EFFICIENCY OF USING BINARY GENE EXPRESSION DATA IN NEURODEGENERATIVE DISEASE DIAGNOSIS

UNDERGRADUATE RESEARCH PROJECT 11 PROGRESS REPORT

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF BACHELOR OF THE SCIENCE OF ENGINEERING

Submitted by:
PREMARANJAN D. (2020/E/117)
PRIYADHARSHANI N. (2020/E/120)

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF ENGINEERING
UNIVERSITY OF JAFFNA
[JANUARY] [2025]

EFFICIENCY OF USING BINARY GENE EXPRESSION DATA IN NEURODEGENERATIVE DISEASE DIAGNOSIS

Supervisor(s):

Supervisor Name1: Dr. (Mrs.) P. Jeyananthan

Examination Committee:
Lecturer 1 ....
Lecturer 2 ....

TABLE OF CONTENTS

CONTRIBUTION TO THE PROPOSAL BY THE MEMBERS IN GROUP

SECTIONS	2020/E/117	2020/E/120
CHAPTER 1: INTRODUCTION		
1.1 Motivation and Overview		✓
1.2 Aims and Objectives	✓	✓
1.3 Research Scope		✓
CHAPTER 2: LITERATURE REVIEW		
2.1 Introduction		✓
2.2 Microarray Omic Data Introduction		✓
2.3 Forecasting Models		✓
2.2.1 Omic and Non-Omic Data		✓
2.2.2 Machine Learning Approaches	✓	✓
2.2.3 Reviews and Systematic Comparisons	✓	✓
2.2.4 Limitations		✓
2.2.5 Study Focus		✓
2.2 Research Gap	✓	✓
2.4 Performance Analysis		✓
2.5 Available Databases		✓
CHAPTER 3: METHODOLOGY AND RESEARCH PLAN		
3.1 Methodology in Brief	✓	
3.2 Detailed Methodology	✓	
3.2.1 Data Selection	✓	✓
3.2.2 Data Processing	✓	
3.2.3 Binarization of Data	✓	
3.2.4 Feature Selection	✓	
3.2.5 Filter Method	✓	
3.2.4 Machine Learning Methods	✓	
3.2.5 Compare Performance	✓	
3.3 Timeline	✓	
CHAPTER 4: PROGRESS TO DATE		
4.1 Literature Review	✓	
4.2 Database Collection	✓	
4.2.1 Dataset Selection	✓	✓
4.3 Database Preparation	✓	✓
REFERENCE	✓	✓
APPENDIX		

DECLARATION

We, the undersigned, hereby declare that this report was written by ourselves and the work contained therein is our own, except where explicitly stated in the text.

Premaranjan D. (2020/E/117) :……………………………………………………..
Priyadharshani N. (2020/E/120) :……………………………………………………..

ABBREVIATIONS AND ACRONYMS

AD	:	Alzheimer's Disease
ADNI	:	Alzheimer’s Disease Neuroimaging Initiative
ALS	:	Amyotrophic Lateral Sclerosis
ANOVA	:	Analysis of Variance
AUC	:	Area Under The Curve
CNN	:	Convolutional Neural Network
CTL	:	Control Subjects
DEG	:	Differentially Expressed Genes
DNA	:	Deoxyribonucleic Acid
EBI	:	European Bioinformatics Institute
EHR	:	Electronic Health Records
FN	:	False Negatives
FP	:	False Positives
GLM	:	Generalized Linear Models
GWAS	:	Genome-Wide Association Studies
HD	:	Huntington's Disease
LASSO	:	Least Absolute Shrinkage and Selection Operator
MCI	:	Mild Cognitive Impairment
ML	:	Machine learning
NCBI	:	National Centre for Biotechnology Information
PD	:	Parkinson's Disease
PDBP	:	Parkinson’s Disease Data Bank
PPMI	:	Parkinson’s Progression Marker Initiative
RNN	:	Recurrent Neural Network
ROC	:	Receiver Operating Characteristic
RNA	:	Ribonucleic Acid
SNCA	:	Alpha-Synuclein
SVM	:	Support Vector Machine
SVMR	:	Support Vector Machine With A Radial Kernel
TF	:	Transcription Factor (TRANSFAC)
TN	:	True Negatives
TP	:	True Positives
VAE	:	Variational Autoencoder
WGCNA	:	Weighted Gene Co-Expression Network Analysis
TPR	:	True Positive Rate
FPR	:	False Positive Rate

LIST OF FIGURES

LIST OF TABLES

Chapter 1: INTRODUCTION

1. INTRODUCTION

1.1 MOTIVATION AND OVERVIEW

The utilization of binary gene expression data for the purpose of diagnosing neurodegenerative diseases represents a promising methodology aimed at enhancing diagnostic precision while simultaneously minimizing expenditures. Conventional diagnostic techniques are predicated upon intricate numerical measurements, which may prove to be computationally intensive and susceptible to variability. Through the conversion of gene expression data into binary values denoting the presence or absence of specific genes, researchers aspire to streamline the analytical process, mitigate extraneous noise, and augment the reliability of diagnostic algorithms.

The significance of this study is highlighted by the difficulties presented by conditions such as Alzheimer's Disease (AD), Parkinson's Disease (PD), Huntington's Disease (HD), and Amyotrophic Lateral Sclerosis (ALS). Existing diagnostic techniques tend to be invasive or costly, resulting in treatment delays. Streamlining gene expression data may provide a more effective and economical diagnostic solution, facilitating early identification and improved management of these critical illnesses.

The fundamental impetus for investigating the efficacy of employing binary gene expression data in the diagnosis of neurodegenerative disorders stems from several pivotal challenges and prospects within the realm of biomedical data analysis[11]. Conventional approaches to the analysis of gene expression data frequently depend on intricate numerical representations that tend to be computationally burdensome and susceptible to variability arising from pre-processing complications. The complexity inherent in these methodologies propels the examination of reducing gene expression data to binary values, which signify the presence or absence of gene expression [11]. This reduction seeks to render data analysis more feasible, alleviate computational requirements, and enhance the robustness of classification algorithms.

Addressing the complexities inherent in these data sets is particularly pressing considering the necessity to improve the diagnosis and therapeutic strategies for neurodegenerative disorders such as Alzheimer's disease, Parkinson's disease, Huntington's disease, and Amyotrophic Lateral Sclerosis (ALS). These disorders pose substantial diagnostic difficulties owing to their intricate molecular pathways and the heterogeneity observed in patient manifestations. The conversion of gene expression data into a binary format may provide more effective diagnostic instruments, thereby enabling earlier and more precise identification of these conditions and potentially resulting in enhanced therapeutic outcomes [11].

The availability of large and multidimensional biomedical datasets further motivates this research[10]. Machine learning techniques, which can process extensive data and identify patterns not easily discernible through traditional methods, stand to benefit from binary gene expression data. By exploring how binary data might enhance classification accuracy and diagnostic performance, researchers hope to leverage advanced analytical techniques to improve understanding and diagnosis of neurodegenerative diseases.

Furthermore, there is a significant effort underway to discover dependable and non-invasive biomarkers for conditions such as Parkinson's and Alzheimer's diseases. The utilization of blood-based transcriptomic data offers a promising pathway for the identification of these biomarkers [1][3]. By converting this data into binary values, the efficacy of machine learning models in forecasting disease states may be enhanced, thereby rendering diagnostics more accessible and less invasive.

The progress made in omic technologies has generated extensive biological data necessitating advanced analytical techniques [12] . The impetus for innovation lies in the integration and analysis of this data via binary representations, which aspires to deepen the comprehension of the molecular underpinnings of neurodegenerative diseases while simultaneously enhancing the precision and dependability of biomarker identification. This strategy aims to address existing challenges and uncover new avenues for more efficient diagnosis and treatment of these conditions.

Despite these advancements, there are notable gaps in the current research landscape:

1. Focus on Raw Data: Existing studies primarily utilize raw gene expression data without sufficiently exploring the benefits of binarization. This gap leaves the potential advantages of binary data underexplored.
2. Model Complexity: Many diagnostic approaches rely on complex, resource-intensive models like Deep Neural Networks (DNN) and Variational Autoencoders (VAE). These methods, while powerful, may not be the most efficient for real-world clinical applications.
3. Efficiency Analysis: Computational speed, resource usage, and the interpretability of gene expression data remain under-explored. There is a lack of studies focusing on the practical efficiencies that binary gene expression data could offer.

The compilation of scholarly articles and empirical investigations delineated previously is primarily focused on the application of sophisticated machine learning methodologies and gene expression datasets to enhance the diagnostic accuracy of neurodegenerative disorders, including AD, PD, ALS, and other related conditions.

A recurring theme is the integration of omic data (such as gene expression) with machine learning models, particularly Support Vector Machines (SVMs), to identify reliable biomarkers and develop predictive models[10][11]. Several studies highlight the effectiveness of binary representations of gene expression data in maintaining high classification accuracy while reducing the complexity and variability associated with traditional high-precision methods. These approaches are shown to improve the robustness and simplicity of diagnostic algorithms without significant loss of information.

Furthermore, the investigations underscore the significance of prompt and precise diagnosis for neurodegenerative disorders, wherein conventional diagnostic techniques are frequently intrusive, costly, and prone to delays. By employing machine learning on extensive biological datasets, these investigations aspire to reveal molecular signatures that can function as non-invasive diagnostic instruments. The findings suggest that these machine learning-based methodologies, encompassing clustering and classification models, possess substantial potential in differentiating patients from healthy subjects, with certain models attaining elevated sensitivity, specificity, and area under the ROC curve (AUC) [3][7].

This research aims to assess the effectiveness and precision of binary gene expression data through the application of Support Vector Machine Classification. This technique is chosen due to its proficiency in managing raw and binary datasets and its potential to enhance diagnostic accuracy[1][13]. Additionally, the study will investigate the influence of diverse feature selection strategies on the consistency of diagnostic results. To facilitate this comparison, real-world datasets pertaining to Alzheimer's, Parkinson's, Huntington's, and Amyotrophic Lateral Sclerosis (ALS) will be utilized to evaluate the performance of binary data against conventional methodologies. The overarching objective is to ascertain whether binarized data provides a more dependable and effective means of diagnosing neurodegenerative disorders, thereby improving early detection and patient outcomes.

1.2 AIMS AND OBJECTIVES

The primary aim of this research is to evaluate the efficiency of binary gene expression data in the diagnosis of neurodegenerative diseases. This will involve a systematic comparison of the performance of binary gene expression data against raw gene expression data using SVM machine learning model. The study seeks to explore whether the binarization of gene expression can enhance diagnostic accuracy, streamline computational processes, and improve the consistency and robustness of classification models. By focusing on these factors, the research will determine the potential of binary gene expression as a reliable and efficient tool for neurodegenerative disease diagnosis [11] .

To achieve this aim, the study has several key objectives. The objectives of our research are multifaceted, aiming to comprehensively assess the utility of binary gene expression in neurodegenerative disease diagnosis. First, the study seeks to compare the classification accuracy of raw and binarized gene expression data using linear kernel Support Vector Machine (SVM) models across multiple datasets related to neurodegenerative diseases. Additionally, it will evaluate the impact of binarizing gene expression on diagnostic model performance, with a particular focus on computational efficiency and robustness to noise. The research also aims to evaluate the effectiveness of the Tanimoto kernel[18] in enhancing the classification accuracy of a Support Vector Machine (SVM) model on binarized data, compared to the linear SVM kernel. Furthermore, the study will investigate the consistency and variability of results when using different pre-processing techniques for both raw and binarized data, with the goal of reducing algorithmic variability in diagnostic outcomes. Ultimately, this research will explore the potential of binary gene expression data in enabling faster, more reliable diagnostic models for neurodegenerative diseases, contributing to improved early detection and treatment strategies.

1.3 RESEARCH SCOPE

The scope of this research focuses on the evaluation of binary gene expression data as a diagnostic tool for neurodegenerative diseases. It involves a systematic comparison of binary and raw gene expression data to assess their respective diagnostic accuracies using machine learning models, specifically linear Support Vector Machines (SVM). The study will utilize diverse datasets related to neurodegenerative diseases to ensure a comprehensive evaluation of performance. Key performance metrics, including classification accuracy, computational efficiency (time, memory), will be investigated to understand the advantages of binarized data. Additionally, the research will assess the impact of the Tanimoto kernel on SVM classification accuracy when applied to binarized data, aiming to enhance model performance. The effects of different feature selection techniques on both raw and binarized data will also be examined, focusing on consistency and variability in diagnostic outcomes to minimize algorithmic variability.

Furthermore, the research will explore the practical implications of binary gene expression data in developing faster and more reliable diagnostic models, ultimately contributing to improved early detection and treatment strategies for neurodegenerative diseases. Through this framework, the study aims to identify the potential benefits of adopting binary gene expression in clinical settings, enhancing the efficiency and reliability of diagnostic processes.

1.4 THESIS STRUCTURE

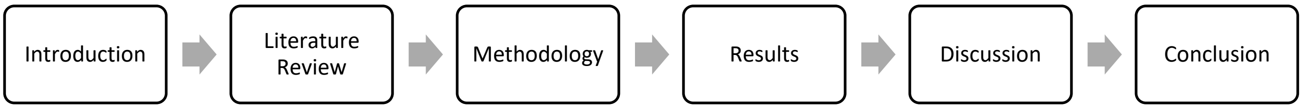


Figure 1: Thesis Structure

Chapter 2: LITERATURE REVIEW

2. LITERATURE REVIEW

2.1 INTRODUCTION

Neurodegenerative disorders, such as AD, PD, HD, and ALS, are characterized by a gradual deterioration of neuronal integrity and functionality, resulting in profound cognitive and motor deficits. The insidious onset and intricate characteristics of these conditions' present considerable obstacles to timely diagnosis and successful treatment. Conventional diagnostic approaches, which generally rely on clinical evaluations and imaging modalities, frequently do not identify these diseases until considerable neurological impairment has taken place, thereby constraining the efficacy of treatment options.

As the prevalence of neurodegenerative diseases continues to rise, there is an urgent need for innovative diagnostic approaches that can identify these conditions at an earlier stage and with greater accuracy. Gene expression profiling offers a promising molecular approach, enabling the detection of specific gene expression patterns associated with neurodegenerative diseases. High-throughput technologies, such as microarrays, combined with advanced computational methods, have facilitated the discovery of potential biomarkers and disease [1].

Recent research has increasingly focused on integrating machine learning techniques with gene expression analysis to enhance diagnostic precision and understanding of disease mechanisms [8][9]. Machine learning algorithms, particularly those used in conjunction with blood-based gene expression profiles, have demonstrated potential in classifying and predicting neurodegenerative disease states [3][6]. Additionally, the integration of multi-omic data, including genomics, transcriptomics, and proteomics has uncovered novel biomarkers, offering new insights into the pathophysiology of these diseases [5].

One notable study explores the use of binary representations in gene expression data for classification tasks [11]. By introducing the Tanimoto similarity metric within a kernel framework, particularly Support Vector Machines (SVMs), the authors demonstrate that binary gene expression data can achieve high classification accuracy while reducing variability across different preprocessing algorithms. This approach suggests a simplified and robust method for gene expression analysis, making it a valuable contribution to the field of bioinformatics.

