI analysis, phylogenetic analysis, or validation analysis, which could have provided stronger clinical insights.
In [4] evidence suggests that COVID-19 may increase the risk of developing neurodegenerative diseases (NDGDs) like stroke, Alzheimer's disease, epilepsy, Parkinson's disease, and multiple sclerosis. GEO microarray datasets for COVID-19 and these NDGDs were analyzed to uncover shared molecular patterns. The study observed that COVID-19 shared 19, 26, 20, 19, and 22 DEGs with epilepsy, stroke, multiple sclerosis, Alzheimer's disease, and Parkinson's disease, respectively. They mapped disease-gene relationships, explored dysregulated pathways, and built PPI and PDI networks, validating their results.
The study in [4] also investigates the genetic and pathogenetic similarities between 2019-nCoV (COVID-19) and other coronaviruses, particularly SARS-CoV. They identified hundreds of dysregulated genes using genome alignment, DNA-DNA hybridization, and gene expression comparisons. They constructed an infectome-diseasome network of up- and down-regulated genes, PPI networks, protein-chemical interactions (PCI), and analyzed pathways and gene ontologies. This work aims to understand shared mechanisms between COVID-19 and related viruses for drug repurposing. In contrast, our work focuses on uncovering shared molecular mechanisms between beta-thalassemia and its comorbidities.
The study in [6] explored the molecular relationships between COVID-19 and its comorbidities, including lung cancer, hypertension, myocardial infarction, and diabetes mellitus. They identified 93 upregulated and 15 downregulated genes in COVID-19, with overlaps of 28 shared genes with diabetes mellitus, 17 with lung cancer, 6 with myocardial infarction, and 7 with hypertension. They performed signaling pathway analysis, GO analysis, PPI and hub protein analysis, PDI analysis, and constructed networks of dysregulated genes. However, this study did not validate their work or perform phylogenetic analysis.
The study in [7] investigates the genetic connections between gastric cancer and its common comorbidities, including kidney disease, diabetes, stroke, and liver cancer.

Using mRNA-seq and microarray datasets, they identified matching shared genes, constructed gene-disease networks, analyzed pathways, ontologies, and protein interactions, and validated their work with benchmark databases. This computational work highlights significant genetic associations between gastric cancer and these comorbid conditions. However, it did not analyze PDI interactions or phylogenetic analysis.

In [8] the focus was on identifying influential genes (IFGs) in glioblastoma using the Cancer Genome Atlas (TCGA) dataset to understand its genetic links with various comorbidities. They identified 26 dysregulated IFGs from over 16,261 genes through statistical analysis, conducting further analyses including protein-protein and protein-drug interactions, comorbidity networks, and phylogenetic analysis. However, they did not analyze signaling pathways, GO, or validation networks, limiting their work compared to ours.

The work in [9] identified matching genes among welding fume (WF) and respiratory system diseases (RSDs) by developing a quantitative framework. Using microarray data for WF and RSDs (e.g., asthma, lung cancer, chronic bronchitis, pulmonary edema), they focused on identifying common genes, their networks, pathway analysis, GO analysis, and PPI analysis, validating their results.

The study in [10] analyzes gene expression to reveal genetic links between Parkinson's disease (PD) and other neurodegenerative diseases (Alzheimer's, ALS, Huntington's, and multiple sclerosis). They identified shared dysregulated genes, pathways, GO analysis, phylogenetic analysis, and protein interactions, validating their findings to highlight PD's potential role in the progression of these disorders.

In [13] the paper showed bidirectional connections between T2D and breast cancer using GSE 29231, GSE70905, and GSE50586 for diabetes, malignant breast tissue, and both biopsies, respectively. They identified 94 common DEGs, constructed a PPI network, performed limited pathway and ontology analysis, and identified hub proteins and survival construction.

In [14] differentially expressed genes (DEGs) were identified for colorectal cancer (CRC) and eight related comorbidities. Protein interaction analysis uncovered four sub-networks and eight key hub genes as potential therapeutic targets, predicting clinical outcomes and highlighting genes linked to CRC progression and patient survival. The study reviews machine learning and network-based methods for discovering genetic risk factors for CRC.

The study in [19] applied bioinformatics and systems biology methods to identify risk factors for cardiovascular disease (CVD) progression. They found 32, 17, 53, 70, and 89 common DEGs between CVD and its associated risk factors, identifying potential biomarkers through PPI analysis, pathway analysis, and ontology analysis, validated using benchmark databases.

2.2 Brief Summary of Related Work

The existing related works, their methods, datasets, techniques, and limitations are highlighted in Table 2-1 below. The proposed work aims to overcome these limitations after preprocessing microarray and mRNA-seq datasets.

Table 2.1: Table 2-1 Brief Summary of Related Work

Authors Name	**Datasets**	**Methods**	**Limitations of Work**
Malik, S.E., Kanwal, S., Javed, J., Hidayat, W., Ghaffar, T. and Aamir, A.H., 2023 [1]	135 Beta-Thalassemia Major (BTM) patients	Statistical analysis and laboratory methods	No genetic and molecular relationships
Akiki, N., Hodroj, M.H., Bou-Fakhredin, R., Matli, K., Taher, A.T., 2023 [2]	Secondary data	Summarizing patterns, mechanisms, diagnostic methods, preventive measures, and treatments reported in the literature	No relationships shown among disease and its comorbidities
Podder, N.K., Rana, H.K., Azam, M.S., Rana, M.S., Akhtar, M.R., Rahman, M.R., Rahman, M.H. and Moni, M.A., 2020 [3]	GEO microarray datasets	Statistical methods, quantitative model, z-transform, and multi-layered topologies	Fails to calculate large-scale datasets; No drug protein identification; No phylogenetic analysis
Podder, N.K., Shill, P.C., Rana, H.K., Omit, S.B.S., Al Shahriar, M.M.H. and Azam, M.S., 2021 [4]	mRNA and microarray datasets	Statistical methods and algorithm, and Benjamini-Hochberg algorithm	No phylogenetic analysis; Fails to validate their selected comorbidities

Table 2.1: Table 2-1 Brief Summary of Related Work (Continued)

Authors Name	**Datasets**	**Methods**	**Limitations of Work**
Datta, R., Podder, N.K., Rana, H.K., Islam, M.K.B. and Moni, M.A., 2020 [7]	Microarray and mRNA-seq datasets	Statistical methods (t-test), z-transformation, multilayer topology, and neighborhood benchmark method	No drug protein identification; No phylogenetic analysis
Podder, N.K. and Shill, P.C., 2022 [8]	TCGA datasets	Statistical and bioinformatics model	No significant identification of signaling pathway and GO analysis; Fails to validate their work
Rana, M.S., Podder, N.K., Rana, H.K., Hasan, M.I., Azam, M.S., Rahim, M.A., Iqbal, S.H.S. and Saha, S., 2023 [10]	mRNA and microarray datasets	Benjamini-Hochberg algorithm, neighborhood-based benchmarks, and multilayer topology	Fails to identify protein-drug intersections
Durrani, I.A., Bhatti, A. and John, P., 2023 [13]	Microarray and mRNA-seq datasets	Integrated in silico analyses approach	No phylogenetic analysis; Fails to identify drug protein
Talihati, Z., Abudurousuli, K., Hailati, S., Han, M., Nuer, M., Khan, N., Maihemuti, N., Simayi, J., Zhang, W. and Zhou, W., 2025 [14]	TCGA database, genomic database, transcriptomic, and GEO datasets	Limma package in R, STRING database, molecular docking, etc.	Fails to analyze PDI, phylogenetic, and validation analysis

Table 2.1: Table 2-1 Brief Summary of Related Work (Continued)

Authors Name	**Datasets**	**Methods**	**Limitations of Work**
Barua, J.D., Omit, S.B.S., Rana, H.K., Podder, N.K., Chowdhury, U.N. and Rahman, M.H., 2022 [19]	GEO microarray datasets	Z-transformation, neighborhood-benchmark, and multilayered topology	Fails to identify drug protein; No phylogenetic analysis
Rahman, M.H., et al., 2023 [25]	GEO datasets	Design matrix model, fit-linear, and Bayesian model	Fails to identify hub and drug proteins; Fails to validate the work; Fails to calculate large-scale data

2.3 Research Gap

Our proposed work offers a more comprehensive level of analysis compared to the referenced studies, as it integrates all analytical approaches previously applied separately in the existing literature. This study comparatively investigates all analyses to explore the deep genetic and molecular correlation of beta-thalassemia with associated endocrine and cardiac diseases. This is shown in Table 2-2.

