PERFORMANCE COMPARISON OF DIFFERENT MOLECULAR DATA IN THE IDENTIFICATION OF DIABETIC RETINOPATHY

UNDERGRADUATE RESEARCH PROPOSAL SUBMITTED
IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF
BACHELOR OF THE SCIENCE IN ENGINEERING

Submitted by:
Ashfa A.G.F. [2019/E/011]
Chandrasiri H.V.B.L. [2019/E/023]

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF ENGINEERING
UNIVERSITY OF JAFFNA
[AUGUST] 2024

PERFORMANCE COMPARISON OF DIFFERENT MOLECULAR DATA IN THE IDENTIFICATION OF DIABETIC RETINOPATHY

Supervisor(s):

Supervisor : Dr. (Mrs.) P. Jeyananthan

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

KEYWORDS

Diabetic retinopathy
Performance comparison
Machine Learning
DNA methylation
Gene expression
miRNA data
Feature selection
Machine learning

ABSTRACT

Diabetic retinopathy is a serious issue caused by diabetic that can damage the eye retina. This can lead to many complications up to the level of blindness if untreated. Even though this can be diagnosed using dilated eye exams, current technologies left the world with many different digital data such as medical images and omics data. Medical images are widely used in the diagnosis of this disease using machine learning algorithms. However analyzing the omics data has much more advantageous compared to image processing such as less complex, low time and computer power, especially biomarker identification can be done during the study of omics data. In the same way, this study uses three different omics data such as DNA methylation, total RNA and small RNA in the diagnosis of diabetic retinopathy using different machine learning algorithms. Four different feature selection algorithms are used with each data individually to select the biomarkers of the study and the best set of features are used with different machine learning algorithms to reveal the model with the highest accuracy. Comparing the accuracies between models shows that best 14 total RNA features selected using Random Forest Feature Importance along with Naïve Bayes algorithms outperforms other model with the accuracy value of 0.9625. Further to this selected set of features are biologically validated using gene ontology (GO) analysis.

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.

Ashfa A.G.F. (2019/E/011) :……

Chandrasiri H.V.B.L. (2019/E/023) :…….....

CONTRIBUTION TO THE PROPOSAL BY THE MEMBERS IN GROUP

Sections	2019/E/011	2019/E/023
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 Forecasting Models & Prediction Models	✓	
2.3 Performance Analysis	✓	
2.4 Available Databases	✓	
2.5 Research Gap		✓
Chapter 3: Methodology And Research Plan		
3.1 Methodology in Brief	✓	✓
3.2 Detailed Methodology		✓
3.3 Timeline	✓	
CHAPTER 4: EXPERIMENTAL FRAMEWORK		
4.1 Introduction		✓
4.2 Database Collection	✓	
4.3 Database Preparation		✓
4.4 Feature selection	✓	✓
4.5 Prediction method	✓	✓
4.6 Evaluation method	✓	✓
CHAPTER 5: EXPERIMENTAL RESULT		
5.1 Introduction	✓	
5.2 Experimental Result	✓	✓
CHAPTER 6: RESULT ANALYSIS		
6.1 Introduction	✓	
6.2 Result and Discussion	✓	✓
CHAPTER 7: CONCLUSION FUTURE DIRECTIONS		
7.1 Conclusion		✓
7.2 Future Work	✓	✓
REFERENCE	✓	✓

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AUC	:	Area Under the Curve
AI	:	Artificial Intelligence
APTOS	:	Asia Pacific Tele-Ophthalmology Society
CNN	:	Convolutional Neural Network
DCNN	:	Deep Convolutional Neural Networks
DME	:	Diabetic Macular Edema
DL	:	Deep Learning
DNA	:	Deoxyribonucleic Acid
DR	:	Diabetic Retinopathy
DM	:	Diabetes Mellitus
DME	:	Diabetic Macular Edema
FN	:	False Negative
FP	:	False Positive
GEO	:	Gene Expression Omnibus
GRU	:	Gated Recurrent Unit
GWAS	:	Genome-Wide Association Study
HER	:	Electronic Health Record
KNN	:	K-Nearest Neighbors Algorithm
LASSO	:	Least Absolute Shrinkage And Selection Operator
LSTM	:	Long Short-Term Memory
ML	:	Machine Learning
NADPH	:	Nicotinamide Adenine Dinucleotide Phosphate
NCBI	:	National Center for Biotechnology Information
NN	:	Neural Networks
NOX4	:	NADPH Oxidase 4
NPDR	:	Non-Proliferative Diabetic Retinopathy
PCA	:	Principal Component Analysis
PDR	:	Proliferative Diabetic Retinopathy
ROC	:	Receiver Operating Characteristic Curve
RNA	:	Ribonucleic acid
RNN	:	Recurrent Neural Network
SVM	:	Support Vector Machine
SNP	:	Single Nucleotide Polymorphisms
T2DM	:	Type 2 Diabetes Mellitus
TN	:	True Negative
TP	:	True Positive
UPLC-MS Spectrometry	:	Ultrahigh-Performance Liquid Chromatography Mass
VGG	:	Visual Geometry Group