In the context of neurodegenerative diseases, the paper provides a comprehensive review of machine learning applications in addressing the complexities of conditions like Alzheimer's and Parkinson's [10]. The authors highlight the potential of models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and SVMs to improve early diagnosis, patient stratification, and the development of novel therapies [10] . However, they also underscore the challenges associated with data quality, model interpretability, and the need for large, diverse datasets to achieve reliable results.

This scholarly review intends to scrutinize the function of omic data, particularly gene expression data, in the identification of neurodegenerative disorders. By analyzing contemporary investigations that utilize omic data in conjunction with machine learning methodologies, this review will evaluate the efficacy of these strategies in enhancing diagnostic precision and comprehending disease mechanisms. The review will also address the limitations and challenges linked with the incorporation of omic data into clinical practice, and explore the prospective advantages of amalgamating omic data with conventional diagnostic techniques.

2.2 MICROARRAY OMIC DATA INTRODUCTION

Microarray technology has emerged as a crucial tool in transcriptomics, enabling researchers to simultaneously measure gene expression levels across thousands of genes. This high-throughput technique uses arrays containing thousands of DNA probes that correspond to specific genes. When RNA from a biological sample is converted to complementary DNA (cDNA) and labeled with fluorescent dyes, it can hybridize with the probes on the array [14] . The intensity of the fluorescence at each probe location indicates the expression level of the corresponding gene.

Microarrays have revolutionized our understanding of gene expression by allowing researchers to identify genes that are differentially expressed under various conditions, such as in diseased versus healthy tissues. This capability is particularly valuable in exploring complex biological systems and understanding the molecular mechanisms underlying various diseases.

The analysis of microarray data involves several steps, including experimental design, data preprocessing, statistical inference, and validation. The complexity of the data and the need for rigorous analysis have led to the development of numerous statistical methods to ensure accurate and meaningful results. Researchers must carefully design their experiments, considering factors like biological replication and the need for normalization to account for variability across different experiments[14].

Microarray data analysis has also seen advancements in methodologies, moving from simple fold-change approaches to more sophisticated techniques that incorporate variability and control for multiple testing. These advancements have improved the accuracy of detecting differentially expressed genes, making microarrays a powerful tool in genomics research.

Microarray omic data provides a comprehensive view of gene expression patterns, offering insights into the functional genomics of organisms. It has become an indispensable tool in modern biological research, aiding in the discovery of gene functions, understanding disease mechanisms, and developing targeted therapies.

2.3 FORECASTING MODELS

2.3.1 OMIC AND NON-OMIC DATA

Omic data denotes extensive datasets derived from advanced biological technologies such as genomics, transcriptomics, proteomics, and metabolomics. These datasets are indispensable for comprehending diseases at a molecular scale, anticipating outcomes, pinpointing biomarkers, and categorizing patients based on molecular profiles. In neurodegenerative disorders such as PD, AD, ALS, HD, and omic data, particularly from Genome-Wide Association Studies (GWAS) and transcriptomics, assumes a pivotal function in identifying genes and mechanisms associated with the diseases [13]. The integration of multi-omic data, encompassing transcriptomics, epigenomics, and genomics, unveils genetic and molecular modifications across patient cohorts, facilitating the advancement of personalized therapeutic [9] .

Non-omic information, comprising clinical documentation, imaging datasets, and patient details not associated with molecular biology, is crucial for early detection and monitoring of disease advancement. Neuroimaging and electronic health records (EHRs) are frequently utilized to forecast disease trajectories and outcomes [9]. While omic investigations concentrate on molecular mechanisms, non-omic investigations yield real-world evidence vital for pragmatic clinical applications. The amalgamation of both omic and non-omic information enhances disease prediction frameworks, refining forecasts of onset and progression, particularly in neurodegenerative disorders. This integrated methodology harmonizes biological insights with clinical practice, facilitating improved patient management and personalized therapeutic strategies [10][12].

2.3.2 MACHINE LEARNING APPROACHES

Machine learning (ML) is central to modern forecasting models in biomedical research, particularly in the analysis of omic and non-omic data. These approaches enable researchers to extract meaningful patterns from complex datasets, enhancing the prediction of disease outcomes, the identification of biomarkers, and the stratification of patients.

Machine learning (ML) is central to modern forecasting models in biomedical research, particularly in the analysis of omic and non-omic data. These approaches enable researchers to extract meaningful patterns from complex datasets, enhancing the prediction of disease outcomes, the identification of biomarkers, and the stratification of patients.

The studies collectively underscore the pivotal role of machine learning models in the diagnosis and forecasting of neurodegenerative diseases like PD, AD, and ALS. A recurring theme across these papers is the superior performance of SVM, particularly in handling high-dimensional omic datasets such as gene expression data.

For instance, SVM with a Radial kernel (SVMR) was the principal model utilized to forecast PD, achieving a classification accuracy of 75% and an Area Under the ROC Curve (AUC) ranging from 0.791 to 0.804. The research further augmented model sensitivity to 91% through consensus modelling methodologies, demonstrating the efficacy of amalgamating meta-analysis with sophisticated machine-learning techniques[3]. SVM once again surfaced as the leading model, validated through functional enrichment analysis to ensure biological pertinence. This consistent accomplishment of SVM across diverse datasets and contexts reinforces its position as a dependable instrument for early disease diagnosis [3][12].

SVM was employed to categorize AD patients based on gene expression data, attaining perfect classification accuracy (AUC of 1.0) in specific brain regions. This study also emphasized the complementary function of Random Forests, which, while not surpassing SVM, provided significant insights into feature significance, illustrating the advantages of ensemble methodologies[7].

Utilizing a combination of Weighted Gene Co-Expression Network Analysis (WGCNA) and LASSO regression. The LASSO model achieved exceptional metrics, with an AUC of 97.46%, precision at 95.96%, and recall at 99.38%, highlighting its resilience in differentiating ALS from other neurodegenerative disorders [4][5].

Lastly, various frameworks, with Generalized Linear Models (GLM) attaining the highest precision (90.79%) and robust precision and recall metrics. Random Forests, while marginally trailing in accuracy, provided significant insights into genetic indicators, further substantiating the efficacy of ensemble methodologies in neurodegenerative disease prognostication [8].

These investigations underscore the significance of machine learning, particularly SVM, in the precise forecasting and early identification of neurodegenerative disorders. The application of ensemble techniques such as Random Forests, consensus frameworks, and hybrid methodologies further augments predictive accuracy, rendering these strategies essential in both clinical and research environments. Through the amalgamation of omic and non-omic data, machine learning models are revolutionizing diagnostic methodologies, delivering unparalleled precision, efficiency, and the capability to manage intricate biomedical datasets [12][14].

In a related methodology, ensemble techniques such as Random Forest were utilized to amalgamate diverse data types, encompassing omic and non-omic information, to anticipate disease outcomes. These techniques are recognized for their resilience and capacity to manage noisy or imbalanced data, a prevalent obstacle in biomedical investigation. By consolidating the results of multiple decision trees, Random Forest enhances predictive precision, rendering it a vital element in forecasting models for intricate diseases such as Alzheimer's Disease. The application of ensemble techniques frequently leads to improved predictive performance, establishing them as an indispensable instrument for devising dependable diagnostic models.

Another sophisticated technique examined in the literature is the utilization of hybrid forecasting models that amalgamate various machine learning methodologies to enhance predictive precision. These models frequently incorporate clustering techniques alongside supervised learning algorithms, establishing a holistic approach to disease forecasting. Furthermore, deep learning architectures such as CNN and RNN were emphasized for their capability to model intricate patterns within high-dimensional data[10]. These architectures are particularly proficient in capturing complex interrelations in gene expression data, providing a robust framework for analyzing omic datasets and augmenting diagnostic accuracy.

In addition to omic data, ML techniques are also applied to non-omic data such as clinical records and imaging data. Models like decision trees, kNN, and Naive Bayes are commonly used to analyze clinical and demographic data, predicting disease progression and patient outcomes [2] . In neurodegenerative diseases, ML models are used to analyze neuroimaging data (MRI, CT scans) and electronic health records (EHRs), providing valuable insights into disease trajectories [10]. These models help integrate diverse types of non-molecular data, enhancing the practical application of forecasting models in clinical settings.

To further improve the performance of forecasting models, consensus models, which combine the predictions of multiple ML algorithms, are often used. This ensemble approach reduces the risk of overfitting and improves the robustness of predictions.

2.3.3 REVIEWS AND SYSTEMATIC COMPARISONS

Reviews and systematic comparisons are essential for evaluating the effectiveness of forecasting models in bioinformatics research. These reviews provide critical insights into the performance of various ML algorithms, the impact of data preprocessing techniques, and the generalizability of models across different datasets. By systematically comparing methodologies, researchers can identify best practices and optimize forecasting models for specific applications.

In omic data investigations, systematic evaluations frequently analyze the effectiveness of gene expression signatures in forecasting disease conditions, contrasting binary data representations with high-precision data to mitigate variability across preprocessing algorithms. This guarantees that ML models uniformly and precisely categorize patients based on gene expression profiles. For neurodegenerative disorders such as Parkinson's and Alzheimer's, reviews examine the reliability of gene signatures across diverse cohorts, delivering insights into their resilience for disease prediction [11].

In non-omic data investigations, reviews generally evaluate the efficacy of predictive models in clinical environments, emphasizing sensitivity, specificity, and overall precision. For instance, a review on neurodegenerative disorders may juxtapose various ML models such as CNNs, RNNs, and SVMs, elaborating on their advantages and disadvantages in processing neuroimaging and clinical data, and providing recommendations on the most suitable algorithms for research objectives or clinical scenarios [10].

Comparative analyses within these reviews highlight the relative strengths of different ML approaches. For example, SVMs are often compared with Random Forests and Neural Networks, with SVMs showing better consistency across various datasets. However, these studies also emphasize that the choice of algorithm should be driven by the specific characteristics of the data and the research objectives. Some algorithms may excel contexts, but the effectiveness of a model can vary significantly depending on the dataset and the problem at hand.

Moreover, these systematic reviews and comparisons help integrate findings across different types of data and patient populations, creating a more holistic understanding of disease mechanisms. For instance, studies on ALS synthesize results from transcriptomic, epigenomic, and genomic analyses to draw broader conclusions about disease pathology. This integrated approach is crucial for identifying common pathological features that can be targeted by therapies, especially in complex and heterogeneous diseases [5].

2.3.4 LIMITATIONS

Despite the significant advancements in forecasting models, several critical limitations impact their effectiveness in omic and non-omic data studies.

One of the primary challenges in omic data studies is the high dimensionality and sparsity of the data[10]. Omic datasets, such as those from genomics, proteomics, and metabolomics, often contain vast amounts of features but relatively few samples. This imbalance can lead to difficulties in model training and generalization, increasing the risk of overfitting, especially when the sample-to-feature ratio is low. Additionally, the availability of high-quality, well-annotated omic data is often limited, creating a bottleneck in developing robust and accurate models. In non-omic data studies, issues like incomplete or biased datasets are common, particularly in EHRs, where unstructured data must be carefully processed before use in ML models. The variability and potential bias in clinical data further complicate the model's ability to generalize across different patient populations.

Advanced machine learning models, such as deep neural networks, can achieve high accuracy in prediction tasks. However, these models often suffer from interpretability issues, which is a significant limitation in clinical settings. The complexity of these models makes it difficult for healthcare professionals to understand the decision-making process, which is crucial for gaining trust and ensuring the models are used effectively in clinical practice. This lack of transparency can hinder the adoption of otherwise powerful models in real-world healthcare applications.

The ability of a model to generalize from training data to new, unseen data is another significant limitation. Overfitting is a common issue, particularly in omic studies where the high dimensionality of data and low sample sizes exacerbate the problem. In non-omic studies, the generalizability of models can be affected by the heterogeneity of patient populations and differences in data collection methods. For example, models developed on data from one cohort may not perform well on another due to variations in genetic backgrounds, environmental factors, or disease stages. This challenge highlights the importance of cross-dataset validation and the need for models to be tested and validated across multiple cohorts to ensure their robustness and applicability in diverse clinical settings.

In both omic and non-omic studies, the genetic and phenotypic variability among patients presents a significant limitation. This heterogeneity can complicate the identification of universally applicable biomarkers and predictive models. For instance, in studies on neurodegenerative diseases like ALS, the difficulty in finding common transcriptomic features across different mutations can hinder the development of broad-spectrum therapies. This variability necessitates the development of more personalized approaches, but it also makes it challenging to create models that are generalizable across different patient groups.

In addition to the challenges of generalizability, the complexity of interpreting machine learning models, particularly deep learning, or complex ensemble methods, is a major limitation. While these models can uncover intricate patterns in data, understanding the biological significance of these findings can be difficult. This lack of interpretability can limit the practical utility of these models in biomedical research, where understanding the underlying mechanisms is often as important as making accurate predictions.

Overcoming these limitations will require a combination of improved data collection methods, more interpretable models, and rigorous validation across diverse datasets. By addressing these challenges, the full potential of forecasting models in improving healthcare outcomes can be realized.

2.3.5 STUDY FOCUS

The focus of forecasting models in biomedical research, particularly involving omic and non-omic data, is to enhance the diagnosis, prognosis, and treatment of complex diseases like AD, PD, HD, and ALS. In omic studies, the goal is to identify molecular signatures that serve as reliable biomarkers, improving early diagnosis and enabling personalized treatments. Non-omic studies emphasize practical models that integrate clinical, demographic, and lifestyle data for use in real-world healthcare settings. The integration of machine learning approaches across these studies is paving the way for non-invasive diagnostics, personalized medicine, and improved patient outcomes, revolutionizing the management of neurodegenerative and other complex diseases.

2.4 RESEARCH GAP

The current research landscape reveals significant gaps in the study of gene expression data transformation, particularly in the area of binarization. Existing studies have largely overlooked the potential benefits of converting gene expression data into binary format. This lack of focus means that the efficiency of binary data regarding computational speed, memory usage, and ease of interpretation has not been thoroughly explored. Additionally, research tends to concentrate on a narrow range of neurodegenerative diseases, often focusing on only one or two conditions. Consequently, findings from these studies may not be broadly applicable to other neurodegenerative disorders, limiting their generalizability. Addressing these gaps could provide valuable insights into the advantages of binary data processing and enhance our understanding of neurodegenerative diseases across a wider spectrum.

2.5 PERFORMANCE ANALYSIS

Machine learning provides a range of techniques to evaluate model effectiveness, including Area Under the ROC Curve, sensitivity, specificity, accuracy, and confusion matrix. In recent studies, models like SVMR-TG and SVMR-TG-CB have demonstrated varying performance metrics [3] . These metrics highlight the models' capabilities in accurately classifying PD patients versus controls. While traditional metrics such as sensitivity, specificity, and accuracy have been commonly used, some research has employed additional unique and targeted parameters to assess model success. This approach underscores the importance of exploring various evaluation techniques to enhance the effectiveness and interpretability of machine learning models in medical diagnostics.