Table 2.2: Comparison of Proposed Work with Related Studies

Related Work	**Datasets**	**Diseases**	**DGN, Pathways, GO, PPI, Validation Network**	**PDI, Phylogenetic Analysis**
[3]	Microarray datasets	T2D vs. comorbidities; Gastric cancer vs. comorbidities; Welding fumes vs. respiratory system	YES	NO
[4],[6]	mRNA and microarray datasets	COVID-19 vs. comorbidities	YES but Validation: NO	NO but PDI: YES

Table 2.2: Comparison of Proposed Work with Related Studies (Continued)

Related Work	**Datasets**	**Diseases**	**DGN, Pathways, GO, PPI, Validation Network**	**PDI, Phylogenetic Analysis**
[7],[9]	Microarray datasets	T2D vs. comorbidities; Gastric cancer vs. comorbidities; Welding fumes vs. respiratory system	YES	NO
[8]	TCGA datasets	Glioblastoma vs. comorbidities	NO but DGN: YES; PPI: YES	YES
[10]	mRNA and microarray datasets	Parkinson's vs. neurodegenerative	YES	NO but Phylogenetic: YES
Proposed Work	mRNA and microarray datasets	Beta-thalassemia vs. endocrine and cardiac diseases	YES	YES

Previous studies on any main diseases and its comorbidities have performed individual analyses such as gene expression profiling, pathway analysis, gene ontology, protein-protein interaction or drug interaction were applied separately and not in an integrated manner. No existing work has combined all of these analytical approaches together to provide a comprehensive genetic, molecular and evolutionary understanding of beta-thalassemia with its associated endocrine and cardiac complications. This gap limits the discovery of common biomarkers, therapeutic targets, and evolutionary insights. Where our work have performed these analyses in combined manner to give a clear and straight visions of BT and its comorbidities connections.

2.4 Conclusion

This chapter has explored previous research relevant to beta-thalassemia and those research that relevant to system biological approaches. While earlier studies have provided valuable insights into gene expression with its clinical outcomes in a separated manner. But this literature review highlights the absence of an integrated framework that connects genetic profiling with pathway enrichment, ontology, interaction networks and evolutionary perspectives. Recognizing these gaps has guided the

direction of the present study, which aims to bring these analyses together to achieve a more clear understanding of beta-thalassemia and its related comorbidities.

CHAPTER 3

METHODOLOGY

In this chapter, the methodology has been clarified. For discussion convenience, there are a total of 3 sub-chapters under chapter 3, which we introduced. In section 3.1, we discussed the collected dataset; in section 3.2, we discussed data pre-processing; in section 3.3, we discussed proposed deep learning architectures.

3.1 Introduction

This chapter outlines the proposed model and methodology for analyzing the genetic profiling between BT and its comorbidities. The study integrates bioinformatics and system biological techniques to process gene expression data of BT and its complications such as PCOS, hypothyroidism, hypogonadism, diabetes, cardiomyopathy and arrhythmia.
The methodology involves data preprocessing, genetic profiling, pathway enrichment, functional ontology, protein-protein and protein-drug interaction networks, phylogenetic analysis and validation analysis. These approaches aim to identify key biomarkers, therapeutic targets and evolutionary insights between the genes of BT and its comorbidities. The details of these analytical approaches are presented in the following sections.

3.2 Overview of Analytical Approaches

This chapter show a standard analytical procedure to gain the genetic link between beta-thalassemia and its comorbidities by analyzing the microarray and mRNA sequence datasets. We have used Gene Expression Omnibus (GEO) datasets for each diseases where each datasets has two groups: normal tissue (healthy or control samples) and malignant tissue (affected samples). Comparing these normal and malignant tissues by using Benjamini-Hochberg method to make sure that the significant genes are reliable and control the false discovery rate by adjusting p-value. A Limma package of R language to identify differentially expressed genes (DEGs) between BT and PCOS, T2D, hypothyroidism, hypogonadism, ACM, arrhythmia respectively. After applying the Benjamini-Hochberg correction to control the false discovery rate, this study applies statistical thresholds to differentiate up-regulated and down-regulated genes. And then by finding common DEGs, disease-gene network was constructed to visualize their co-relations. This study performed ontological analysis, pathway analysis, PPI, PDI, phylogenetic analysis and their respective networks. After that it performed validation analysis using gold benchmark datasets including dbGaP and OMIM.
A working flowchart representing this quantitative method in Figure 3-1.

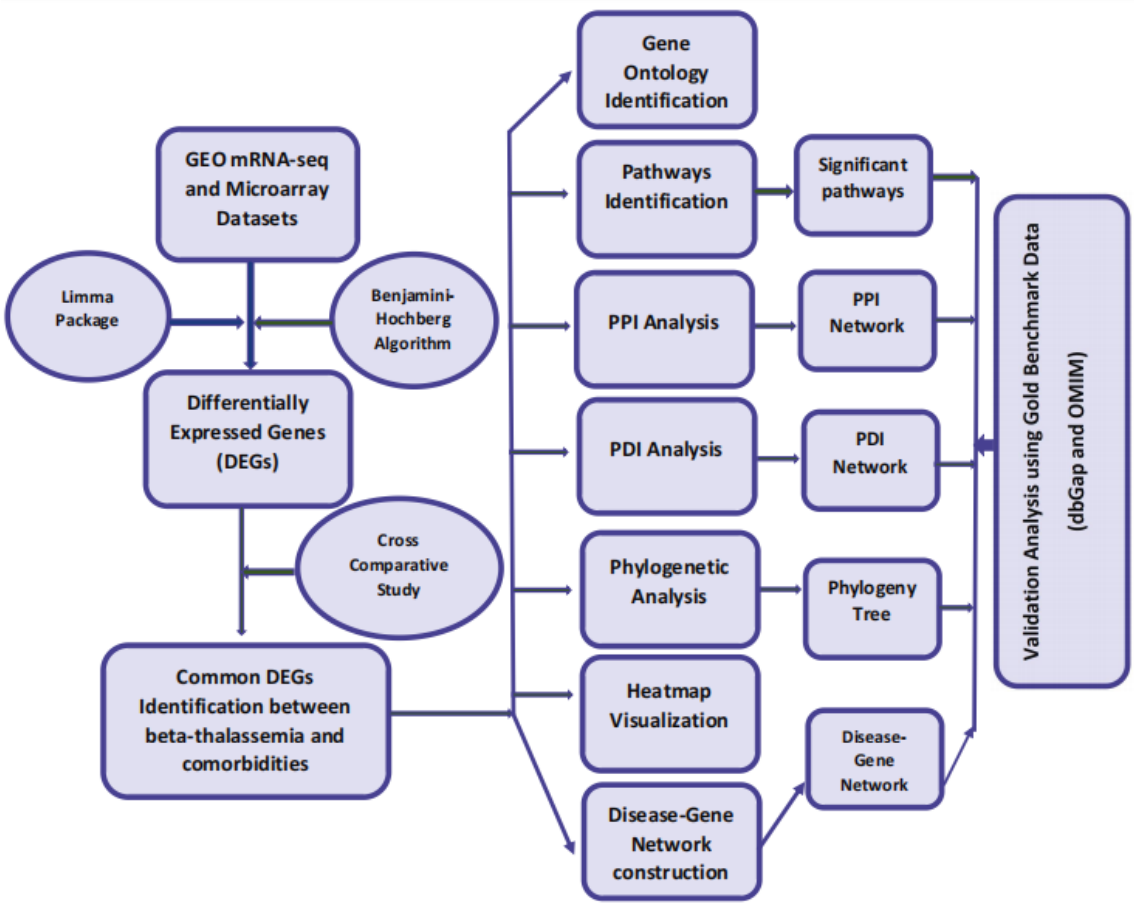


Figure 3.1: Working flow diagram of this investigation

3.3 Dataset Information

This study investigated mRNA-seq datasets for BT, PCOS, hypothyroidism, ACM, and arrhythmia, and microarray datasets for T2D and hypogonadism, available in Gene Expression Omnibus (GEO), which is maintained by National Center for Biotechnology Information (NCBI) [3], to identify the genetic link between beta-thalassemia and its comorbidities. Each and every datasets has two group normal tissue and malignant or affected tissue. This work compares these normal and malignant samples to indicate which genes are expressed differentially. Those differences can then point out the possible genetic links between BT and its comorbidities.