CHAPTER 1: Introduction

1.1 Motivation and Overview

Diabetes mellitus (DM) is becoming more common in emerging and wealthy nations. It is estimated that by 2045, there will be 629 million people worldwide with diabetes [1]. Diabetes mellitus (DM) causes a medical disorder called diabetic retinopathy (DR). DR is a serious condition that can lead to severe blindness by damaging the human retina [2]. DM is a chronic condition due to problems with glucose metabolism and various issues with blood vessels [2]. Early detection and accurate diagnosis of DR are crucial for effective treatment. Thanks to advancements in molecular data analysis tools, the detection and categorization of disorders like DR have improved in recent years [2][3]. However, more research is needed to determine which molecular data methods are the most reliable and accurate for diagnosing DR [4]

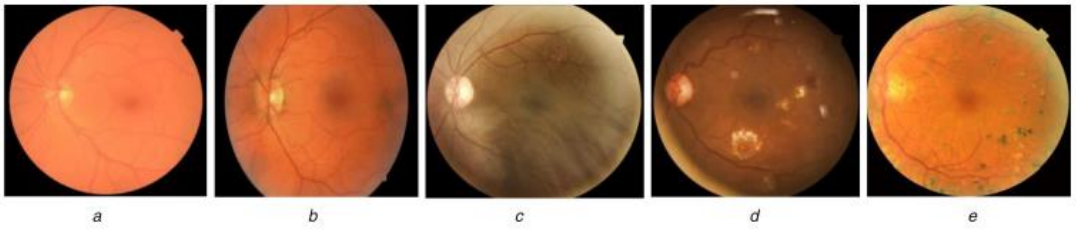


Figure 1 : Images arranged in increasing severity levels of DR [11]
(a) No DR, (b) Mild NPDR, (c) Moderate NPDR, (d) Severe NPDR, (e) PDR

Addressing the significant information gap is the main focus of this research. We aim to make substantial medical diagnostics and customized therapy advancements by comparing and evaluating diverse molecular data types in identifying diabetic retinopathy (DR) [7]. Understanding the pros and cons of various data modalities can aid in the development of more accurate and effective DR detection systems [9]. This, in turn, can lead to improved patient outcomes and advancements in medical procedures. By exploring multiple data sources and their implications for DR identification, we hope to contribute to better medical practices and enhanced patient care [7].

To identify diabetic retinopathy, the suggested research compares the effectiveness of numerous molecular data sets in great detail. The application of several molecular data modalities in the context of DR diagnosis has been studied in several important research publications, which we shall examine and synthesize to achieve this goal [1][2][3][4].

1.2 Aims and Objectives

Our research project aims to use machine learning methods to compare different molecular data sets to find diabetic retinopathy (DR). Early diagnosis and detection of DR are essential for getting the proper treatment on time and avoiding vision loss. Molecular data, like gene expression patterns, protein biomarkers, and epigenetic markers, have much potential for improving the accuracy of diagnosing DR. By looking at these data sets, we hope to improve DR diagnosis and help patients do better.

Our research's main objectives are:

- One of the objectives is to examine how helpful gene expression patterns, protein biomarkers, and epigenetic markers are for detecting and differentiating diabetic retinopathy.
- To assess and compare the performance of various machine learning methods, such as support vector machines, random forests, and deep neural networks, in identifying diabetic retinopathy using different molecular data.
- Another objective is data analysis and interpretation to gain insights into the molecular mechanisms and pathways associated with diabetic retinopathy.