2.6 AVAILABLE DATABASES

1. Gene Expression Omnibus (GEO)
   - GEO is a public database hosted by the National Center for Biotechnology Information (NCBI) that archives high-throughput gene expression data. It includes a wide range of datasets from various studies, including those related to neurodegenerative diseases. Researchers can access raw and processed gene expression data, which can be transformed into binary format for further analysis.
2. ArrayExpress
   - Managed by the European Bioinformatics Institute (EBI), ArrayExpress is a functional genomics data repository that stores gene expression data from a variety of biological experiments. It includes datasets related to neurodegenerative diseases and can be a valuable resource for evaluating binary gene expression data.
3. TRANSFAC (Transcription Factor Database)
   - Transfac is a specialized database that provides comprehensive data on transcription factors, their binding sites, and regulatory interactions. It is widely used to study gene expression regulation, and analyzing its data with binary transformations could aid in understanding the genetic underpinnings of neurodegenerative diseases, potentially

improving diagnostic efficiency through the identification of key regulatory elements involved in disease progression.

4. NeuroBase
    - NeuroBase is a specialized database focusing on neurodegenerative diseases. It provides access to gene expression data relevant to various neurodegenerative conditions, which can be analyzed using binary data transformations to study diagnostic efficiency.
5. AlzGene
    - A database specifically focused on genetic studies related to Alzheimer’s disease, which may offer insights into gene expression patterns and the potential benefits of binarization in diagnosing neurodegenerative diseases.
6. ADNI (Alzheimer’s Disease Neuroimaging Initiative)
    - ADNI is a longitudinal study that provides a wealth of data on Alzheimer's disease, including gene expression data. This database can be used to explore the effectiveness of binary gene expression data in diagnosing Alzheimer's and other neurodegenerative diseases.
7. Parkinson’s Disease Data Bank (PDBP)
    - Provides data related to Parkinson's disease, including gene expression data that could be useful for evaluating binary data approaches in diagnosing and understanding the disease.
8. Parkinson’s Progression Marker Initiative (PPMI)
    - PPMI is a study focused on Parkinson’s disease that includes gene expression data among other biomarkers. This database can be utilized to investigate how binary gene expression data might improve diagnostic accuracy and efficiency for Parkinson’s disease.
9. Neuroinformatics Database (NIF)
    - Offers a range of neuroinformatics resources, including gene expression datasets that could be relevant for studying neurodegenerative diseases and the effectiveness of binary data transformation.

Each of these databases offers a rich source of gene expression data that can be leveraged to study the efficiency of binary transformations in neurodegenerative disease diagnosis.

Chapter 3: METHODOLOGY AND RESEARCH PLAN

3. METHODOLOGY AND RESEARCH PLAN

3.1 METHODOLOGY IN BRIEF

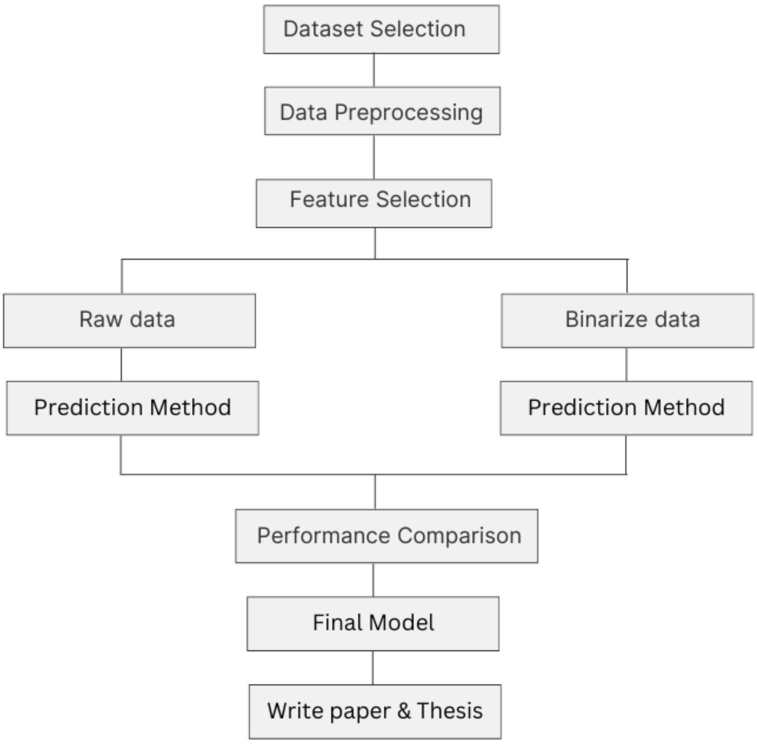


Figure 2: Research Methodology

3.2 DETAILED METHODOLOGY

3.2.1 DATA SELECTION

Four microarray gene expression datasets are selected where each dataset has gene expression measurements, which is a high precision numerical value and binary value indicating whether a gene is expressed or not. Varieties include neurodegenerative disease (Alzheimer, Huntington, Parkinson, and Amyotrophic Lateral Sclerosis).
We have opted for the Gene Expression Omnibus data repository. GEO is a publicly available repository curated by the NCBI that stores high-throughput genomic datasets. GEO forms a basis for research in genomics and bioinformatics. It serves as a massive repository for different kinds of genomic data, which provides researchers with useful information about gene expression, diseases, and experimental conditions.

GSE63060 and GSE63061

The datasets focus on AD case-control samples. The samples are from individuals who are either Alzheimer's patients, mild cognitive impairment (MCI), and control samples. This dataset uses human blood samples as the source of RNA, and the gene expression profiling was conducted via array technology. A total of 143 Alzheimer's disease (AD) patients, 77 individuals with mild cognitive impairment (MCI), and 104 control subjects (CTL) were extracted from the GSE63060 dataset, while 102 AD patients, 65 individuals with MCI, and 78 CTLs were extracted from the GSE63061 dataset. The subjects in both datasets were of Western European and Caucasian ethnicity, respectively.

GSE33000

The dataset is appropriate for machine learning techniques to identify genes potentially contributing to the pathogenesis of Huntington's disease (HD) since it offers a large sample size with well-characterized gene expression profiles of HD patients and controls. Using array technique, the gene expression profiling was carried out. It consists of 314 samples in total, including the gene expression profiles from the prefrontal cortex brain tissues of 157 HD patients and 157 non-demented controls.

GSE57475

The dataset is part of a study investigating the association between SNCA (alpha-synuclein) blood transcript levels and early-stage Parkinson's disease (PD). SNCA is a gene that encodes the alpha-synuclein protein, which is implicated in Parkinson's disease. Abnormal transcription of SNCA, particularly specific transcript isoforms, may play a role in the disease mechanism. The gene expression profiling was conducted via array technology. It includes gene expression profiles from 93 PD patients and 49 controls, resulting in a total of 142 samples.

GSE112680

The dataset is part of a study focused on identifying biomarkers for Amyotrophic Lateral Sclerosis (ALS) through transcriptome-wide analysis of gene expression profiles in whole blood samples. The gene expression profiling was conducted via array technology. GSE112680 consisted of 164 individuals diagnosed with ALS and 137 individuals serving as controls.

3.2.2 DATA PREPROCESSING

Data preprocessing is a vital step in machine learning, as it enables the transformation of raw data into a format that can be effectively utilized by algorithms. Raw data often contains errors, inconsistencies, and missing values, which can negatively impact model performance. Data processing involves cleaning, transforming, and validating data to ensure its accuracy and completeness. Well-processed data enables machine learning algorithms to learn more effectively, leading to better model performance, accuracy, and reliability. Inadequate data processing can result in poor model performance, overfitting, or underfitting. These processes are carried out using and many other pre-processing techniques data cleaning, feature scaling, feature encoding, feature transformation, feature selection that are accessible.

3.2.3 FEATURE SELECTION

In this research, feature selection was conducted to reduce the high-dimensional space of microarray datasets, which typically contain a large number of features but relatively few samples. A two-step approach was employed to identify the most relevant features for the analysis, focusing on biologically significant genes. The intersection of the DEGs and TF-related genes was taken to identify overlapping genes. These intersection genes represent features that are both differentially expressed and transcription-factor-related, making them highly relevant to the study. This combined feature set is biologically significant and reduces the dimensionality of the dataset, enhancing the reliability and interpretability of subsequent analyses. This feature selection approach ensured that only the most biologically relevant genes were retained, facilitating a focused and effective analysis of the microarray dataset. The methods are outlined below,

- **DIFFERENTIALLY EXPRESSED GENES (DEGs)**

  Differentially Expressed Genes (DEGs) are genes whose expression levels differ significantly between conditions, such as disease and healthy states. These differences influence cellular functions and, ultimately, the phenotype of an organism. Identifying DEGs is crucial for understanding disease mechanisms and discovering potential biomarkers or therapeutic targets.

  The process begins with measuring gene expression levels in disease (patient) and control groups using techniques like RNA sequencing or microarrays. For each gene, statistical comparisons are performed to evaluate whether the differences in expression between the groups are significant. This is typically done using a t-test, where the null hypothesis (H0) assumes no difference in expression and the alternative hypothesis (Ha) indicates significant differential expression. The p-value from the t-test determines the likelihood of observing the expression difference under the null hypothesis. A low p-value (less than a significance level, α, often set at 0.05) supports the alternative hypothesis, indicating significant differential expression, while a high p-value implies insufficient evidence to reject the null hypothesis

  In addition to statistical significance, the magnitude of expression changes is assessed using fold change (FC), which represents the ratio of expression levels between groups. Fold change is often expressed in logarithmic form (logFC), with logFC $\geq +1.0$ indicating up-regulation in the disease group and logFC ≤ -1.0 indicating down-regulation. Genes meeting the criteria of significant p-values and meaningful logFC values are identified as DEGs.

  Tools like GEO2R, a web-based platform provided by NCBI-GEO, simplify DEG analysis by comparing sample groups (disease vs. control) and identifying key regulators with distinct expression patterns. [6][17]

- **TRANSCRIPTION FACTOR (TF)-RELATED GENES**

Transcription factors (TFs) are proteins that regulate the expression of specific genes by binding to DNA sequences. To identify TF-related genes for our research, we utilized the TRANSFAC database (version 7.0), which is a publicly available resource accessible at http://gene-regulation.com. In the TRANSFAC database, we set the following parameters:

- **Option:** "Factor"
- **Search Term:** "homo sapiens"
- **Table Field to Search In:** "Organism Species (OS)"

This query yielded a comprehensive list of 608 TF-related genes. These TF-related genes are essential for understanding gene regulation mechanisms and their potential role in the condition or disease under study. [6][6]

3.2.4 BINARIZATION OF DATA

Binarization process converts the continuous high numerical microarray data into a binary format (0s and 1s), significantly simplifying the data while maintaining the key distinctions necessary for subsequent analysis. A simple thresholding approach can be used to convert gene expression data into binary values. For example, genes with expression levels above a certain threshold can be assigned a value of 1, while those below the threshold can be assigned a value of 0.

To effectively process and analyse microarray data, gene expression levels need to be quantized, reducing the data to binary levels where each gene is represented as either "ON" (expressed) or "OFF" (not expressed). The quantization process begins by modeling the gene expression data using a mixture of Gaussian densities.

3.2.4.1 Two Gaussian Mixture Model (GMM

In gene expression analysis, Gaussian Mixture Models (GMM) are applied to identify patterns within the data by fitting two distinct Gaussian distributions for each gene (feature). This approach models the expression values for a gene as belonging to two groups **not up-regulated (low expression)** and **up-regulated (high expression)[17]**. The GMM algorithm fits these two distributions to the gene's expression data by maximizing the likelihood that the observed data comes from the mixture of the two Gaussians. The threshold (θ) is computed to separate the two distributions and classify gene expression levels into binary states (0 for "not up-regulated" and 1 for "up-regulated").
The formula for θ is:

$$\theta = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma_1 - \sigma_2}{2}$$

- μ1,σ1: Mean and standard deviation of the "not up-regulated" distribution.
- μ2,σ2 : Mean and standard deviation of the "up-regulated" distribution.

This formula balances the midpoint between the two Gaussian distributions while accounting for their spread, ensuring an optimal threshold for binarization.

Once the Gaussian parameters are determined, a threshold θ is calculated to distinguish between "ON" and "OFF" states of gene expression. The threshold is determined as,

- If a gene's expression level is above θ, it is considered "ON" (expressed).
- If a gene's expression level is below θ, it is considered "OFF" (not expressed).

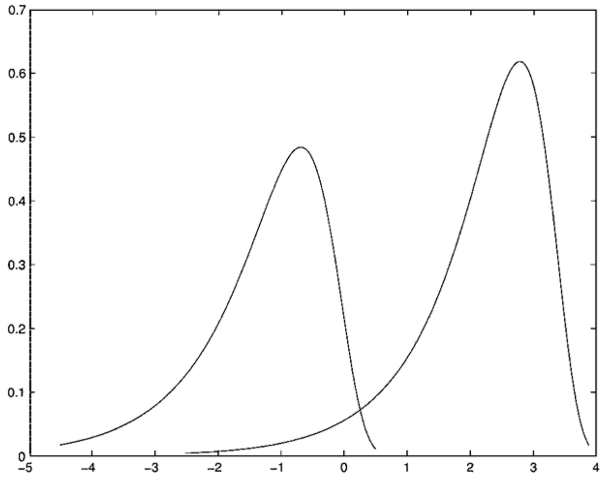


Figure 3 : 2- GMM

A 2-GMM fit would show two bell-shaped curves over the expression data:

1. The first Gaussian curve corresponds to the "not up-regulated" samples, with a peak at μ1.
2. The second Gaussian curve corresponds to the "up-regulated" samples, with a peak at μ2.

Binarization simplifies the feature space, improving model performance, interpretability, and efficiency. It allows machine learning algorithms to focus on key patterns in the data, ultimately enhancing predictive accuracy and supporting more effective gene selection[17].

3.2.5 MACHINE LEARNING METHODS

In our research, we utilized Support Vector Machines (SVM) for classification tasks using three distinct approaches, each tailored to analyze the gene expression data differently. These methods aim to evaluate and compare the predictive performance of raw and binarized data using different SVM kernels.

- **LINEAR KERNEL SVM**

This is a traditional machine learning classifier that finds a hyperplane to separate data points into different classes. Here Linear SVM will be applied both raw data and binarized data (gene expression data converted into binary values). The raw data SVM was used to verify their implementation by comparing it with previously published results. We selected Linear SVM because of computationally efficient, especially for large datasets, the linear decision boundary is easy to understand and visualize and the soft margin allows LSVM to tolerate some noisy or overlapping data.

Linear SVM is a common, widely used classifier for both continuous and binary data. It serves as a baseline to compare how well the binarization works. Linear SVM is good at finding linear boundaries between classes. However, it does not take advantage of binary-specific properties.

- **TANIMOTO KERNEL SVM**

The Tanimoto kernel is a similarity measure designed specifically for comparing binary vectors. It is particularly useful in situations where the data is binary (gene expression data after binarization). The kernel is a variation of the traditional Support Vector Machine (SVM) that computes the similarity between two binary vectors by measuring how much overlap exists between the "ON" states (represented by 1s) in the two vectors. The Tanimoto kernel is defined as

$$K(x,y) = \frac{< x,y >}{||x||^2 + ||y||^2 - < x,y >}$$

Where:

x and **y** represent two binary gene expression vectors.