For BT, GSE117221 GEO mRNA-seq dataset are selected which has total 49 samples. 17 samples are normal tissue that means healthy patient and 32 samples are malignant tissue where (15 samples for thalassemia intermediate and 17 samples for thalassemia major) both type of patients are either female or male.

For T2D, GSE25724 expression profiling by array genes are selected with 13 samples where 7 samples for type 2 diabetes patients and 6 samples for non-diabetic patients.

GSE216609 mRNA-seq datasets are selected for PCOS with total 7 samples where 4 sample for control and 3 samples for polycystic ovary syndrome.
With 8 samples GSE176153 GEO mRNA genes are selected for hypothyroidism where 4 samples for healthy patients and 4 samples for malignant patients for both male and female. For ACM RNA-seq datasets are selected from NCBI. Its GEO accession is GSE233780 with 12 samples where 6 samples for ACM and other 6 for healthy control.
GSE26966 GEO microarray datasets are used for hypogonadism with 23 samples where 9 samples for normal pituitary and 14 samples for gonadotrope tumor.
For arrhythmia GSE175944 mRNA-seq datasets are collected which introduced the most common arrhythmia is atrial fibrillation with 6 samples where 3 samples for control patient and 3 samples for case samples.
Table 3-1 provides a detailed representation of datasets including GEO accession id, Gender (no. of male and no. of female), samples (divided into case and control groups), associated diseases and source name that indicates the tissue type, cell type or experimental conditions.

Table 3.1: Representation of Datasets Information

S. No.	**Disease Name**	**GEO Accession Id.**	**Gender (Male + Female)**	**Sample (Case + Control)**	**Disease status (case, control)**	**Source Name**
1	Beta-thalassemia	GSE-117221	49 total, (23 + 26)	49 samples, (32 + 17)	Thalassemia intermediate (TI) and Thalassemia major (TM), and Healthy patient	Healthy-ErPCs, TI-ErPCs, TM-ErPCs
2	T2D	GSE-25724	13 total, (07 + 06)	13 samples, (06 + 07)	Non-diabetic, and Type 2 diabetes	Human islets, non-diabetic and Human islets, diabetic
3	PCOS	GSE-216609	07 total, (07 female)	07 samples, (03 + 04)	Control, and PCOS	Cumulus granule cells

Table 3.1: Representation of Datasets Information (Continued)

S. No.	**Disease Name**	**GEO Accession Id.**	**Gender (Male + Female)**	**Sample (Case + Control)**	**Disease status (case, control)**	**Source Name**
4	Hypothyroidism	GSE-176153	08 total, (02 + 06)	08 samples, (04 + 04)	Control, and Hypothyroidism	Whole blood
5	ACM	GSE-233780	12 total, (10 + 02)	12 samples, (06 + 06)	ACM, and Healthy control	Cardiac mesenchymal stromal cells
6	Hypogonadism	GSE-26966	23 total, (12 + 11)	23 samples, (09 + 14)	Normal pituitary (NP), and Gonadotrope tumor (GT)	NP at autopsy within 2–18 hr. of death, and GT at time of transsphenoidal surgery
7	Arrhythmia	GSE-175944	Not mentioned	06 samples, (03 + 03)	Control, and PITX2 knockout	Cell line of human induced pluripotent stem cells (hiPSC)

3.4 Methodology

This section provides a detailed overview of the methodology applied in this study. Different analytical approaches have been applied to investigate the genetic profiling of beta-thalassemia and its associated comorbidities. The process includes data collection, preprocessing, identification of important functional genes, enrichment analysis, and network-based evaluations such as protein-protein and protein-drug interactions. Additionally, evolutionary relationships has been incorporated to strengthen the findings. By combining these diverse methods, the study aims to ensure a comprehensive and reliable analysis framework that captures both the molecular and clinical perspectives of beta-thalassemia.

3.4.1 Selection of mRNA-seq and Microarray Datasets

For BdSL classification, this study used pre-trained convolutional neural networks, ResNet-50 for feature extraction and evaluating two classifiers SVM (Support Vector Machine) and CNN (Convolutional Neural Network) for the final classification job. For this study, a reliable and relevant data was selected for investigating the genetic association between BT and its comorbidities. Two types of datasets were utilized that are mRNA-seq datasets and microarray datasets which were selected from the Gene Expression Omnibus (GEO), a public functional genomics data repository maintained by the National Center for Biotechnology Information (NCBI) [new ref].

- mRNA-seq datasets (gene expression that is profiling by high throughput sequencing) were selected for BT, PCOS, hypothyroidism, ACM and arrhythmia. These datasets provide a comprehensive view of transcriptional expression.
- Microarray datasets (gene expression that is profiling by array) were selected for T2D and hypogonadism from GEO. Microarray datasets remain valuable for differential expression analysis for experimental designs.

3.4.2 Differential Gene Expression Analysis

Differential gene expression analysis is a base of bioinformatics for understanding the molecular and genetic inter-relations of complex diseases like beta-thalassemia and its comorbidities. In this study, we performed gene expression analysis on microarray and mRNA-seq datasets to identify differentially expressed genes (DEGs) between case disease and control samples for BT, PCOS, hypothyroidism, hypogonadism, T2D, ACM and arrhythmia. The analysis was conducted using the GEO2R tool on NCBI Gene Expression Omnibus (GEO) website and the Limma package in R which are widely used for strong statistical frameworks.

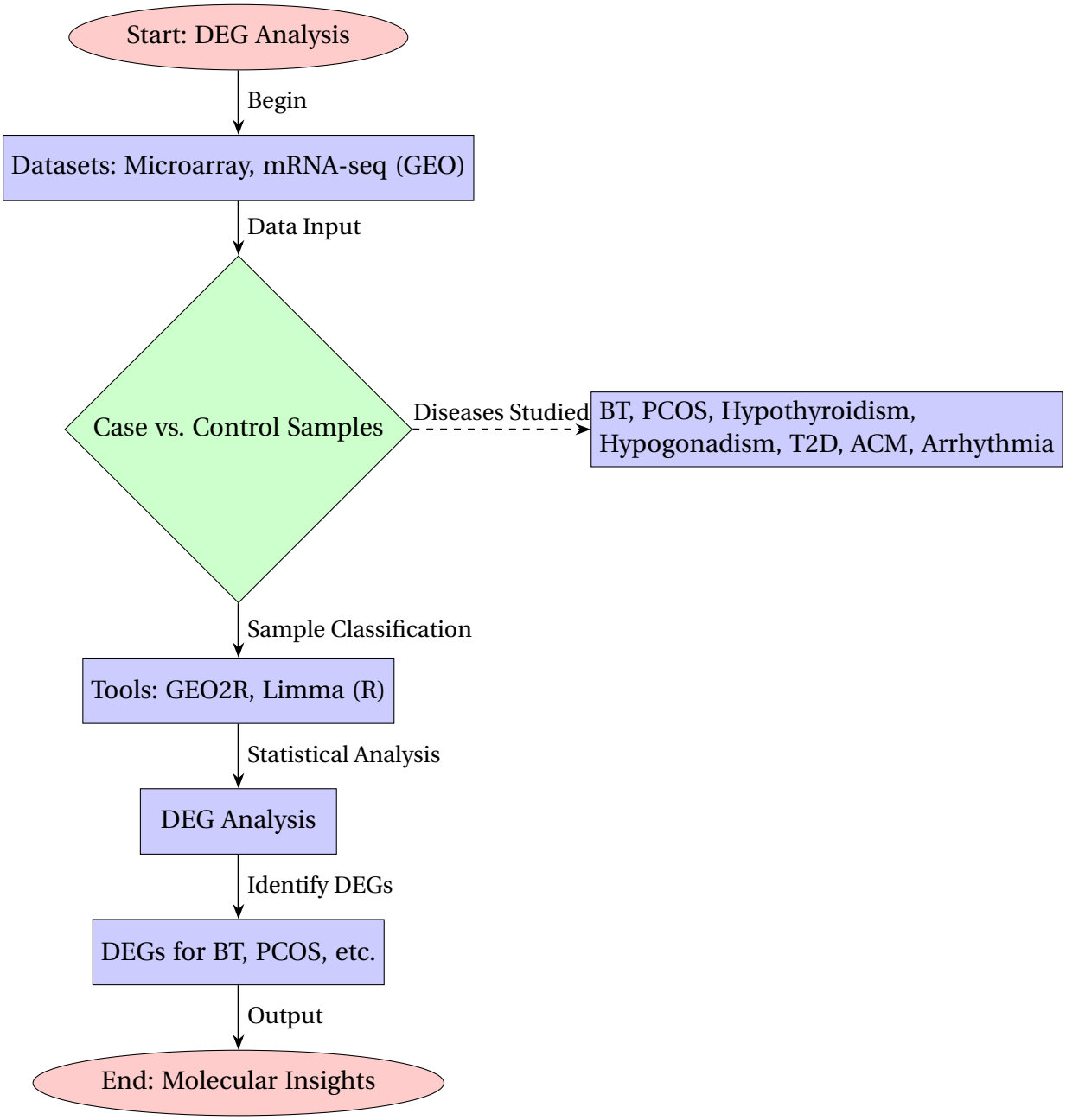


Figure 3.2: Workflow for Differential Gene Expression Analysis of Beta-Thalassemia and Comorbidities.