1.3 Research Scope

Our research will focus on gathering and preparing different molecular datasets related to diabetic retinopathy, such as gene expression profiles and protein markers. We will use machine learning methods to test how well they can detect diabetic retinopathy using the collected molecular data. Ethical concerns will also be considered, and the study will be conducted within a specific timeframe, considering data availability and sample size limitations. This research aims to identify potential biomarkers for detecting diabetic retinopathy and proposes using machine learning-based diagnostic tools in real-world medical settings. By doing this study, we hope to shed light on better ways to diagnose and manage diabetic retinopathy for improved patient care.

CHAPTER 2 : Literature Review

2.1 Introduction

In recent years, there has been much interest in using retinal imaging to find and diagnose diabetic retinopathy due to its non-invasive nature and ability to provide detailed visual information about the retina [9][13]. Several studies have explored the potential of artificial intelligence (AI) and machine learning algorithms to predict diabetic retinopathy from retinal images with high accuracy and sensitivity [6]. However, accurately distinguishing between different disease stages, like non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR), remains a challenge [11][14]. Understanding the chemicals in the blood of people with diabetic retinopathy is essential, as they may play a role in the disease's progression [1][21]. By identifying these substances and their contributions to DR growth, we can develop focused, therapeutic interventions and personalized treatment plans [19]. Utilizing plasma metabolites, amino acids, and other molecular markers can provide valuable insights into the disease and its mechanisms [1][10]. This research can significantly improve our understanding and management of diabetic retinopathy.

- Plasma Metabolites

Much has been learned about the relationship between diabetic retinopathy and plasma metabolites, including amino acids and other chemicals in the blood. These studies aim to determine what is different about the substances in the blood of people with diabetic retinopathy and how these chemicals might be linked to the progression of the disease [1]. Targeted methods have been used in metabolomics studies to measure the number of serum metabolites in people with type 2 diabetes and find significant differences between the metabolomics profiles of different analysis groups [12]. These results reveal potential metabolite markers for DR progression in people with type 2 diabetes. [12].

- Multi-Omics Data

Genomic, transcriptomic, proteomic, and metabolomic data are increasingly used in diabetic retinopathy studies [2][5][6]. By looking at these different molecular datasets, researchers can learn a lot about the biological processes involved in the growth and spread of DR [5]. For example, DNA methylation and gene expression data were used to find diagnostic biomarkers for cervical cancer. This shows the promise of multi-omics data to improve disease diagnosis and risk assessment [2]. Integrative study of chromosome copy number variation and gene expression has also been used to examine the molecular changes linked to cervical carcinoma [8].

In support of the current trend in diabetic retinopathy research, our study aims to add to the growing body of literature by using omics data, such as genomes, transcriptomics, proteomics, and metabolomics, to predict diabetic retinopathy. By using machine learning methods on this multidimensional data, we hope to improve the accuracy and specificity of diabetic retinopathy prediction, especially when telling the difference between NPDR and PDR. Our study aims to add to existing methods based on retina imaging and, in the long run, move the field toward a more accurate and effective way to diagnose and treat diabetic retinopathy.

2.2 Forecasting Models

Ultrahigh-performance liquid chromatography-mass spectrometry (UPLC-MS) was used in the study by [1] to examine how plasma molecules changed in people with diabetic retinopathy. The study used machine learning techniques like the Least Absolute Shrinkage and Selection Operator (LASSO) and logistic regression to find important metabolites linked to DR. These metabolites could be treatment targets. If these compounds are correctly identified, it might be possible to develop more effective ways to treat diabetic retinopathy.

In the same way, [2] used support vector machines (SVMs), decision trees, and random forests to determine whether gene expression, protein expression, lipid profile, and microRNA data could serve as biomarkers for diabetic retinopathy. The results showed how important these types of molecular data are for telling the difference between people with and without DR. Using different data in machine learning models makes it possible to get a complete picture of how complicated the disease is and helps make personalized treatment plans.

Additionally, researchers have explored various ways of classifying DR. For instance, [4] proposed an optimized hybrid ML classifier that combined neural networks (NN) and deep convolutional neural networks (DCNN) to classify the severity of DR using smartphone-based retinal imaging accurately. This method demonstrated the potential of using portable devices for testing and monitoring diabetic retinopathy, especially in resource-limited areas.