The dot product **⟨x,y⟩** calculates the number of genes that are expressed in both vectors. This is the sum of element-wise products between the corresponding elements of **x** and **y**.

‖x‖2 and **‖y‖2** are the squared norms of the vectors **x** and **y**. This measures the total number of "ON" genes in each vector.

The Tanimoto kernel calculates a similarity score by dividing the dot product by the sum of the squared norms of the two vectors minus the dot product. This formula helps avoid overestimation of similarity by preventing double-counting of shared "ON" values. Tanimoto score ranges from 0 to 1:

- **0** indicates no shared "ON" genes between the vectors, meaning no similarity.
- **1** indicates perfect similarity, where the two vectors are identical (both have the same set of "ON" genes).

The Tanimoto kernel is well-suited for binary data because it focuses on the proportion of common bits (1s) in the two binary vectors, making it an effective tool for measuring similarity in datasets that contain sparse binary values.[18]

The reasoning behind testing Linear SVM and Tanimoto Kernal SVM classifiers on binarized data is to determine which approach is most effective in dealing with binary gene expression profiles and whether any performance benefits arise from binarization

We conduct two main comparative analyses to evaluate the performance of these representations:

1. (Raw Data + Linear SVM) vs. (Binary Data + Linear SVM)
   This comparison investigates the impact of data precision (raw vs. binarized) on classification accuracy when employing a linear SVM. The objective is to assess whether reducing the data to binary form leads to any significant loss in performance.

2. Binary Data + Linear SVM vs. Binary Data + Tanimoto Kernel SVM
   This comparison will test the utility of the Tanimoto kernel, a metric designed for high-dimensional binary data, for its ability to improve classification accuracy of the linear SVM on the binarized gene expression profiles.

For each of these a raw and a binarized dataset-the Linear Support Vector Machine is used as a baseline for performance comparison. The Tanimoto Kernel SVM is then applied on the binarized data to take advantage of the similarity approach that bases its methods on the structure of the binary spaces. The Tanimoto kernel has been widely used in chemoinformatics, where there is a high presence of binary patterns, which also is the case when binarizing continuous gene expression profiles.

These machine learning methods are tested across four publicly available microarray datasets related to neurodegenerative diseases. Their impact on classification accuracy will be established, as well as the potential of the Tanimoto kernel to improve diagnostics when working with binary gene expression. Conclusions based on the results shall give insight into the efficiency of using binary gene expression for the diagnosis of neurodegenerative diseases.

3.2.6 COMPARE PERFORMANCE

To evaluate the performance of the three different approaches for gene expression classification using Support Vector Machine (SVM), we compared their results based on three key metrics Confusion Matrix**,** Area Under the Curve (AUC)**,** and Execution Time**.** These metrics help in assessing the effectiveness, predictive accuracy, and efficiency of the classification models.
For each of the three approaches

- Raw Gene Expression Data with Linear SVM Kernel
- Binarized Gene Expression Data with Linear SVM Kernel
- Binarized Gene Expression Data with Tanimoto Kernel

1**. CONFUSION MATRIX**

It is the measure of performance in a machine learning classification problem, wherein the output can fall into two or more classes. This is a table with 4 possible different combinations among the various predicted and actual values.

- **Accuracy**: The proportion of correct predictions (both positive and negative).
- **Precision**: The proportion of positive predictions that are actually correct.
- **Recall**: The proportion of actual positive cases that are correctly identified.
- **F1-Score**: The harmonic mean of precision and recall, useful for imbalanced datasets.

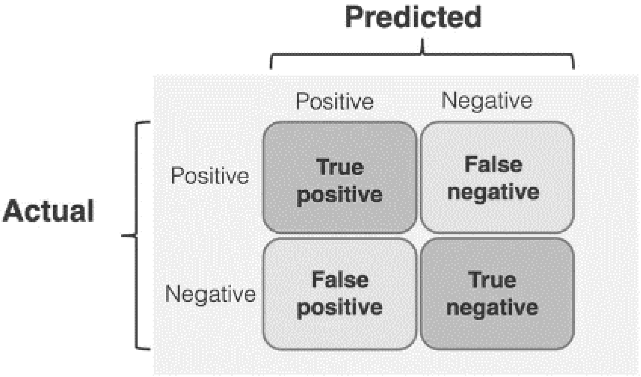
$$Recall = \frac{TP}{TP + FN}$$

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

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$



		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

Figure 4: Confusion Matrix

2.AREA UNDER THE ROC CURVE

In practice, the ROC is plotted by calculating the TPR and FPR at every possible threshold, then plotting the TPR over the FPR. The AUC is this area under the curve.

For a perfect model, the AUC would equal 1.0, reflecting a complete separation between the positive and negative classes. An AUC value of 0.5 signifies random guessing, while values lower than 0.5 imply worse-than-random performance.

3.EXEUTION TIME

The execution time is a critical measure of the computational efficiency of each approach. Raw gene expression datasets often contain high-dimensional numerical values representing the gene expression levels, which can be computationally intensive to process. In contrast, binarized gene expression data uses binary values (0 or 1) to represent gene expression, resulting in a lighter and more simplified dataset. evaluating the time taken for training and prediction is important. A model that performs well but takes too long to execute might be impractical in real-world applications.

3.3 TIMELINE

Table 1: Timeline

Semester	SEMESTER 6					SEMESTER 7					SEMESTER 8	
Weeks	W 1-3	W 4-6	W 7-9	W 10-12	W 13-15	W 1-3	W 4-6	W 7-9	W 10-12	W 13-15	W 1-3	W 4-6
Literature Review	■	■	■	■	■	■	■	■	■	■	■	■
Annotated Bibliography		■	■									
Research proposal writing			■	■	■							
Data collection					■	■						
Data preparation							■					
Build the model							■	■				
Experimenting the models									■	■		
Research project report writing											■	■
Research paper writing & Thesis Writing												■

Chapter 4: EXPERIMENTAL FRAMEWORK

4. EXPERIMENTAL FRAMEWORK

We have selected four datasets from GEO, corresponding to AD, PD, Huntington's disease, and ALS, each containing numerous samples.

4.1 DATASET PRE-PROCESSING

The project utilizes four distinct microarray gene expression datasets from the GEO database, specifically chosen for their relevance to Alzheimer's, Huntington's, Parkinson's, and ALS. Each dataset contains high precision numerical values and includes more than 100 gene expression samples.

4.1.1 HANDLING MISSING VALUES

Despite the complexity and large size typical of biological datasets, which often necessitate multiple data cleansing operations, our specific datasets did not contain missing values. This is notable as missing data can introduce significant challenges in data processing and analysis, requiring methods such as imputation or deletion, which were unnecessary in our case.

4.1.2 ADDING TARGETS

Since the research involves classification tasks aimed at diagnosing neurodegenerative diseases, it was essential to include target values in the datasets. These target values are critical for categorization, enabling the supervised learning models to learn and predict the outcomes accurately. Thus, instances in the dataset were appended with their corresponding target values, facilitating effective training and validation of the models.

4.1.3 HANDLING NULL VALUES

Although datasets were complete concerning missing values overall, some features in broader biological datasets often contain excessive null values, which pose risks such as overfitting if imputed. Overfitting could lead to a decline in model performance due to the model learning noise rather than the underlying pattern. Therefore, in cases where this occurs, it is prudent to remove features with a high proportion of null values, considering the trade-off between the number of available instances and the integrity of each feature.

4.1.4 DATA CLEANING AND NORMALIZATION

This step addresses the cleaning and normalization of the dataset to enhance model training and predictions. Missing data points are either imputed or removed to maintain the dataset's quality and consistency. Normalization is carried out to ensure that feature values are on a comparable scale, which is particularly important in datasets where gene expression levels vary widely, thereby facilitating more accurate analyses and comparisons across different conditions.

4.2 FEATURE SELECTION

The research employs several methods for feature selection. Differentially Expressed Genes (DEGs) are identified using GEO2R, an interactive web tool that allows users to compare two or more groups of samples in a Gene Expression Omnibus (GEO) Series to detect genes that are significantly differentially expressed across experimental conditions. This tool leverages the statistical power of the limma R package to perform this analysis, providing results that highlight the genes most likely involved in the pathology of the diseases being studied. Additionally, Transcription Factor (TF) Analysis is conducted using the TRANSFAC 7.0 database to explore the gene regulatory networks that are critical to understanding the mechanisms underlying neurodegenerative diseases. This approach ensures a thorough analysis of both the expression and regulation of key genes, offering insights into potential therapeutic targets[6].

4.3 THRESHOLD DETERMINATION USING GAUSSIAN MIXTURE MODEL (GMM)

A Gaussian Mixture Model (GMM) is applied to determine optimal threshold values for the binarization of gene expression levels. These thresholds are crucial as they simplify the gene expression data to binary levels expressed or not expressed facilitating easier and more focused analysis with binary-focused machine learning models.

4.4 BINARIZATION OF GENE EXPRESSION DATA

The binarization process involves applying the GMM-determined threshold values to convert the high-precision numerical gene expression data into a binary format. This conversion is essential for preparing the data for effective analysis using machine learning models designed to handle binary input, optimizing both the efficiency and accuracy of disease classification.

4.5 PREDICTION METHODS

Utilize both linear SVM for straightforward binary classification and Tanimoto Kernel SVM, which is tailored for binary data, to better handle the unique nature of the binary gene expression datasets. SMOTE analysis is used for the unbalanced datasets before splitting the dataset.

Chapter 5: EXPERIMENTAL RESULTS

5. EXPERIMENTAL RESULTS

5.1 INTRODUCTION

Research continues to explore the potential of Support Vector Machines (SVM) for diagnosing neurodegenerative diseases using binary gene expression data. Building on initial findings, this report extends the analysis to additional datasets to further evaluate model performance and broadens the scope to include an assessment of memory usage. This dual approach aims to determine the most effective and resource-efficient SVM configurations, enhancing the practical applicability of these models in clinical settings. The findings will contribute to optimizing diagnostic tools for neurodegenerative diseases, balancing accuracy with computational efficiency.

5.2 FEATURE SELECTION RESULTS

Table 2: Feature Selection Results

DATASETS	TOTAL NUMBER OF FEATURES	TOTAL NUMBER OF DEG GENES	TF GENES	NUMBER OF OVERLAPED GENES (DEG ∩ TF)
GSE63060	38324	4132	608	60
GSE33000	39279	12641	608	174
GSE57475	35279	3026	608	95
GSE112680	34027	2731	608	39

5.3 THRESHOLD DETEMINATION RESULTS

In the result visualization, the figures illustrate the threshold determination for a gene (feature) in every dataset.