Data Pre-processing and Normalization

To reduce experimental differences and keep the data consistent, gene expression values in each dataset were normalized using the Z-score method. This approach standardizes the expression values to make them comparable across different samples [**?**]. The Z-score transformation for a gene expression value is calculated as:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{3.1}$$

where:

- z_{ij}: Normalized gene expression value for gene i in sample j,
- x_{ij}: The value of gene expression i in sample j,
- μ_i: Mean value of gene expression i in all samples,
- σ_i: The standard deviation of gene expression values i.

The normalization method was applied to both microarray (for T2D and hypogonadism) and mRNA-seq (for BT, hypothyroidism, PCOS, ACM, and arrhythmia) datasets to ensure reliable identification of differentially expressed genes (DEGs).

Statistical Analysis for DEG Identification

We identified the differentially expressed genes (DEGs) by applying the Benjamini-Hochberg (BH) correction method, which minimizes false positives by controlling the false discovery rate (FDR). The BH formula is:

$$\text{BH adjusted } p\text{-value} = \frac{i}{m}Q \tag{3.2}$$

where:

- i: Rank of individual p-value,
- m: Total number of tests,
- Q: False discovery rate threshold.

The adjusted p-value threshold of BH was set to ≤ 0.05 to ensure statistical significance. These genes were classified based on the $\log_2$ fold change ($\log FC$) values by applying the following conditions:

- To identify up-regulated genes: **log FC** > threshold,
- To identify down-regulated genes: **log FC** < threshold.

These threshold values were used to identify significant DEGs between case and control groups.

3.4.3 Gene Diseases Network Analysis

This study constructed Disease-Gene Networks (DGNs) using a neighborhood-based benchmarking and topological approach to investigate the genetic associations

between beta-thalassemia and its comorbidities. These networks of up- and down-regulated genes show the relationships between diseases and differentially expressed genes (DEGs), providing a visual framework to identify shared genetic signatures.
Each node represents either a disease or a gene, where diseases are source nodes and their associated genes are target nodes in the DGN network. This disease-gene connection forms a bipartite graph. A connection exists if at least one significant DEG is shared between a disease and beta-thalassemia. Let D represent the set of diseases and Z represent the set of dysregulated genes identified from the gene expression analysis, where a gene $z \in Z$ is connected to a disease $d \in D$ if z is significantly dysregulated in the dataset for d. If associated diseases d_1 and d_2 have sets of significant dysregulated genes Z_1 and Z_2, respectively, then the number of shared genes is defined as:

$$|Z_1 \cap Z_2| \tag{3.3}$$

By using the Jaccard Coefficient, this study measured the similarity between disease pairs, which evaluates the overlap relative to the total unique genes. The edge weight or similarity score between diseases d_1 and d_2 is calculated as:

$$X(d_1, d_2) = \frac{|Z_1 \cap Z_2|}{|Z_1 \cup Z_2|} \tag{3.4}$$

Here, $|Z_1 \cap Z_2|$ represents the set of common DEGs, and $|Z_1 \cup Z_2|$ represents the union of DEGs for the two diseases. The Jaccard Coefficient $X(d_1, d_2)$ ranges from 0 to 1, where a higher value indicates greater similarity in the genetic profiles of the diseases [**?**].
Two separate networks were constructed for up-regulated and down-regulated DEGs to show the distinct regulatory patterns. The DGNs were visualized and analyzed using the Cytoscape platform. Cytoscape enabled the mapping of disease-gene associations by optimizing the layouts to highlight connectivity patterns, such as hub genes across multiple comorbidities.

3.4.4 Pathway and Functional Association Analysis

Pathway analysis involves the systematic identification of enriched biological pathways from sets of differentially expressed genes (DEGs) that are shared between beta-thalassemia and its comorbidities. This study identifies significant pathways using Enrichr, a public tool developed by the Ma'ayan Laboratory for Computational Systems Biology for gene set enrichment analysis (GSEA). Enrichr works by comparing input gene lists with collections of curated biological databases. The method applies statistical tests to measure significance, ensuring reliable and meaningful insights into the underlying molecular mechanisms.

A pathway is a sequence of molecular interactions that results in a specific change or product in a cell. It is a standard method for understanding the connections between complex diseases [**?**]. We investigated dysregulated gene pathways across four databases using Enrichr: Reactome, KEGG, WikiPathways, and BioCarta.

1. **Preparation of Input Gene Sets**
   Enrichr compares the input gene list against a reference set that includes all known annotated genes in the human genome (approximately 20,000–30,000), depending on each library. Using this background, it ensures that the analysis is fair and considers the full range of genes that could potentially be expressed.

2. **Selection of Pathway Libraries**
   Enrichr compares the input gene list against multiple pathway-focused databases. This study selects four databases based on their relevance to biological processes:

   - **KEGG** (Kyoto Encyclopedia of Genes and Genomes) provides correlated sets of genes involved in metabolic, signaling, and disease pathways.
   - **Reactome** is an open-source database of manually curated biological pathways, focusing on reactions and interactions.
   - **WikiPathways** is a community-curated resource with pathways linked to drug interactions and biological processes.
   - **BioCarta** focuses on gene interaction models in signaling and metabolic pathways.

Using Enrichr, we analyzed the common DEGs identified from the cross-comparative analysis of beta-thalassemia and its comorbidities. Pathways were considered significant if their adjusted p-value was ≤ 0.05, ensuring robust statistical reliability. The analysis identified key pathways for each comorbidity, such as GPCR ligand binding for PCOS, interferon gamma signaling for hypogonadism, and zinc homeostasis for arrhythmia, among others.

3.4.5 Gene Ontology Analysis

Gene Ontology (GO) analysis is a structured framework for describing genes by their molecular functions, biological processes, and cellular components, enabling a deeper understanding of their roles in disease mechanisms [**?**]. In this study, we conducted GO analysis to explore the functional associations of common differentially expressed genes (DEGs) shared between beta-thalassemia and its comorbidities. The analysis was performed using Enrichr, a web-based enrichment tool developed

by the Ma'ayan Laboratory, which queries gene sets against standardized ontology databases to identify significant functional terms [**?**].

GO Analysis Workflow The common DEGs identified through cross-comparative analysis of GEO datasets were used as input for Enrichr. We focused on two key ontology databases:

- **GO Biological Process (GO Term)** categorizes genes based on their involvement in coordinated biological processes, such as signaling pathways or metabolic activities.
- **Human Phenotype Ontology (HP Term)** links genes to phenotypic abnormalities observed in diseases, facilitating the identification of clinical manifestations associated with DEGs.

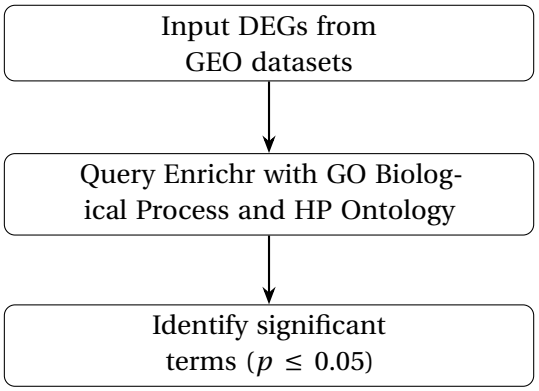


Figure 3.3: Workflow for Gene Ontology (GO) analysis using Enrichr to identify significant functional terms associated with DEGs.