Also, [17] looked into the role of some genes in people with type 2 diabetes who have severe diabetic retinopathy. The work used genotyping and imputation to figure out how epigenetic mechanisms are involved in glucose-induced transcription during DR. Understanding how genetics play a role in DR can give us essential information about how the disease develops and lead to new treatment methods.

The studies we looked at show how important molecular data and genetic factors are in diagnosing and predicting the outcome of diabetic retinopathy. When these different kinds of data are combined with machine learning methods, accurate classification and risk prediction of diabetic retinopathy be possible. By learning more about how genetic predisposition, molecular factors, and machine learning work together, we can find better, more personalized ways to prevent and treat this debilitating disease.

As we continue to compare and analyze these research papers, we hope to learn essential things from how each study was done and what it found. When DNA data, genetic factors, and cutting-edge machine-learning models are used, they could change how diabetic retinopathy is diagnosed and treated. By discovering the best signs in the body and helpful ways to predict diseases, we can provide doctors with the correct information to treat DR (a specific condition) as soon as possible. This will lead to better results for patients in the end.

2.3 Performance Analysis

The researchers aimed to compare how well different genetic data could be used to spot diabetic retinopathy using machine learning. In this study [1], researchers utilized statistical methods like LASSO and logistic regression to analyze the molecules in the plasma of individuals with diabetic retinopathy. ROC curves were created to evaluate the power of risk score and found it as 0.80 [1]. The results were analyzed using ultrahigh-performance liquid chromatography-mass spectrometry and principal component analysis [1]. Support vector machines, decision trees, and deep learning algorithms were used in another way to diagnose diabetic retinopathy [2], and accuracy, sensitivity, and specificity were used as performance measures [2]. Also, the researchers developed a mixed machine learning classifier that used neural networks and deep convolutional neural networks to classify the severity of DR using images of the retina taken with a smartphone [4][18]. In the study [11], they do the DR image classification report using the table. The structure of that table is given in Table 1.

Description	Normal image in Classification	Image Affected by DR in Classification
Normal Image in Actual	TP	FN
Image affected by DR in actual.	FP	TN

Table 1 : Description Of Tp, Fp, Tn, And Fn For Classification Of Retinal Images

2.4 Research Gap

The paper [9] presents an automated approach for diabetic retinopathy (DR) detection using a radial basis function. While this research contributes to the field of DR identification, it also reveals certain limitations that create opportunities for further investigation.

Currently, we mostly rely on retinal images to diagnose diabetic retinopathy. We have yet to explore how other molecular data can help identify this condition. Some automated methods have shown promise, but we must investigate if including different molecular information like genetic markers and proteomic profiles can also help detect and predict diabetic retinopathy.

The previous research used a small sample size and only one type of neural network, so we should now explore more extensive and diverse datasets and try different advanced machine learning algorithms. This way, we can make the detection of diabetic retinopathy more accurate and reliable. Addressing these gaps can improve how we diagnose and manage diabetic retinopathy, ultimately benefiting patients.

2.5 Available Databases

Peripheral venous blood samples [1]

- including 42 DR patients and 32 T2DM patients without DR

  A blood sample dataset was drawn from each patient after ten hours of overnight fasting. [15]