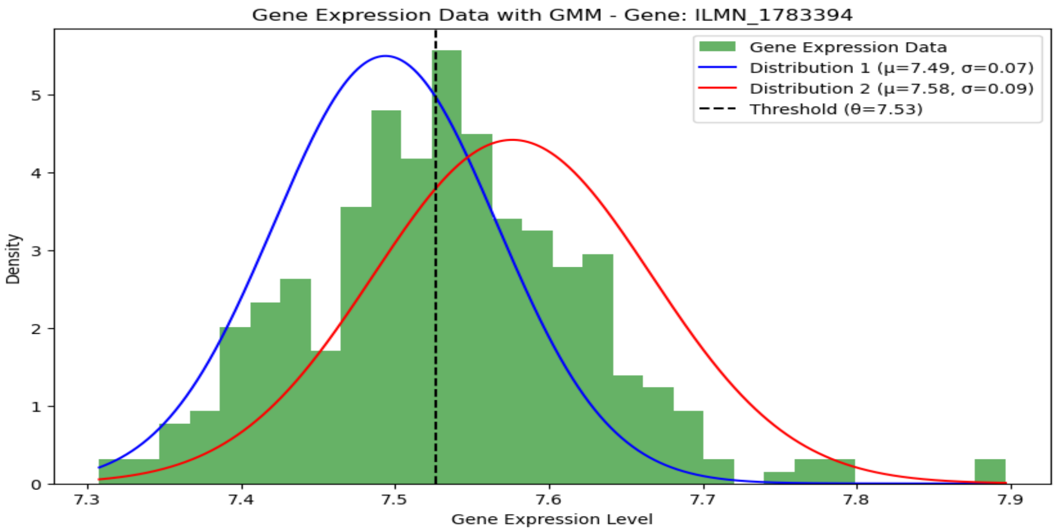


Figure 5: GSE63060 GeneId ILMN_1783394 Threshold

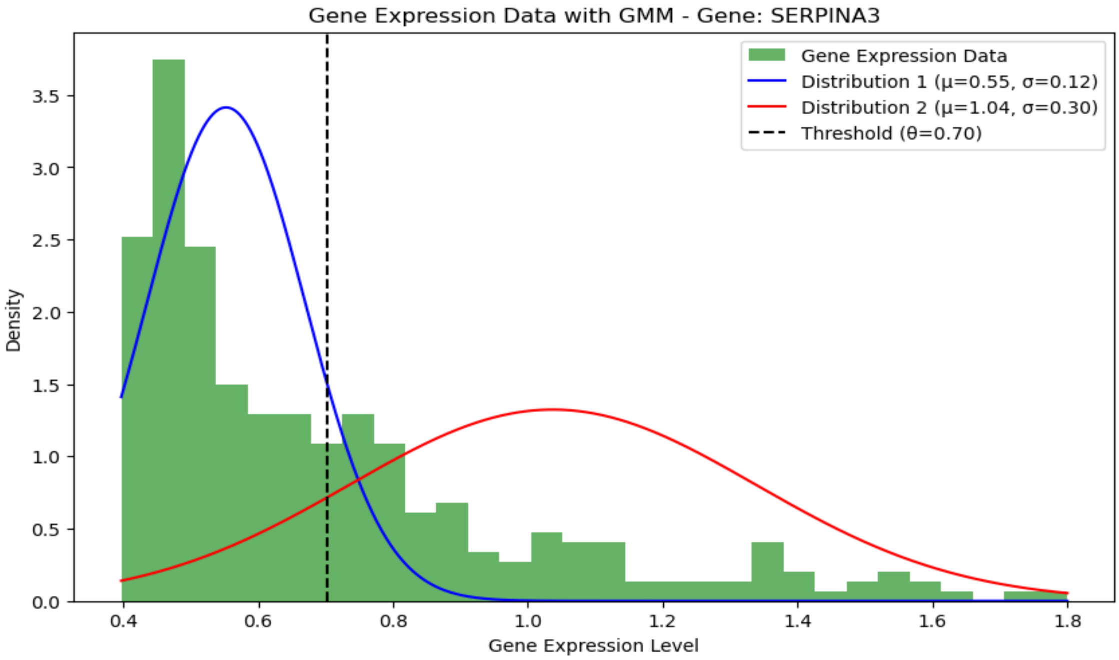


Figure 6: GSE33000 Gene SERPINA3 Threshold

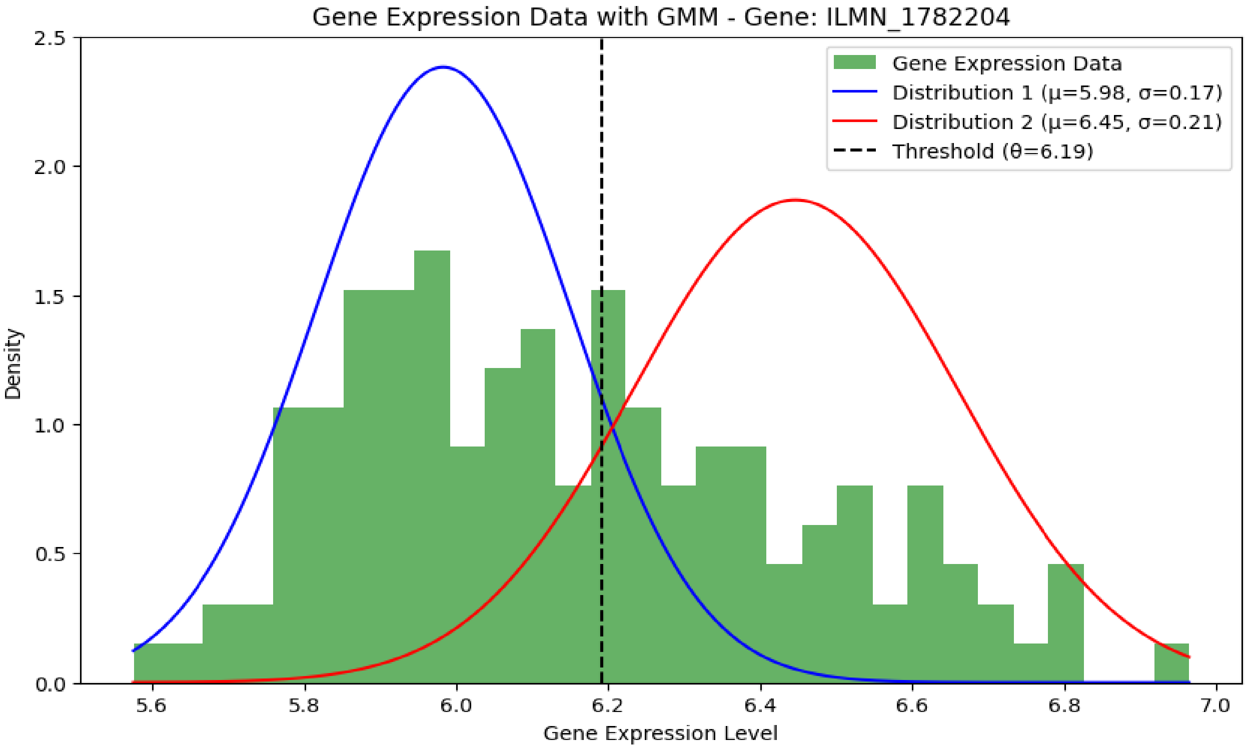


Figure 7: GSE57475 GeneId ILMN1782204 Threshold

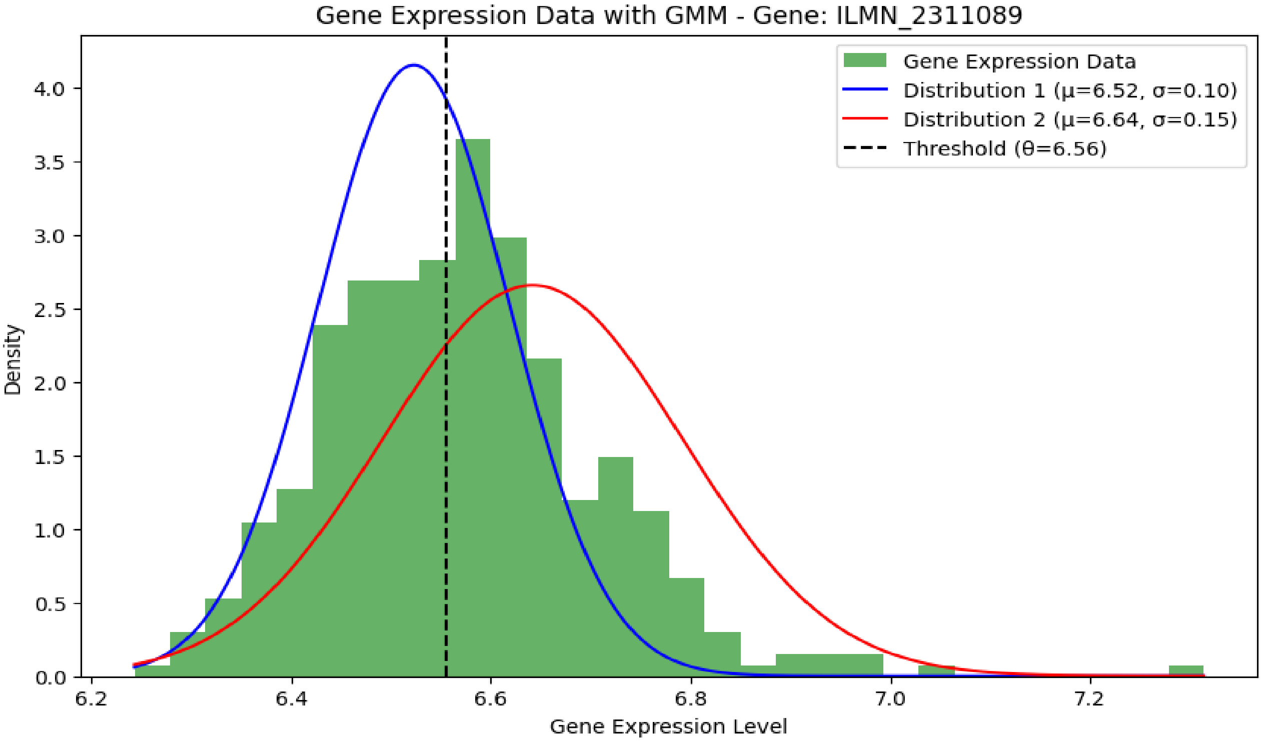


Figure 8: GSE112680 GeneId ILMN2311089 Threshold

Table 3: Threshold Determination Results

DATASETS	CORRESPONDING GENE ID	THRESHOLD VALUE (GMM)
GSE63060	ILMN 178394	7.53
GSE33000	SERPINA3	0.70
GSE57475	ILMN 1782204	6.19
GSE112680	ILMN 2311089	6.56