3.4.6 Up and Down-regulated Identification

Downregulation is the process in which a cell reduces the expression level of a gene compared to a reference in response to an external variable. Upregulation refers to an increase in expression level compared to a reference. Here, the expression level indicates the abundance of biological components, such as RNA or protein [**?**]. These changes are typically determined by comparing two groups: a reference or control sample, often healthy tissue, and a test sample, such as diseased or mutant tissue. The expression values of the test sample are compared against those of the reference sample to generate expression ratios, which is standard practice in gene expression analysis. These ratios are typically transformed into a logarithmic scale for easier interpretation. A positive log value indicates that the gene expression is higher in the test sample compared to the reference, indicating upregulation. A

negative log value signifies reduced expression in the test sample relative to the reference, indicating downregulation. Thus, by comparing the expression profiles of two samples, it is possible to determine whether a gene is up- or down-regulated under specific conditions.

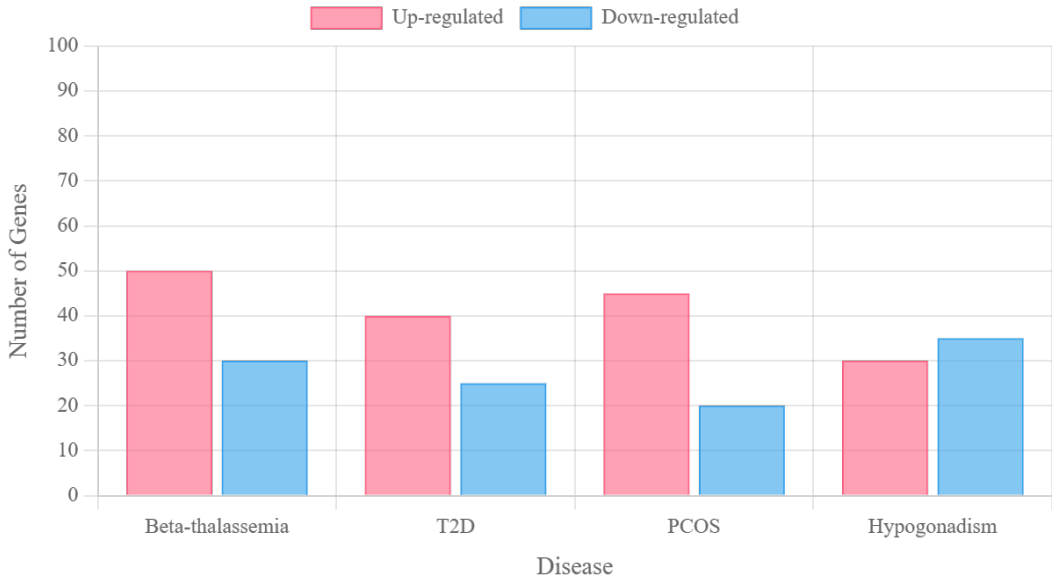


Figure 3.4: Number of up- and down-regulated genes across diseases, identified by comparing expression levels between test and control samples using $\log_2$ fold change thresholds (e.g., $\log FC > 1$ for up-regulated, $\log FC < -1$ for down-regulated, $p \leq 0.05$).

3.4.7 Common Dysregulated Genes

This study identified common differentially expressed genes (DEGs) shared between beta-thalassemia (BT) and its comorbidities by using statistical analysis to explore genetic relationships among these diseases. We retrieved GEO datasets from NCBI for the diseases under investigation. To reject the null hypothesis, a statistical p-value of ≤ 0.05 was used. For up-regulated gene identification, a threshold of $|\log FC| \geq 1$ was applied, and for down-regulated gene identification, a threshold of $|\log FC| \leq -1$ was used [**?**]. By performing an intersection set operation using the R `limma` package, we identified common genes between BT and its comorbidities. Using these datasets and methods, we obtained results that demonstrate the connections between BT and the selected diseases. Other detailed information is discussed in the next chapter.

3.4.8 Network Construction of PPI

Protein-Protein Interaction (PPI) analysis is a primary goal of systems biology, predicting protein functions and drug targets through molecular interactions [**?**]. These

interactions are significant for driving cellular processes and mapping connections between beta-thalassemia (BT) and its comorbidities [**?**]. We constructed PPI networks to explore how proteins encoded by shared differentially expressed genes (DEGs) interact.

We used the STRING database, inputting common DEGs and setting a high confidence score of 0.9 for reliable interactions. The NetworkAnalyst platform, utilizing the Markov Clustering algorithm, facilitated clustering of proteins based on connectivity. The resulting networks were visualized in Cytoscape, with proteins represented as nodes and interactions as edges.

Hub proteins were selected based on topological parameters, specifically a degree greater than 10. The distance between a pair of proteins (i, j) is defined as follows:

$$d(i,j) = 1 - \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \tag{3.5}$$

where N_i and N_j are the neighbor sets of proteins i and j, respectively.

3.4.9 Protein-Drug Interactions Network Analysis

The primary objective of this investigation is to identify potential therapeutic drug particles [**?**]. The Protein-Drug Interaction (PDI) network was constructed using NetworkAnalyst and prepared with the DrugBank database, which is designed for common genes of beta-thalassemia and its comorbidities. To explore potential therapeutic options for beta-thalassemia and its comorbidities, we constructed PDI networks. These networks map how proteins encoded by the shared differentially expressed genes (DEGs) interact with drug molecules, which is used to identify possible treatments [**?**]. Using systems biology-based strategies for studying protein-drug interactions [**?**], the goal of this work was to identify potential drug targets capable of addressing the molecular mechanisms that connect these diseases.

Building the PDI Networks

Firstly, we identified the common DEGs between beta-thalassemia and each comorbidity. These genes were input into the NetworkAnalyst platform, which integrates the DrugBank database to map interactions between proteins and known drug compounds [**?**]. DrugBank provides a comprehensive repository of drugs and their protein targets, including approved, experimental, and investigational compounds, which is used to ensure a broad scope of potential interactions.

Analytical Approach

Following the systems biology framework outlined by Colinge et al. (2012) [**?**], we analyzed the PDI network by:

- Topological Analysis is used to identify hub proteins with multiple drug interactions by using metrics like degree (the number of connections) to prioritize potential therapeutic targets.
- Functional Mapping is used for cross-referenced protein functions to ensure drugs targeted biologically relevant proteins, such as those involved in immune response or signaling pathways linked to beta-thalassemia,Äôs comorbidities.
- Drug Prioritization were ranked based on the number of protein targets and their relevance to disease pathways, ensuring focus on compounds with potential clinical applicability.

The PDI network helps us see which drugs might influence the proteins driving beta-thalassemia and its comorbidities.

3.4.10 Phylogenetic Analysis

Phylogenetic analysis enhances the understanding of the evolutionary relationships among genes, proteins, and diseases [**?**]. We constructed a phylogenetic tree for beta-thalassemia and its comorbidities to illustrate the associative relationships among them by using Molecular Evolutionary Genetics Analysis (MEGA) tools and FASTA sequences of nucleotide datasets from NCBI [**?**]. The tree shows a strong evolutionary relationship between beta-thalassemia and hypogonadism as they belong to a common species. We retrieved nucleotide sequences in FASTA format from the NCBI database for representative genes associated with beta-thalassemia and its comorbidities by selecting sequences based on their relevance to shared differentially expressed genes (DEGs). To ensure comprehensive coverage, we included multiple sequences per disease and prioritized those linked to key pathways like iron metabolism or endocrine signaling. Following standard molecular phylogenetic protocols [**?**]:

- Sequence Alignment: Sequences were aligned using ClustalX to identify homologous regions, accounting for insertions, deletions, and substitutions. This step ensures accurate comparison for evolutionary inferences.
- Tree Construction: We used the Molecular Evolutionary Genetics Analysis (MEGA) software to generate phylogenetic trees. The Neighbor-Joining (NJ)

method was employed, with the Kimura 2-parameter model for distance calculation to handle nucleotide substitutions [**?**].