- In the paper, they showed all datasets in Table 1

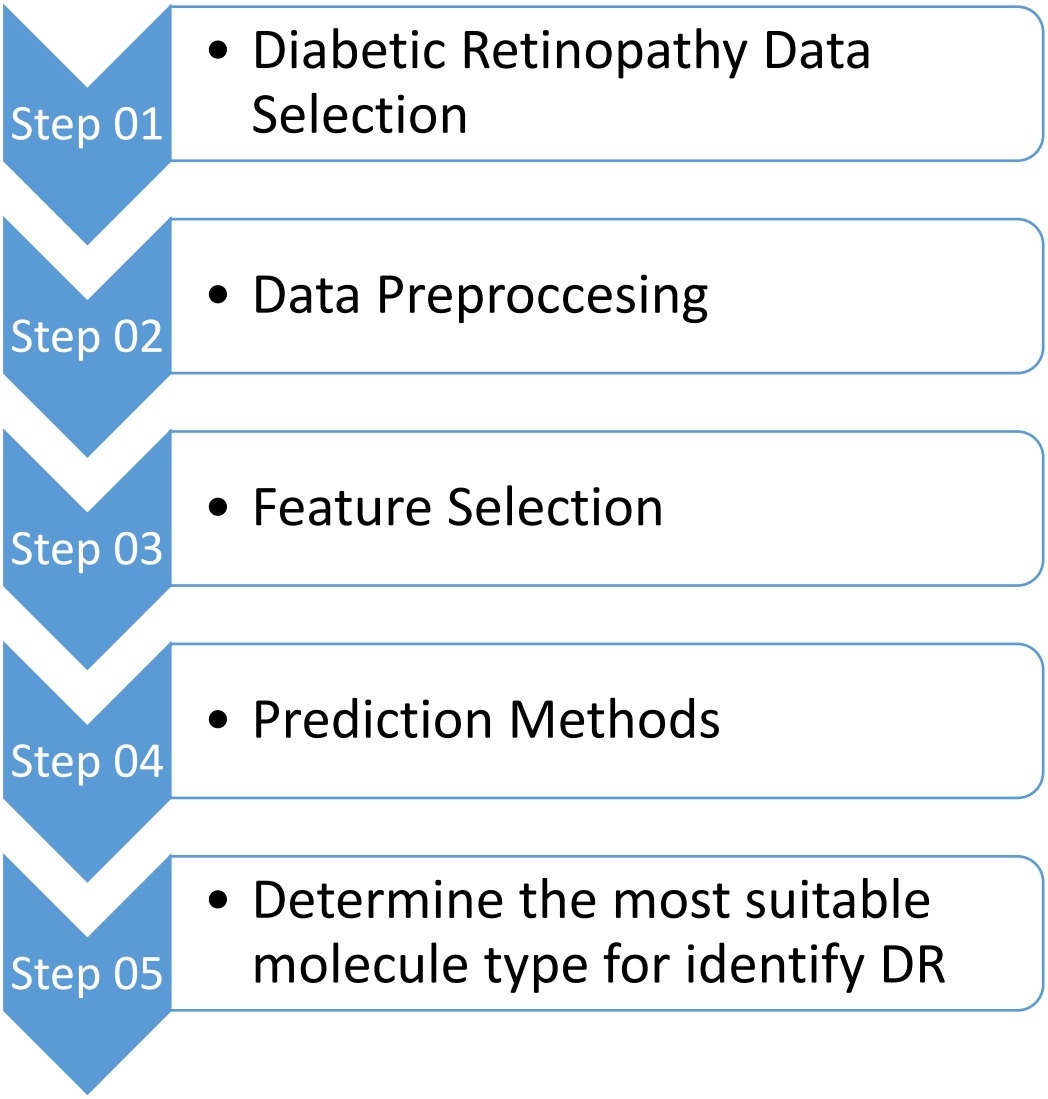
genome-wide association study (GWAS) dataset [16]

- Caucasian Australians with type 2 diabetes were evaluated in a genome-wide association study (GWAS) to compare 270 DME cases and 176 PDR cases with 435 non-retinopathy controls.
- All participants were genotyped by SNP array, and after data cleaning, cases were compared to controls using logistic regression adjusting for relevant covariates.

CHAPTER 3 : Methodology And Research Plan

3.1 Methodology in Brief

This study's methodology involves identifying the best molecule type for diagnosing diabetic retinopathy (DR). The process begins by collecting various kinds of molecules for evaluation. Next, the data undergoes preprocessing to ensure its quality and consistency. Feature selection techniques are then applied to identify the most relevant molecular attributes. Subsequently, multiple models are trained using the preprocessed data. The models' performances are compared to identify the best ones. Finally, the molecule type that yields the highest accuracy in identifying DR using the selected models is determined, providing valuable insights for effective DR diagnosis.



3.2 Detailed Methodology

3.2.1 Data Selection of Diabetic Retinopathy

There are many databases worldwide for selecting datasets for analyzing molecule data. This section discusses the systematic strategy we will use for data selection from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) to assess the performance of different molecular data in detecting Diabetic Retinopathy.

The National Center for Biotechnology Information (NCBI) houses the Gene Expression Omnibus (GEO), an extensive and diverse repository of publicly accessible gene expression data. Researchers from around the globe contribute to GEO, making it a valuable storehouse of genomic information about a wide range of biological processes, diseases, and experimental conditions.