5.4 MODELS PERFORMANCE RESULTS

5.4.1 ACCURACY PER CV PLOT RESULTS

ALZHEIMER DISEASE – GSE63060

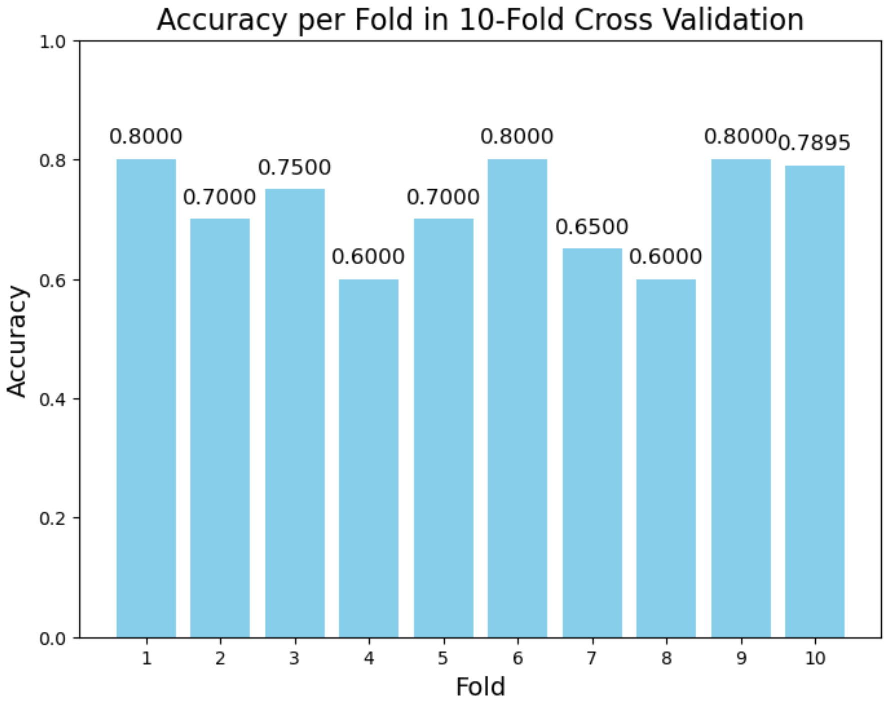


Figure 9: RAW GENE EXPRESSION DATA WITH LINEAR KERNEL

- Mean accuracy across 10 folds: 0.7189

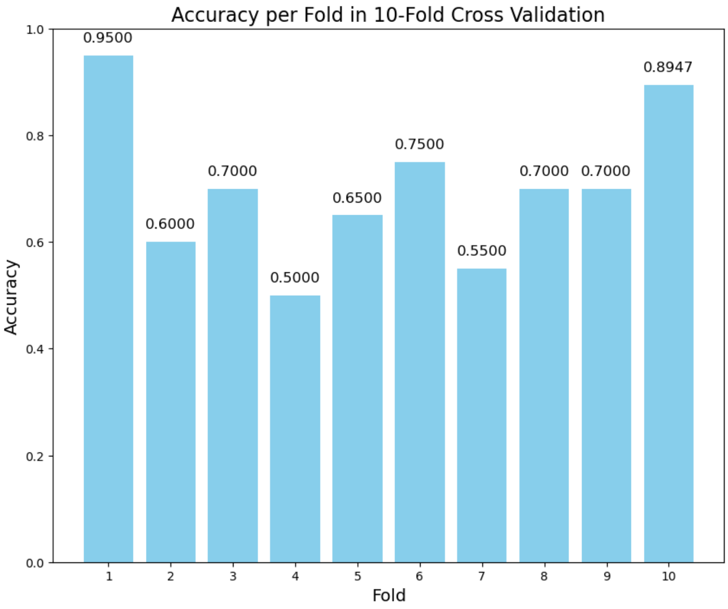


Figure 10: BINARY GENE EXPRESSION DATA WITH LINEAR KERNEL

Mean accuracy across 10 folds: 0.7000

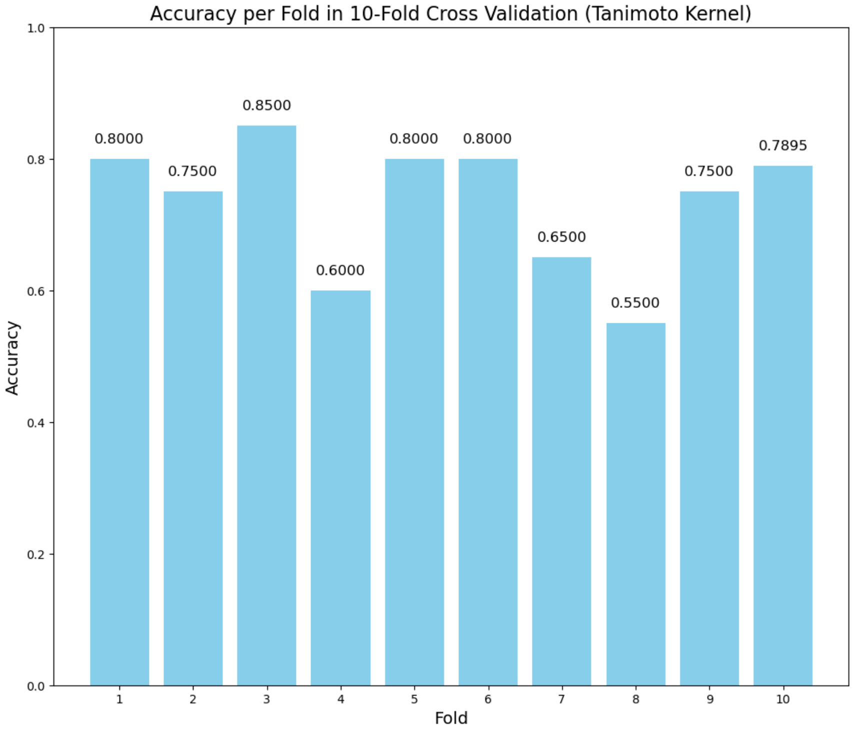


Figure 11: BINARY GENE EXPRESSION DATA WITH TANIMOTO KERNEL

Mean accuracy across 10 folds: 0.7339

HUNTINGTON DISEASE – GSE33000

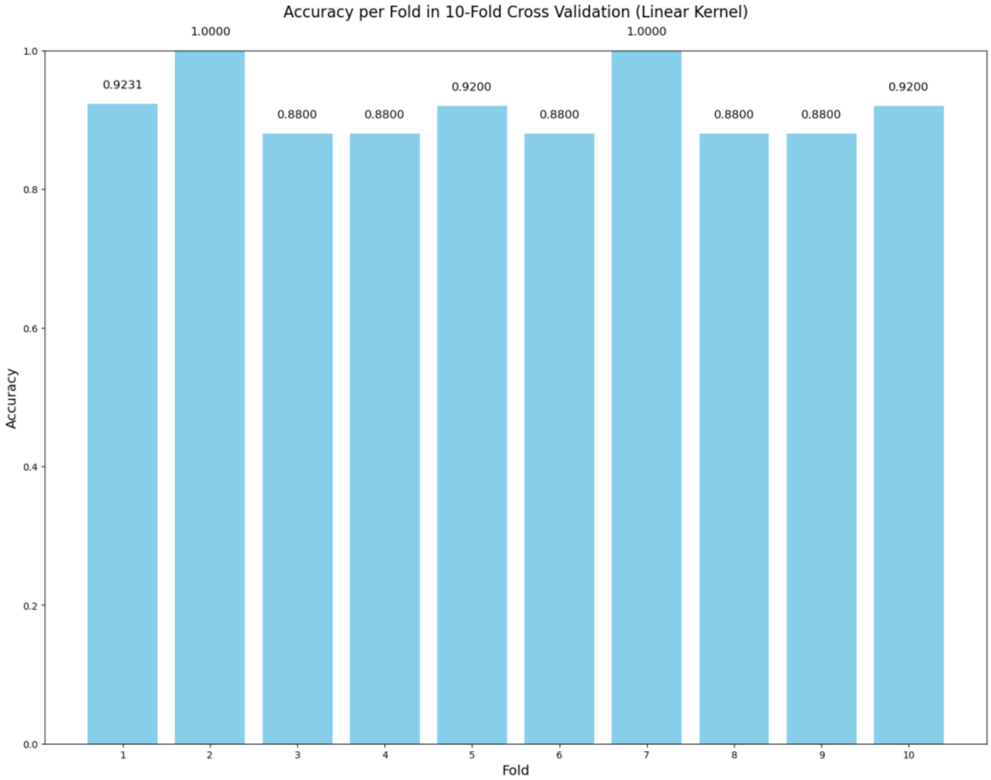


Figure 12: RAW GENE EXPRESSION DATA WITH LINEAR KERNEL

Mean accuracy across 10 folds: 0.9163

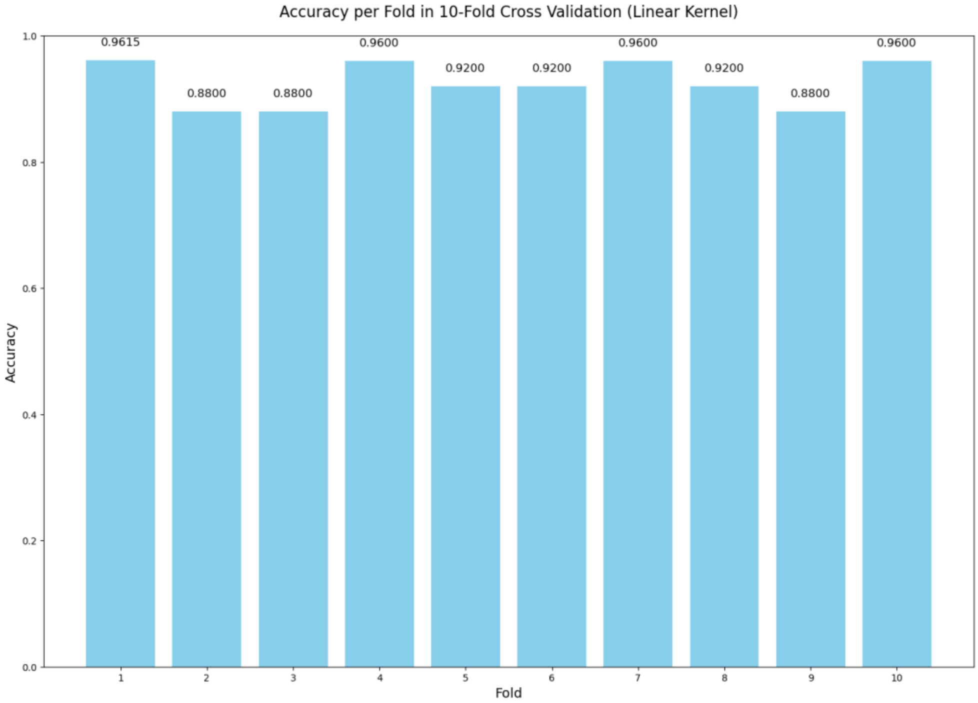


Figure 13: BINARY GENE EXPRESSION DATA WITH LINEAR KERNEL

Mean accuracy across 10 folds: 0.9242

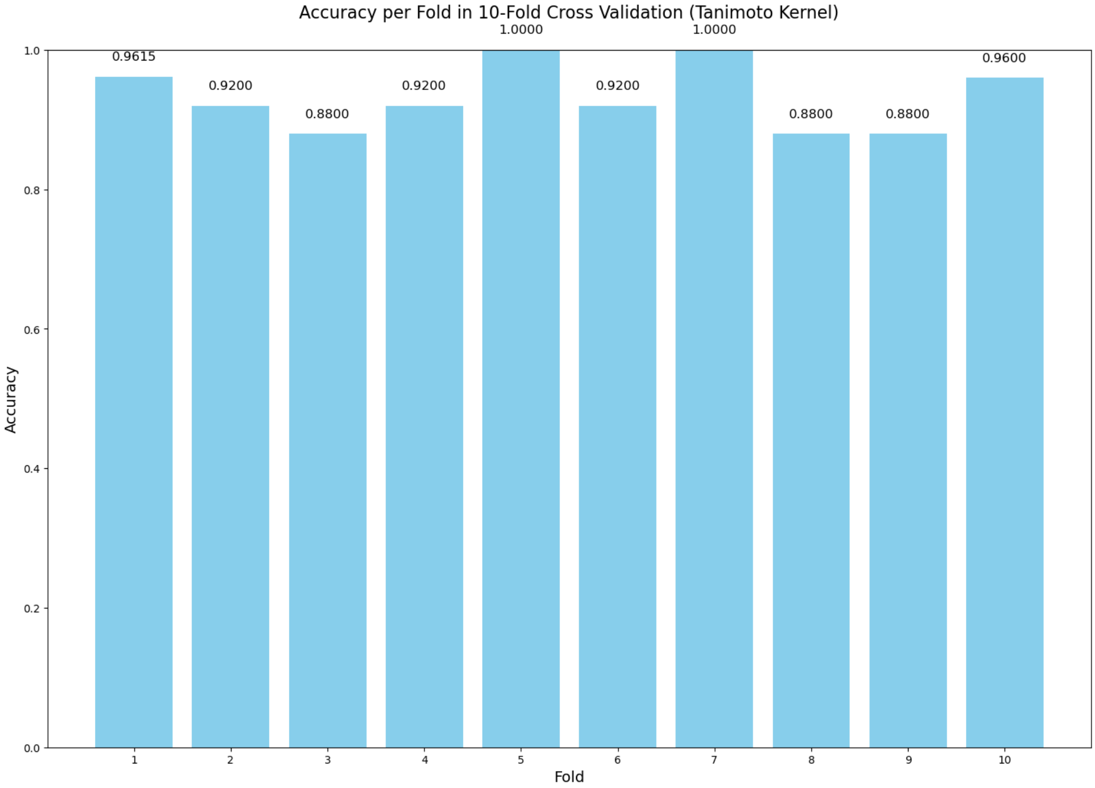


Figure 14: BINARY GENE EXPRESSION DATA WITH TANIMOTO KERNEL

Mean accuracy across 10 folds: 0.9322

PARKINSON DISEASE – GSE57475

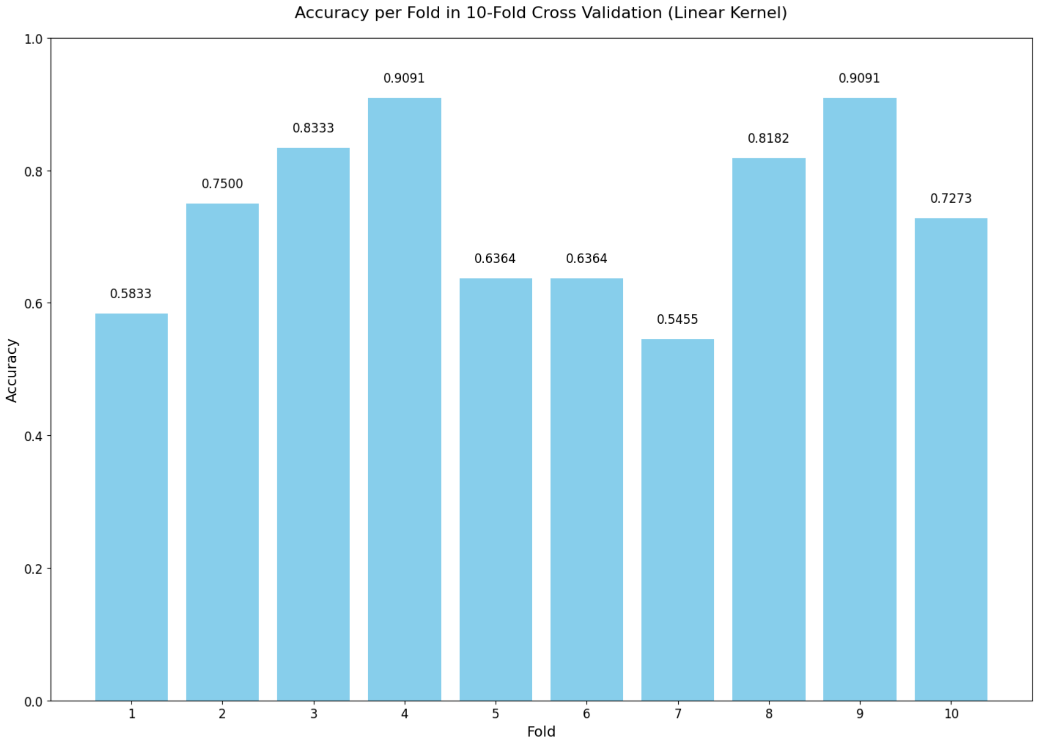


Figure 15: RAW GENE EXPRESSION DATA WITH LINEAR KERNEL

Mean accuracy across 10 folds: 0.7348

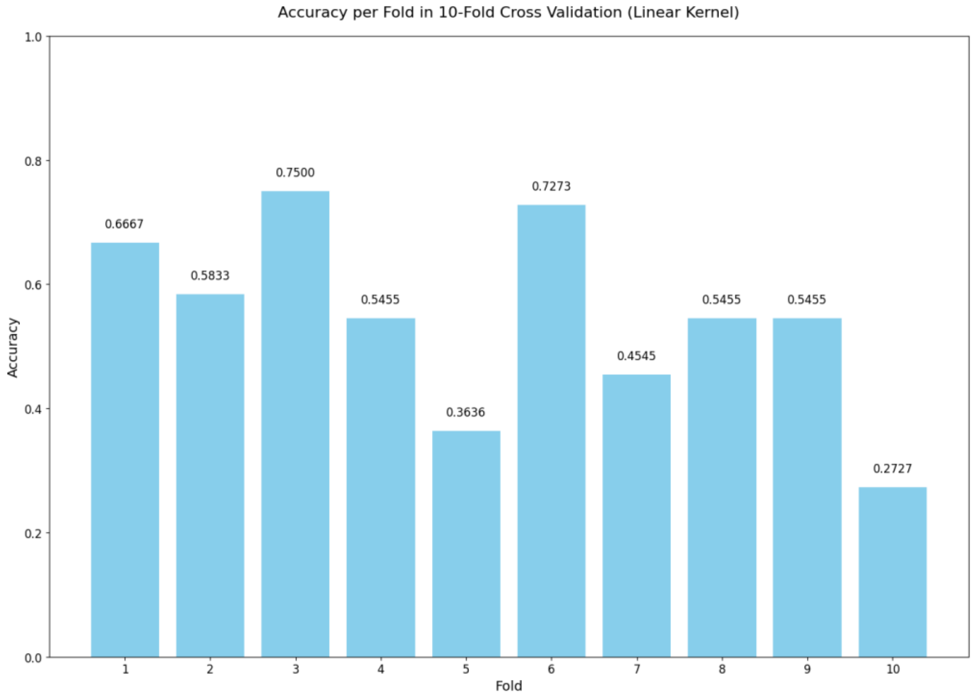


Figure 16: BINARY GENE EXPRESSION DATA WITH LINEAR KERNEL

Mean accuracy across 10 folds: 0.5455

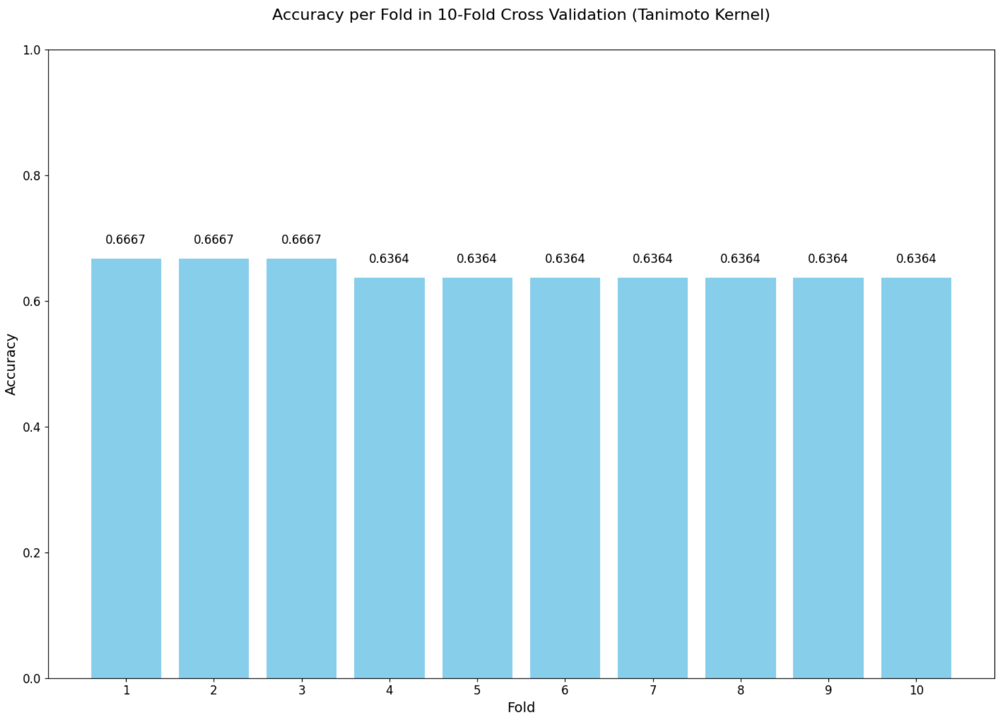


Figure 17: BINARY GENE EXPRESSION DATA WITH TANIMOTO KERNEL

Mean accuracy across 10 folds: 0.6455

AMYOTROPHIC LATERAL SCLEROSIS – GSE112680