3.4.11 Validation Analysis

To ensure the reliability of our findings on the genetic links between beta-thalassemia and its comorbidities, we conducted a validation analysis using gold-standard biomedical databases. This step confirms that the DEGs and their disease associations are consistent [**?**]. We used two benchmark databases, including dbGaP and OMIM from Enrichr tools, on the up-regulated and down-regulated genes of beta-thalassemia to validate our findings [**?**]:

- dbGaP (Database of Genotypes and Phenotypes): This NCBI-hosted database contains genotype-phenotype relationships from large-scale genomic studies. We queried dbGaP with our shared DEGs to identify diseases with overlapping genetic profiles.
- OMIM (Online Mendelian Inheritance in Man): This comprehensive catalog of human genes and genetic disorders was used to verify whether our DEGs are linked to beta-thalassemia or its comorbidities in Mendelian or complex disease contexts.

3.4.12 Working Algorithm

A systematic methodology outlines the entire operational process as a set of structured steps. The following guidelines offer a detailed summary of all procedures carried out in our research [**?**].

3.4.13 Algorithm

Input: Microarray and mRNA-Seq GEO datasets.
Output: Differentially Expressed Genes (DEGs), Common DEGs, Disease-Gene Networks (DGNs), Signaling Pathways, Ontological Pathways, Protein-Protein Interactions (PPIs), Protein-Drug Interactions (PDIs), Phylogenetic Tree, and Validation Network.

1. **Dataset Selection:**
    - Selecting relevant Gene Expression Omnibus (GEO) datasets from the NCBI using disease-specific criteria.

2. **Differential Gene Expression Analysis:**

   - For every dataset $i = 1, 2, 3, \ldots, N$

     a) Load the datasets.

     b) Normalize datasets using Z-score transformation to ensure comparability.

     c) Create a case vs. control design matrix.

     d) Apply the R `Limma` package and GEO2R to compute DEGs using the Benjamini-Hochberg method.

     e) Filter DEGs based on adjusted p-value ≤ 0.05.

     f) Modify $|\log FC| \geq 1$ for Upregulation and $|\log FC| \leq -1$ for Downregulation.

     g) Identify significant DEGs.

3. **Cross-Comparative Analysis:**

   - Compare DEG gene sets between beta-thalassemia and each comorbidity including T2D, PCOS, ACM, hypothyroidism, hypogonadism, and arrhythmia to identify common up- and down-regulated genes.
   - Use the Jaccard Coefficient to quantify similarity between gene sets.

4. **Some Analysis for Common DEGs:**

   - Disease-Gene Networks (DGNs) construction.
   - Enrichment analysis for significant signaling pathways.
   - Enrichment analysis for Ontological pathways.
   - PPI network construction.
   - PDI network construction.
   - Evolutionary phylogenetic Analysis.
   - Plot these pathways in tabular form.

5. **Validation Analysis:**

   - Build a validation network in Cytoscape.

6. **Results:**

   - List of common DEGs.
   - DGNs for up- and down-regulated genes.
   - Heatmap Visualization.

- Enriched signaling and ontological pathways.
- PPI networks and hub proteins.
- PDI networks with potential drug targets.
- Phylogenetic tree showing disease relationships.
- Validated disease-gene associations.

3.4.14 Experimental Setup

To conduct the genetic and molecular analysis of beta-thalassemia and its comorbidities, we used the following computational setup and tools:

- **Hardware:**
  - Device: Desktop PC
  - Processor: Intel Core i5-7200U CPU @ 2.50GHz (2.71 GHz)
  - RAM: 4.00 GB (3.26 GB usable)
  - System: 64-bit Windows 10 Pro, Version 21H2, OS Build 19044.2006
- **Software and Tools:**
  - R Studio: For differential gene expression analysis using the `Limma` package.
  - Cytoscape: For constructing and visualizing disease-gene and validation networks.
  - MEGA: For phylogenetic tree construction.
  - STRING: For protein-protein interaction (PPI) network analysis.
  - Enrichr: For pathway and gene ontology enrichment analysis.
  - Network Analyst: For PPI and protein-drug interaction (PDI) network construction with DrugBank integration.

3.4.15 Conclusion

This chapter outlined the methodology for investigating the genetic and molecular links between beta-thalassemia and its comorbidities. Using GEO datasets, we identified DEGs, constructed DGNs, performed pathway and GO enrichment, built PPI and PDI networks, conducted phylogenetic analysis, and validated findings with dbGaP and OMIM. The systematic approach, supported by tools like R, Cytoscape, Enrichr, STRING, NetworkAnalyst, and MEGA, ensures robust and reproducible results, setting the stage for detailed findings in Chapter 4.

CHAPTER 4

RESULT ANALYSIS AND DISCUSSION

In this chapter, the result analysis and discussion have been clarified. For discussion convenience, there are a total of 4 sub-chapters under chapter 4, which we introduced. In section 4.1, we discussed experimental tools; in section 4.2, we discussed result; in section 4.3, we discussed the discussion part.

4.1 Experimental Elements

There are different types of implementation tools for deep learning. Most of the tools can be used for thesis paper code. When we want to get our thesis outcome, it must need for any researchers. Most of the time, **Jupyter Notebook, Google Colab, and Kaggle Notebook** are used for the implementation of any thesis paper code. With this tool, we can easily run our code without any problems.

We have used **Jupyter Notebook** in our experiment. It is an open-source, web-based interactive computing platform widely used for data analysis, machine learning, and scientific research. Jupyter Notebook allows us to create and share documents containing live code, equations, visualizations, and narrative text. It supports a variety of programming languages, with Python being the most popular. For implementation, no internet connection is required when running it locally, but it can also be hosted on cloud platforms. Jupyter Notebook offers seamless integration with many computational libraries, making it a versatile tool for executing and documenting experiments. The flexibility to visualize data, run code interactively, and document results in a single environment makes Jupyter Notebook a preferred choice for researchers and developers.

4.2 Result

This study applied a two-stage methodology for BdSL classification, utilizing ResNet-50 for feature extraction and evaluating two classifiers SVM (Support Vector Machine) and CNN (Convolutional Neural Network) for the final classification job. After extract feature using ResNet-50 SVM train must faster. And dense layer also works well.

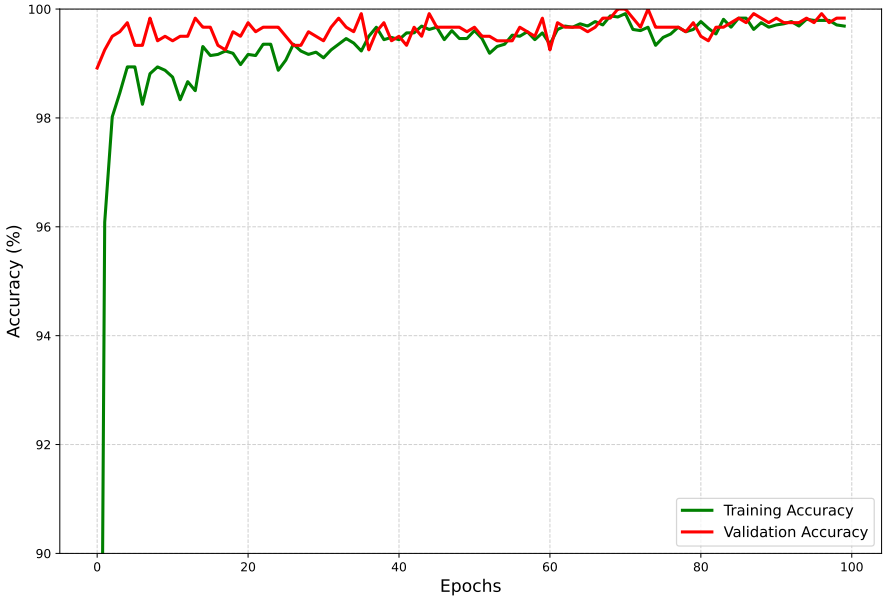


Figure 4.1: Accuracy Curve (ResNet-50 + Dense Layer)