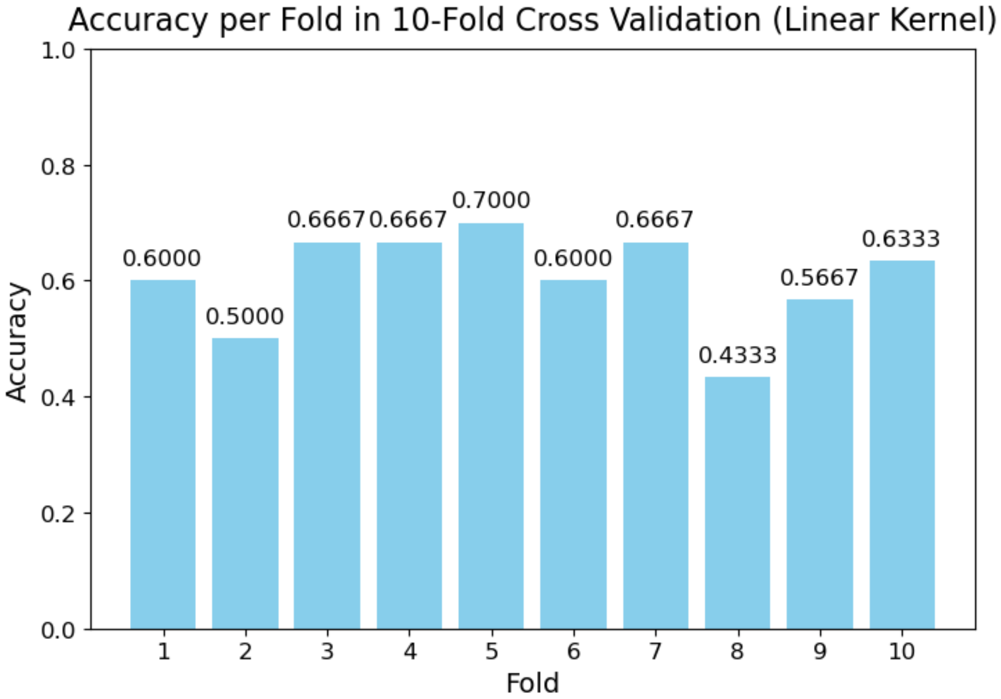


Figure 18: RAW GENE EXPRESSION DATA WITH LINEAR KERNEL

Mean accuracy across 10 folds: 0.6033

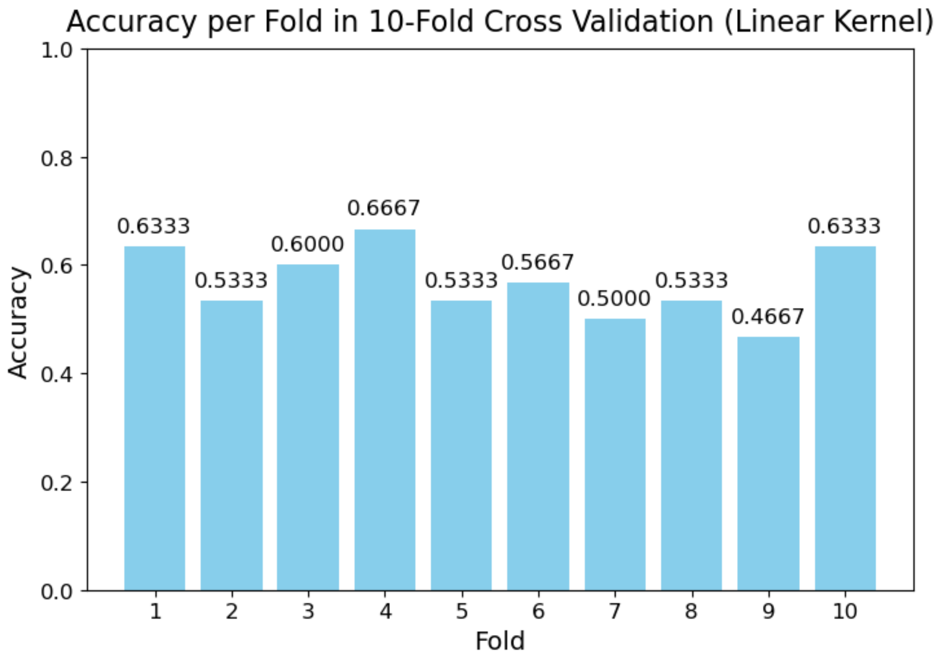


Figure 19: BINARY GENE EXPRESSION DATA WITH LINEAR KERNEL

Mean accuracy across 10 folds: 0.5667

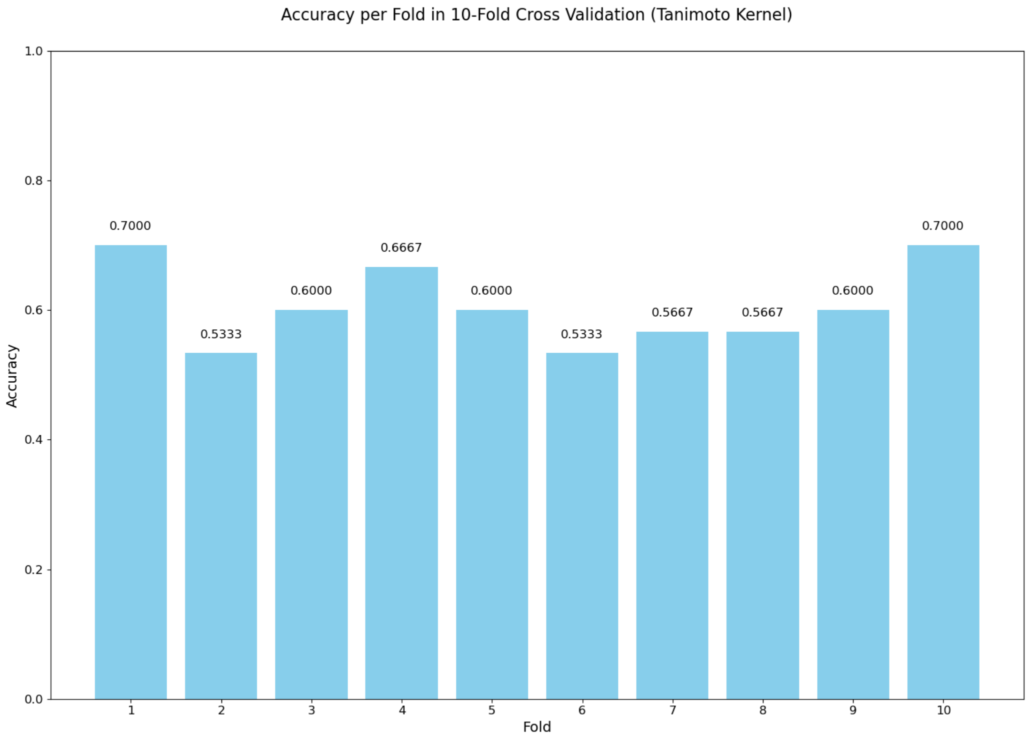


Figure 20: BINARY GENE EXPRESSION DATA WITH TANIMOTO KERNEL

Mean accuracy across 10 folds: 0.6067

5.4.2. MODELS EXECUTION COMPARISON

ALZHEIMER DISEASE – GSE63060

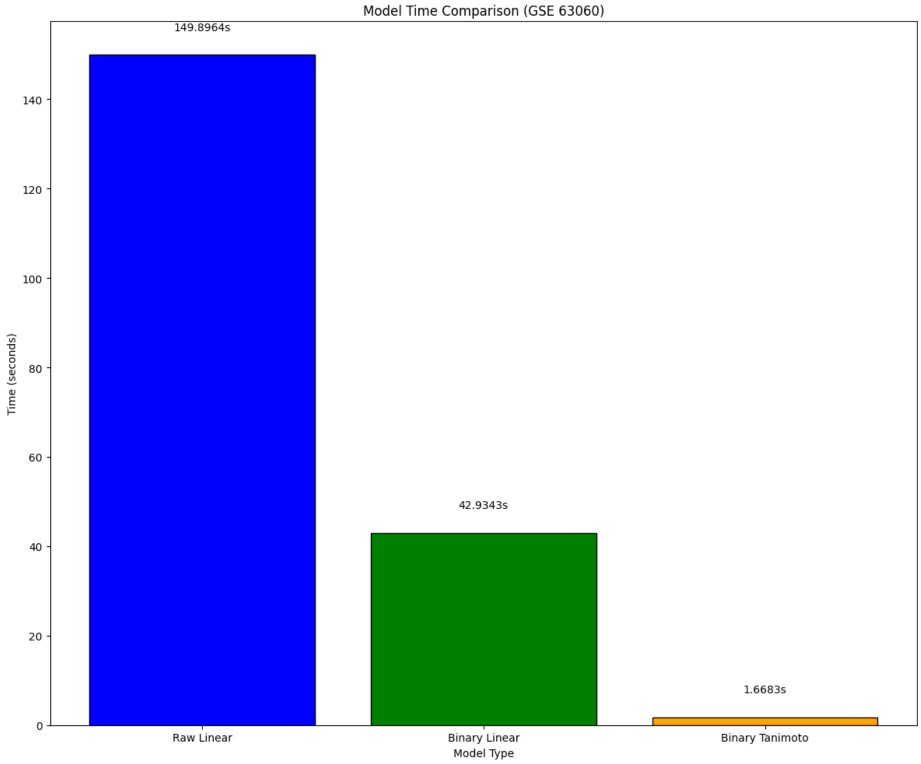


Figure 21: Execution time of GSE63060

HUNTINGTON DISEASE – GSE33000

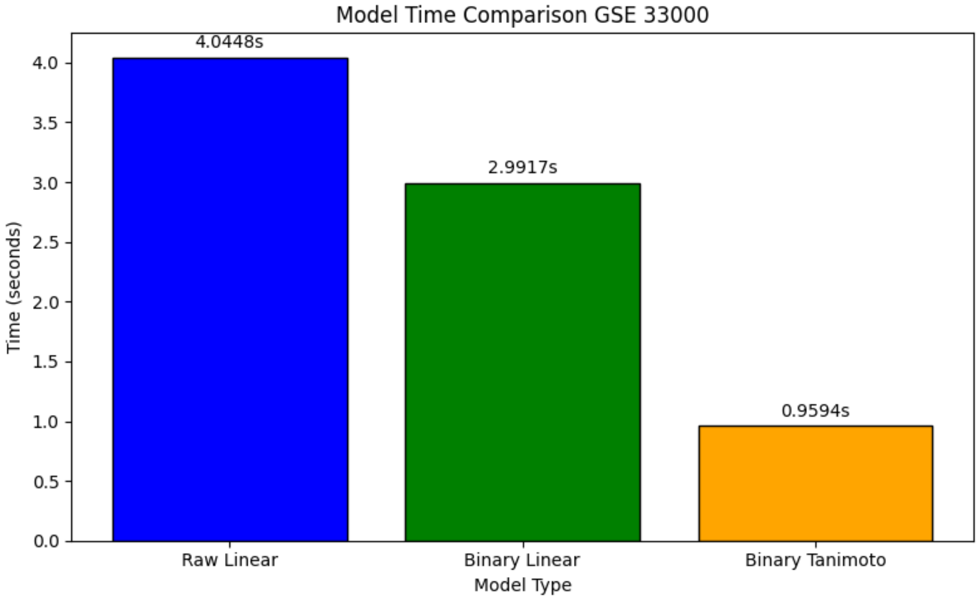


Figure 22: Execution time of GSE33000

PARKINSON DISEASE-GSE57475

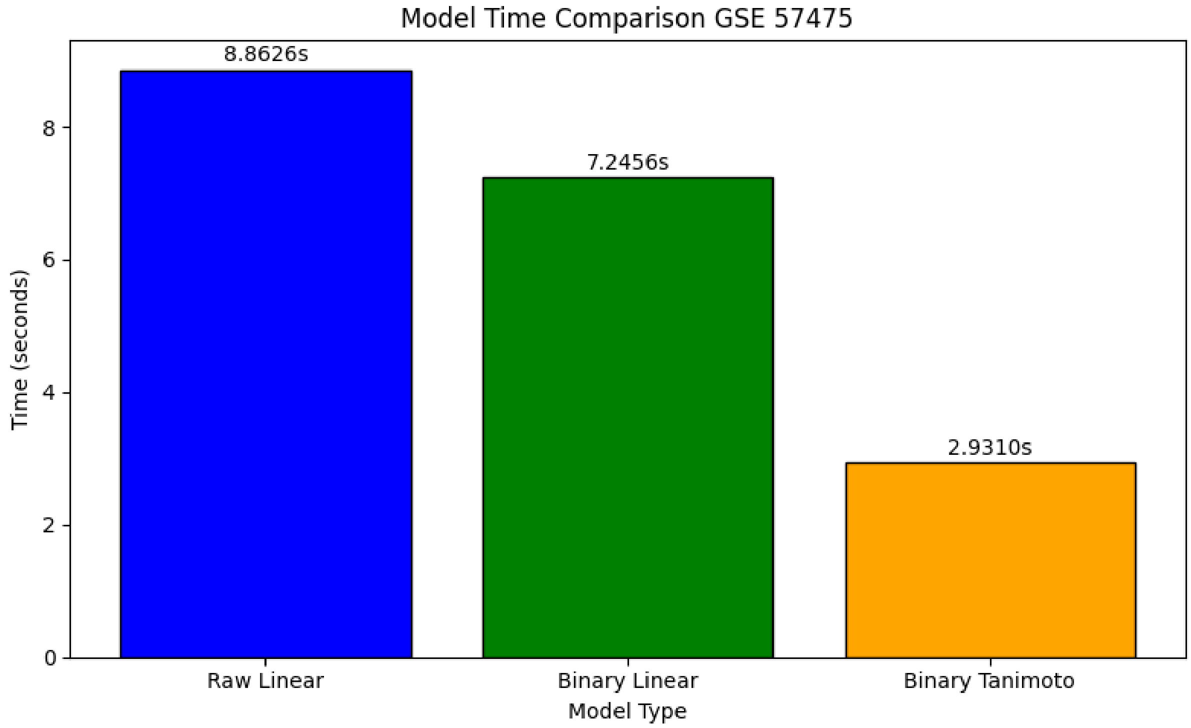


Figure 23: Execution time of GSE57475

AMYOTROPHIC LATERAL SCLEROSIS – GSE112680

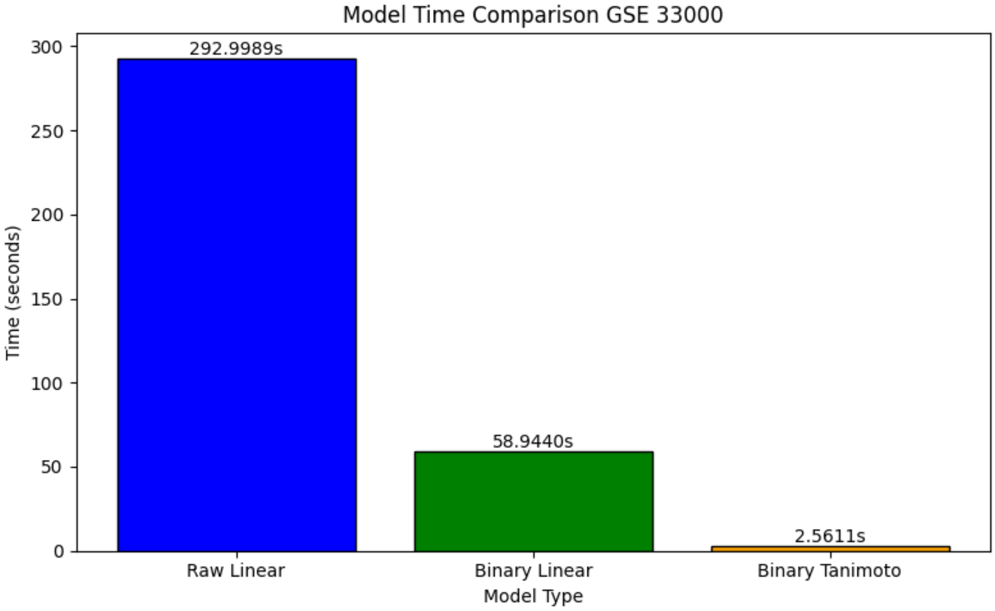


Figure 24: Execution time of GSE112680

5.4.3 MODELS PERFOMANCE USING CONFUSION MATRIX

DATASETS	MODEL	CONFUSION MATRIX
GSE63060	RAW DATA + LINEAR KERNEL	12 11 5 22
	BINARY DATA +LINEAR KERNEL	14 9 6 21
	BINARY DATA + TANIMOTO KERNEL	11 12 3 24
GSE33000	RAW DATA +LINEAR KERNEL	35 3 2 23
	BINARY DATA +LINEAR KERNEL	31 7 4 21
	BINARY DATA + TANIMOTO KERNEL	33 5 1 24
GSE57475	RAW DATA +LINEAR KERNEL	2 7 3 17
	BINARY DATA +LINEAR KERNEL	1 8 2 18
	BINARY DATA + TANIMOTO KERNEL	3 6 4 16
GSE112680	RAW DATA +LINEAR KERNEL	27 12 17 20
	BINARY DATA +LINEAR KERNEL	30 9 24 13
	BINARY DATA + TANIMOTO KERNEL	31 8 27 10