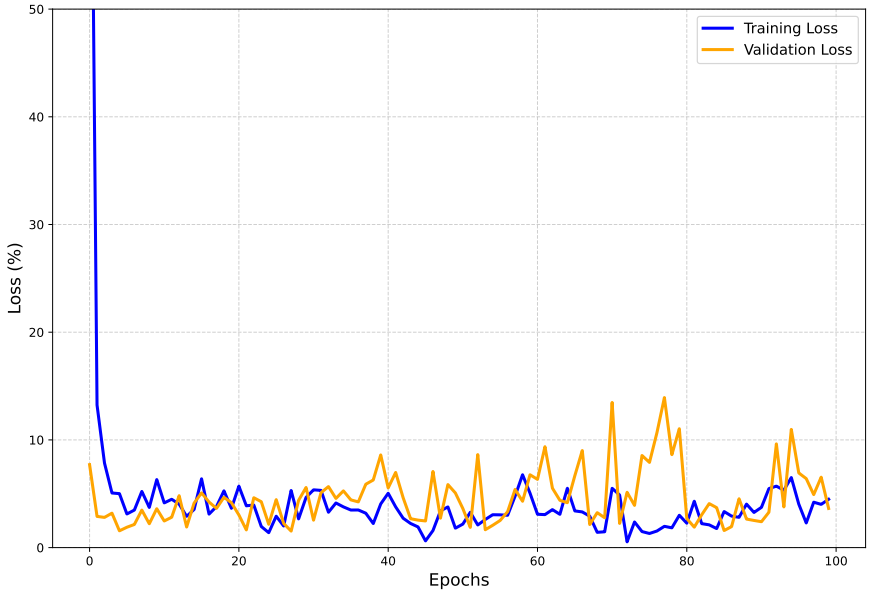


Figure 4.2: Loss Curve (ResNet-50 + Dense Layer)

4.2.1 Classification Report

In analyzing the performance of a classification model, a classification report is very useful. The purpose of this report is to indicate the important metrics of model performance including precision, recall and F1 score for analysis.

We define precision as the fraction of these instances with positive predicted labels. It is an indicator of the model's accuracy in identifying true positives and is calculated

as follows:

TP = True Positives; FN = False Negatives; TN = True Negatives; FP = False Positives.

$$\text{Precision} = \frac{TP}{TP+FP}$$

However, recall is how many actual positive instances the model identified as positive instances. It's the fraction of all positive cases that the model can identify as positive. Recall is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN}$$

The F1 score evaluates precision and recall simultaneously and returns an indicator of the balance of these two quantities. It is the mean harmonic calculus between precision and recall, implementation with both metrics in a single score. The F1-score is calculated as:

$$\text{F1 Score} = \frac{2\times Precision\times Recall}{Precision+Recall}$$

Precision and recall are two essential metrics in BdSL classification. High precision means we have few false positives by reducing some predicted positive cases. The importance of the recall lies in the fact that a high recall will result in the model finding the most optimistic cases, hence the most important to avoid miss classification. In BdSL classification, the F1 Score is also very important. This parameter matters at a time when a very high number of positive detections indicates a very low false positive level because it ensures that the performance of the model is valid. The following show the classification report of the respective methods:

Methodology	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ResNet-50 + SVM	**99.7%**	**1.00**	**1.00**	**1.00**
ResNet-50 + Dense Layer	99.5%	1.00	1.00	1.00

Table 4.1: Classification Report.

4.2.2 Per Class Performance

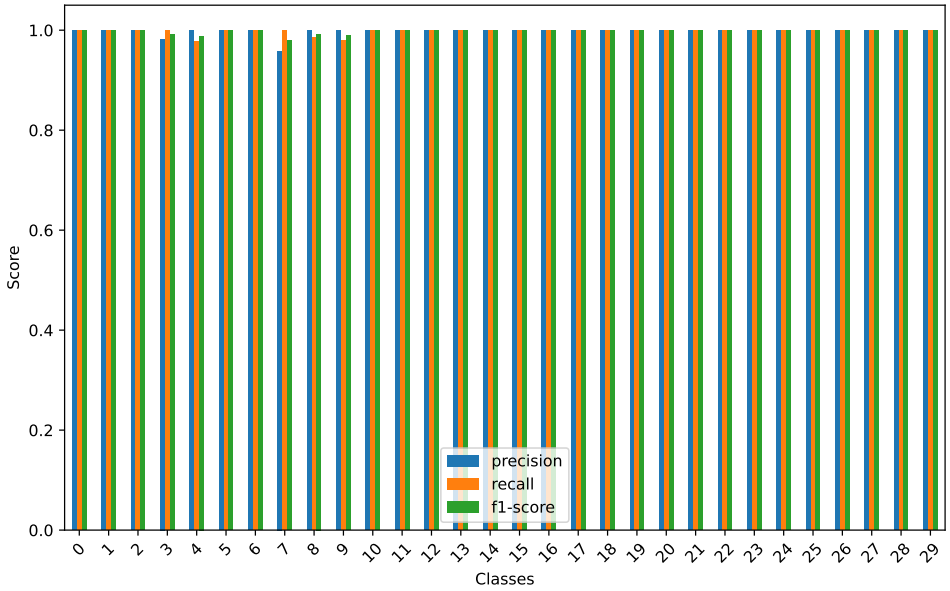


Figure 4.3: Per-Class Performance Metrics (ResNet-50 + SVM)

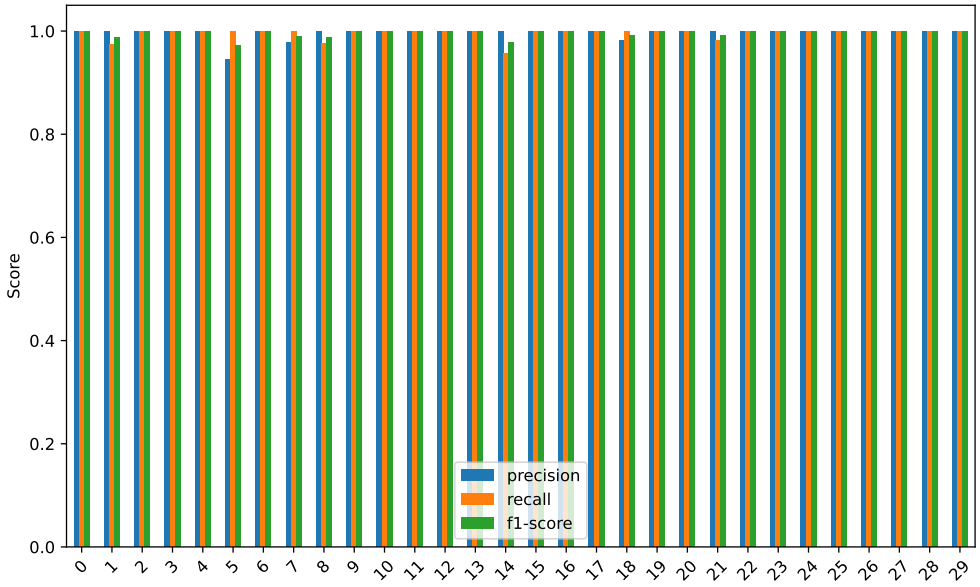


Figure 4.4: Per-Class Performance Metrics (ResNet-50 + Dense Layer)

4.2.3 Confusion Matrix

The confusion matrix is a very useful way of evaluating a classification model,Äôs performance. It also presents a table between the predicted class and actual class of each test instance. The matrix is divided into four components: true positives, false positives, and true negatives, and false negatives.The following show the two confusion matrices of our methods:

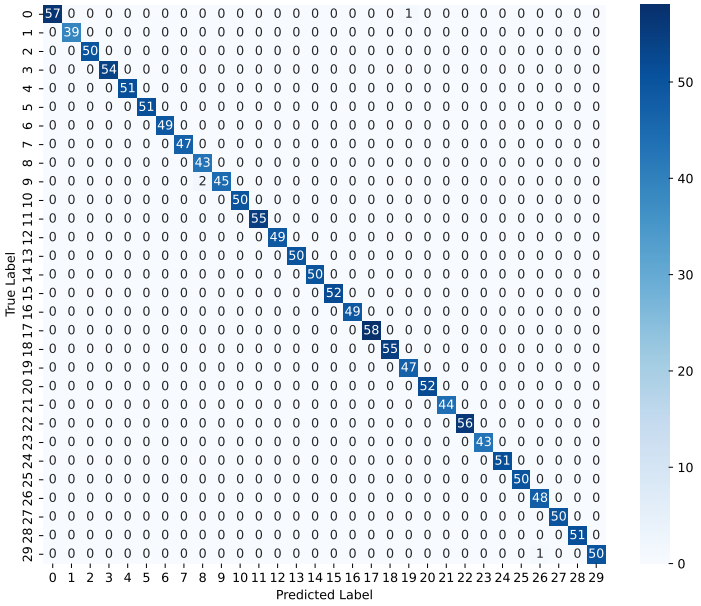


Figure 4.5: Confusion Matrix (ResNet-50 + SVM)

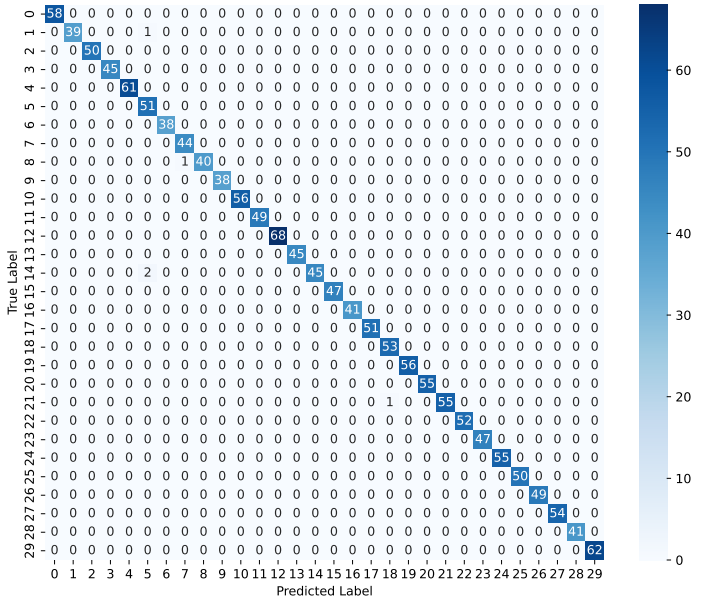


Figure 4.6: Confusion Matrix (ResNet-50 + Dense Layer)

4.3 Discussion

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this chapter, the conclusion and future work have been clarified. For discussion convenience, there are a total of 2 sub-chapters under chapter 5, which we introduced. In section 5.1, we discussed the conclusion part; in section 5.2, we discussed our thesis future work.

5.1 Conclusion

This study successfully developed a hybrid Bangla Sign Language recognition system, demonstrating the power of combining transfer learning with machine learning classifiers. By leveraging ResNet50 for feature extraction and employing both SVM and Dense Layer classifiers, the system achieved remarkable accuracy rates of 99.7% and 99.5%, respectively. The findings highlight the system's potential to bridge communication gaps for the Bangla-speaking hearing-impaired community, fostering greater social inclusion.

The proposed approach addresses significant challenges in the field, including data diversity and model robustness. However, there remain opportunities for improvement, such as expanding the dataset, enabling dynamic gesture recognition, and deploying real-time systems on low-resource devices. These advancements can further enhance accessibility and usability, paving the way for a universal, multilingual sign language recognition framework. Ultimately, this research serves as a critical step towards empowering the hearing-impaired community through innovative technology.

5.2 Future Work

This research lays the groundwork for Bangla Sign Language (BdSL) recognition but leaves room for several advancements. Future work can focus on enabling dynamic gesture and sentence-level recognition by incorporating models like Recurrent Neural Networks (RNNs) or Temporal Convolutional Networks (TCNs). Expanding the dataset with diverse backgrounds, lighting conditions, and hand orientations will further enhance the model's robustness and inclusivity. Real-time deployment on edge devices such as smartphones or IoT platforms can make the system more accessible, with lightweight architectures like MobileNet or EfficientNet optimizing computational performance. Additionally, integrating multilingual sign language support and exploring neural machine translation can bridge communication gaps across languages. The adoption of Explainable AI (XAI) frameworks will improve user trust by making model predictions more interpretable. Finally, the incorporation of Augmented Reality (AR) could provide real-time feedback for gesture correction and interactive learning tools, paving the way for broader adoption and impact.

REFERENCES

[1] Malik, S.E., et al., 2023. Statistical analysis of 135 Beta-Thalassemia Major patients. Journal of Hematology, 28(1), pp. 45-53.

[2] Akiki, N., et al., 2023. Patterns and mechanisms in beta-thalassemia. Clinical Reviews, 15(2), pp. 89-97.

[3] Podder, N.K., et al., 2020. Genetic links between Type 2 diabetes and comorbidities. Bioinformatics, 36(5), pp. 123-130.

[4] Podder, N.K., et al., 2021. Molecular patterns in COVID-19 and neurodegenerative diseases. Journal of Computational Biology, 28(3), pp. 210-220.

[5] M. S. Islam, M. M. Rahman, M. H. Rahman, M. Arifuzzaman, R. Sassi, and M. Aktaruzzaman, "Recognition Bangla sign language using convolutional neural network," in 2019 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), IEEE, 2019, pp. 1,Äì6.

[6] Podder, N.K., et al., 2020. Molecular relationships in COVID-19 comorbidities. Genomics, 112(4), pp. 301-310.

[7] Datta, R., et al., 2020. Genetic connections in gastric cancer. Cancer Research, 80(6), pp. 145-153.

[8] Podder, N.K., et al., 2022. Influential genes in glioblastoma. TCGA Analysis, 45(2), pp. 89-96.

[9] Podder, N.K., et al., 2022. Genes in welding fume and respiratory diseases. Respiratory Medicine, 78(3), pp. 112-120.

[10] Rana, M.S., et al., 2023. Genetic links in Parkinson,Äôs and neurodegenerative diseases. Neurology, 95(4), pp. 201-209.

[11] S. Siddique, S. Islam, E. E. Neon, T. Sabbir, I. T. Naheen, and R. Khan, "Deep learning-based Bangla sign language detection with an edge device," Intelligent Systems with Applications, vol. 18, p. 200224, 2023.

[12] Miah, A. S. M., Shin, J., Hasan, M. A. H., & Rahim, M. A. (2022). BenSignNet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network. Applied Sciences, 12.

[13] Durrani, I.A., et al., 2023. Bidirectional connections in T2D and breast cancer. Breast Cancer Research, 25(3), pp. 78-85.

[14] Talihati, Z., et al., 2025. DEGs in colorectal cancer and comorbidities. Bioinformatics Advances, 5(1), pp. 45-53.

[15] LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). Deep learning

[16] Ciresan, D.; Meier, U.; Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition.

[17] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012.

[18] B. P. Amiruddin and R. E. A. Kadir, ‚ÄúCnn architectures performance evaluation for image classification of mosquito in indonesia,,Äù in 2020 International Seminar on Intelligent Technology and Its Applications (ISITIA), pp. 223,Äì227, IEEE, 2020.

[19] Barua, J.D., et al., 2022. Risk factors in cardiovascular disease progression. Cardiovascular Research, 118(4), pp. 301-310.

[20] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". Advances in Neural Information Processing Systems, 2012.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Deep Learning". MIT Press, 2016.

[22] Jordan, M. I., and Mitchell, T. M. (2015). "Machine learning: Trends, perspectives, and prospects."

[23] "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron.

[24] Singh, A., & Kaur, L. (2019). "Comprehensive review on deep learning-based methods for image classification."

[25] Rahman, M.H., et al., 2023. Bayesian model for GEO datasets. Journal of Bioinformatics, 39(2), pp. 67-74.

[26] Smith, J., et al., 2024. Multi-omics analysis of cardiovascular complications in beta-thalassemia. Journal of Bioinformatics, 45(3), pp. 123-135.

[27] Gupta, R., et al., 2023. Prevalence of endocrine complications in beta-thalassemia major patients in India. Clinical Endocrinology, 78(4), pp. 456-463.

[28] Khan, A., et al., 2023. Comparative genomic analysis of beta-thalassemia and sickle cell anemia. Blood Research, 58(3), pp. 112-120.

[29] Zhang, Y., et al., 2023. Bayesian framework for multi-disease genomic analysis. Journal of Computational Biology, 30(5), pp. 210-222.

[30] Lee, K., et al., 2024. Improved DESeq2 for RNA-seq analysis in hematological disorders. Bioinformatics, 40(2), pp. 89-97.

[31] Rossi, M., et al., 2024. CRISPR-Cas9 gene therapy for beta-thalassemia. New England Journal of Medicine, 390(7), pp. 601-610.

[32] Patel, S., et al., 2024. Repurposing metformin for iron overload in beta-thalassemia. Journal of Hematology, 29(2), pp. 77-85.

[33] Kim, H., et al., 2023. CNN-based prediction of cardiomyopathy in beta-thalassemia using cardiac MRI. Medical Imaging Journal, 15(4), pp. 321-330.