Table 4: GSE63060 Performance Analysis

GSE63060		Precision	Recall	F1 score	support
RAW DATA +LINEAR SVM	0	0.71	0.52	0.60	23
	1	0.67	0.81	0.73	27
BINARY DATA +LINEAR SVM	0	0.70	0.61	0.65	23
	1	0.70	0.78	0.74	27
BINARY DATA + TANIMOTO SVM	0	0.79	0.48	0.59	23
	1	0.67	0.89	0.76	27

Table 5: GSE33000 Performance Analysis

GSE33000		Precision	Recall	F1 score	support
RAW DATA +LINEAR SVM	0	0.95	0.92	0.93	38
	1	0.88	0.92	0.90	25
BINARY DATA +LINEAR SVM	0	0.89	0.82	0.85	38
	1	0.75	0.84	0.79	25
BINARY DATA + TANIMOTO SVM	0	0.97	0.87	0.92	38
	1	0.91	0.96	0.89	25

Table 6: GSE57475 Performance Analysis

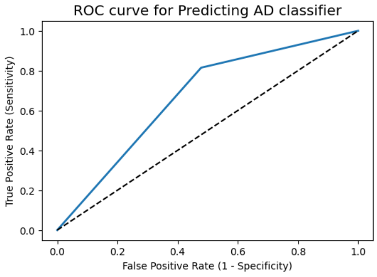
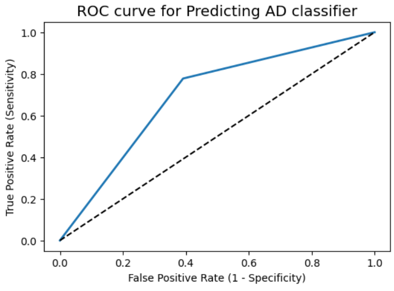
GSE57475		Precision	Recall	F1 score	support
RAW DATA +LINEAR SVM	0	0.40	0.22	0.29	9
	1	0.71	0.85	0.77	20
BINARY DATA +LINEAR SVM	0	0.33	0.11	0.17	9
	1	0.69	0.90	0.78	20
BINARY DATA + TANIMOTO SVM	0	0.00	0.00	0.00	9
	1	0.69	1.00	0.82	20

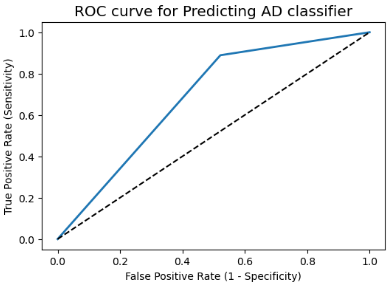
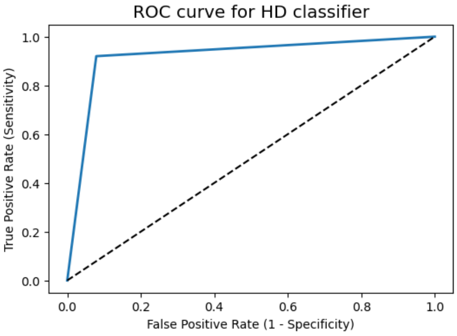
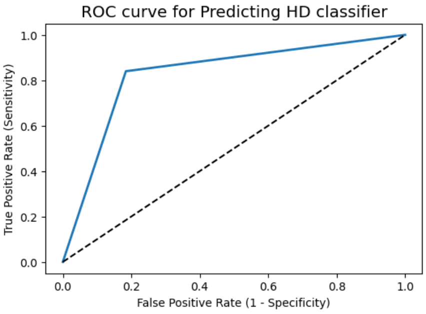
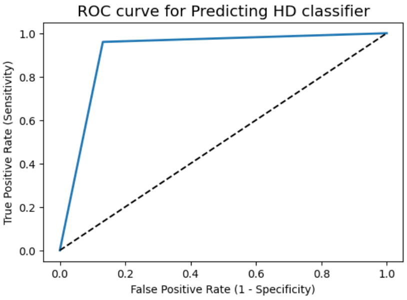
Table 7: GSE112680 Performance Analysis

GSE112680		Precision	Recall	F1 score	support
RAW DATA +LINEAR SVM	0	0.61	0.69	0.65	39
	1	0.62	0.54	0.58	37
BINARY DATA +LINEAR SVM	0	0.56	0.77	0.65	39
	1	0.59	0.35	0.44	37
BINARY DATA + TANIMOTO SVM	0	0.53	0.79	0.64	39
	1	0.56	0.27	0.36	37

5.4.4 ROC CURVE RESULTS

Table 8: ROC-AUC CURVE RESULTS

DATASETS	MODEL	ROC - AUC	CROSS VALIDATED ROC - AUC
GSE63060	RAW DATA +LINEAR SVM  	0.6683	0.7354
	BINARY DATA +LINEAR SVM  	0.6932	0.7222

	BINARY DATA + TANIMOTO SVM  	0.6836	0.7566
GSE33000	RAW DATA +LINEAR SVM  	0.9205	0.9732
	BINARY DATA +LINEAR SVM  	0.8279	0.9795
	BINARY DATA + TANIMOTO SVM  	0.9142	0.9783

Chapter 6: RESULTS ANALYSIS

6. RESULTS ANALYSIS

6.1 INTRODUCTION

The analysis of results highlights the efficiency of using binarized gene expression data over raw gene expression data. The accuracy achieved with binarized data is comparable to or slightly better than that of raw data, demonstrating that binarization does not significantly reduce predictive performance. Additionally, the execution time for binarized models is substantially faster, making them more computationally efficient. This combination of maintained accuracy and reduced execution time suggests that binarized gene expression models provide a **more efficient approach for diagnosing neurodegenerative diseases.**

6.2 RESULTS ACHIEVED

Table 9: Model Accuracy Comparison Results

DATASETS	MODEL	MODEL ACCURACY
GSE63060	RAW DATA + LINEAR SVM	0.7189 ± 0.07
	BINARY DATA +LINEAR SVM	0.6995 ± 0.13
	BINARY DATA + TANIMOTO SVM	0.7339 ± 0.09
GSE33000	RAW DATA +LINEAR SVM	0.9163 ± 0.05
	BINARY DATA +LINEAR SVM	0.9242 ± 0.03
	BINARY DATA + TANIMOTO SVM	0.9322 ± 0.04
GSE57475	RAW DATA +LINEAR SVM	0.6595 ± 0.10
	BINARY DATA +LINEAR SVM	0.6657 ± 0.08
	BINARY DATA + TANIMOTO SVM	0.7124 ± 0.09
GSE112680	RAW DATA +LINEAR SVM	0.6033 ± 0.08
	BINARY DATA +LINEAR SVM	0.5667 ± 0.06
	BINARY DATA + TANIMOTO SVM	0.6067 ± 0.06

Table 10: Model Time Comparison Results

DATASETS	MODEL	MODEL TIME TAKEN (s)
GSE63060	RAW DATA +LINEAR SVM	149.8964
	BINARY DATA +LINEAR SVM	42.9343
	BINARY DATA + TANIMOTO SVM	1.6683
GSE33000	RAW DATA +LINEAR SVM	4.0448
	BINARY DATA +LINEAR SVM	2.9917
	BINARY DATA + TANIMOTO SVM	0.9594
GSE57475	RAW DATA +LINEAR SVM	8.8626
	BINARY DATA +LINEAR SVM	7.2456
	BINARY DATA + TANIMOTO SVM	2.9531
GSE112680	RAW DATA +LINEAR SVM	292.9989
	BINARY DATA +LINEAR SVM	58.9440
	BINARY DATA + TANIMOTO SVM	2.5611

7. CHALLENGES AND SOLUTIONS

Microarray datasets often suffer from high dimensionality, with a vast number of features compared to the sample size, leading to challenges such as overfitting, reduced model performance, and computational inefficiency. To address this, we employed biological feature selection methods by focusing on differentially expressed genes (DEGs) and transcription factor (TF)-related genes [6], ensuring that the selected features are both statistically significant and biologically meaningful, thereby improving interpretability and accuracy. Additionally, traditional thresholding methods for binarization, such as mean or median-based approaches, often result in significant information loss, negatively impacting model performance. To overcome this, we used a 2-Gaussian mixture model [16] to determine optimal thresholds, effectively capturing the natural distribution of the data and preserving critical information. This approach enhances model robustness and accuracy while mitigating the limitations of conventional methods.

8. CONCLUSION

Predictive Performance:

- GSE 33000: The Raw Linear SVM model had the highest overall metrics in terms of precision, recall, and F1 score, with an impressive ROC AUC. However, the Binary Tanimoto SVM also performed well, showing a good balance between precision and recall and achieving a competitive ROC AUC.
- GSE 63060: The performance was generally lower across all models compared to GSE 33000. The Binary Linear SVM model showed a slight improvement in performance metrics over the Raw Linear SVM, indicating that binarization might help in certain contexts to improve the model's ability to generalize from the training data.
- GSE57475: Performance varied across SVM models. The Raw Data + Linear SVM model showed moderate effectiveness. The Binary Data + Linear SVM model improved slightly, suggesting that binarization might enhance the model's generalization. The Binary Data + Tanimoto SVM model performed the best, demonstrating that the specialized Tanimoto kernel effectively utilizes binary data for more accurate predictions.
- GSE112680: The performance of the models on this dataset showed mixed results. The Binary Data + Tanimoto SVM model achieved a slight improvement in performance metrics compared to the Raw Data + Linear SVM model, suggesting that the Tanimoto kernel may capture specific patterns in the binary data more effectively. However, the Binary Data + Linear SVM model showed a decrease in performance.

The analysis of results across all datasets representing different neurodegenerative diseases highlights the effectiveness of using binarized gene expression data with three SVM approaches: Raw Data + Linear SVM, Binary Data + Linear SVM, and Binary Data + Tanimoto SVM. The execution time for binarized models is substantially faster, making them more computationally efficient. The difference in execution time between raw data models and binarized models is significant, with raw data models consistently taking much longer to execute.In terms of accuracy, binarized data models perform comparably to, or slightly better than, raw data models in most cases. For instance, the Binary Data + Tanimoto SVM approach often outperforms the raw data models, demonstrating its ability to capture complex patterns more effectively. While the Binary Data + Linear SVM model shows a slight decrease in accuracy compared to the Raw Data + Linear SVM in some cases, the difference is not substantial and remains within acceptable limits.

Binarized gene expression data does not significantly reduce predictive performance and offers the added advantage of greatly reduced execution time. This combination of maintained accuracy and enhanced computational efficiency suggests that binarized gene expression models, particularly when paired with advanced kernels like Tanimoto SVM, provide a highly efficient and scalable approach for diagnosing neurodegenerative diseases.

9. FUTURE FINDINGS

In the future findings section of our research, will be measure and analyze the memory usage of these models. Understanding the memory consumption is essential for deploying these models in real-world settings, where computational resources can be a constraint. This analysis will help identify the most resource-efficient models, providing insights into their suitability for large-scale or resource-constrained environments. By combining performance metrics with memory usage data, that will gain a comprehensive understanding of the practical viability of each model type.

10. REFERENCES

[1] Zhao, S. *et al.* (2022a) ‘Machine learning-based characterization of cuprotosis-related biomarkers and immune infiltration in parkinson’s disease’, *Frontiers in Genetics*, 13. doi:10.3389/fgene.2022.1010361.

[2] Makarious, M.B. *et al.* (2021) *Multi-modality machine learning predicting parkinson’s disease* [Preprint]. doi:10.1101/2021.03.05.434104.

[3] Falchetti, M., Prediger, R.D. and Zanotto-Filho, A. (2020) ‘Classification algorithms applied to blood-based transcriptome meta-analysis to predict idiopathic parkinson’s disease’, *Computers in Biology and Medicine*, 124, p. 103925. doi:10.1016/j.compbiomed.2020.103925.

[4] Daneshafrooz, N. *et al.* (2022) ‘Identification of potentially functional modules and diagnostic genes related to amyotrophic lateral sclerosis based on the WGCNA and lasso algorithms’, *Scientific Reports*, 12(1). doi:10.1038/s41598-022-24306-2.

[5] Catanese, A. *et al.* (2023) ‘Multiomics and machine-learning identify novel transcriptional and mutational signatures in amyotrophic lateral sclerosis’, *Brain*, 146(9), pp. 3770–3782.

[6] Lee, T. and Lee, H. (2020) ‘Prediction of alzheimer’s disease using blood gene expression data’, *Scientific Reports*, 10(1). doi:10.1038/s41598-020-60595-1.

[7] Wang, L. and Liu, Z.-P. (2019) ‘Detecting diagnostic biomarkers of alzheimer’s disease by integrating gene expression data in six brain regions’, *Frontiers in Genetics*, 10. doi:10.3389/fgene.2019.00157.

[8] Cheng, J. *et al.* (2020) ‘Identification of contributing genes of Huntington’s Disease by Machine Learning’, *BMC Medical Genomics*, 13(1). doi:10.1186/s12920-020-00822-w.

[9] Huseby, C.J. *et al.* (2022) ‘Blood transcript biomarkers selected by machine learning algorithm classify neurodegenerative diseases including alzheimer’s disease’, *Biomolecules*, 12(11), p. 1592. doi:10.3390/biom12111592.

[10] Myszczynska, M.A. *et al.* (2020) ‘Applications of machine learning to diagnosis and treatment of Neurodegenerative Diseases’, *Nature Reviews Neurology*, 16(8), pp. 440–456. doi:10.1038/s41582-020-0377-8.

[11] Tuna, S. and Niranjan, M. (2009) ‘Classification with binary gene expressions’, *Journal of Biomedical Science and Engineering*, 02(06), pp. 390–399. doi:10.4236/jbise.2009.26056.

[12] Su, C., Tong, J. and Wang, F. (2020) ‘Mining genetic and transcriptomic data using machine learning approaches in parkinson’s disease’, *npj Parkinson’s Disease*, 6(1). doi:10.1038/s41531-020-00127-w.

[13] H. Rafieipour, A. A. Zadeh, A. Moradan, and Z. Salekshahrezaee, "Study of Genes Associated with Parkinson Disease Using Feature Selection," *Journal of Biomedical Research*, Oct. 2020. doi: 10.22034/jbr.2020.251812.1035.

[14] Allison, D.B. *et al.* (2006) ‘Microarray data analysis: From disarray to consolidation and consensus’, *Nature Reviews Genetics*, 7(1), pp. 55–65. doi:10.1038/nrg1749.

[15] Lin, R.-H., Wang, C.-C. and Tung, C.-W. (2022) ‘A machine learning classifier for predicting stable MCI patients using gene biomarkers’, International Journal of Environmental Research and Public Health, 19(8), p. 4839. doi:10.3390/ijerph190848.

[16] X. Zhou, X. Wang, and E. R. Dougherty, (2003) Binari zation of microarray data on the basis of a mixture model

[17] Niharika, Roy A, Patra SK. Identification of differentially expressed genes and screening for key genes involved in ovarian cancer prognosis: An integrated bioinformatics and network analysis approach. J Reprod Healthc Med. 2024;5:8. doi: 10.25259/JRHM_6_2024

[18] Ralaivola, Liva, et al. “Graph Kernels for Chemical Informatics.” *Neural Networks*, vol. 18, no. 8, Oct. 2005, pp. 1093–1110, https://doi.org/10.1016/j.neunet.2005.07.009. Accessed 15 Mar. 2020.