We wish to obtain various molecular datasets. As a result, in addition to gene expression data, GEO provides a wide range of omics data, such as microRNA expression, DNA methylation, chromatin accessibility, and more. This integration enables researchers to conduct multi-omics studies, creating a complete knowledge of biological processes and disease causes.

These datasets are provided on GEO and NCBI

- mRNA Gene Expression Datasets
- Datasets of MicroRNA Expression
- Datasets on Epigenetics
- Datasets of ChIP-Seq
- Datasets for Clinical Research
- Datasets for Specific Diseases

We decided to get these types of datasets related to DR

- DNA methylation
- RNA-Seq
  - total RNA
  - small RNA

We took two datasets from GEO and considered the above molecular data

- GEO accession Number: GSE140842

  Title: Alterations of 5-Hydroxymethylcytosines in Circulating Cell-free DNA Reflect Retinopathy in Type 2 Diabetes
  About dataset: This dataset contains genome-wide methylation profiles of circulating cell-free DNA (cfDNA) from 70 Chinese patients with type 2 diabetes mellitus (T2DM), including 35 patients with diabetic retinopathy (DR) and 35 age-, gender-, and diabetic duration-matched controls.

- GEO accession Number: GSE160310

  Title: In-depth transcriptomic analyses Investigating molecular mechanisms underlying diabetic retinopathy
  About the dataset: This is a collection of transcriptomic data from human post-mortem retinal samples. The data was collected from 80 patients diagnosed with various stages of diabetic retinopathy (DR). The data was analyzed using RNA-Seq, a high-throughput sequencing technique that can measure gene expression in a sample.

- total RNA
- small RNA

3.2.2 Data preprocessing

Before using molecular data in machine learning analyses, it is crucial to perform data preprocessing. This step helps to get the data ready and organized for accurate and effective ML analysis. The importance of data preprocessing is highlighted in the sources [1][2][3]. When we do normalization, those methods ensure data consistency [3]. Also, methods like PCA reduce the number of dimensions and help to find the essential parts of the plasma metabolome [1]. All Gene expression, protein expression, lipid profile, and microRNA data are gathered and preprocessed [2] before use. Instead of deleting or removing some data from the dataset, applying these preprocessing methods can give more information for our model. When diagnosing diabetic retinopathy, valuable and clean data can lead to accurate and meaningful results if the data preprocessing is done well. In our selected datasets, we found some null values and replaced them with proper values. Some columns are normalized using Python script.

3.2.3 Feature Selection

In machine learning, feature selection is crucial for choosing the most essential and useful features from the original dataset. It aims to improve model performance, reduce overfitting, and speed up computing. Feature selection helps to simplify the model by figuring out which parts are the most important and keeping them. This makes the model easier to understand and less subject to confusion. Our research will use Information Gain, Correlation Coefficient, Chi-Square, and Feature Importance. Forward feature selection and backward removal are also iterative methods that gradually add or take away features based on how they affect how well the model works. But techniques like theirs take time to select features. In our research, we choose 200, 150, 100, and 50 features using different methods and evaluate them to pick the best feature sets from the datasets.

- Information Gain (mutual_info_classif)

Information Gain measures how much information a target variable gains when a specific feature is present in the model. It measures how much the target variable depends on each feature. This helps find essential features that add a lot to making predictions with less uncertainty.

- Correlation Coefficient (Pearson Correlation)

The Correlation Coefficient measures the linear relationship between two factors. It shows how much one feature changes when the other feature changes. It helps to find features that strongly relate to the goal variable.

- Chi-Square (chi2)

Chi-square is a statistical test determining whether categorical traits and the target variable are statistically related. It checks whether a categorical attribute and the target class are linked meaningfully. Chi2 is often used to choose which features to use in category data, especially when classifying.

- Feature Importance

This method ranks features based on how important they are to the success of the machine learning model. It gives each feature a score that shows how much it adds to the model's accuracy or ability to guess. It helps find the most critical factors that significantly affect the goal variable.

- Forward Feature Selection

This starts with an empty set of features and adds the most important one at a time based on factors for judging performance, such as accuracy or error rate. This process continues until a stopping point, like when a certain amount of model performance is achieved.

- Backward Elimination

This is a way to choose which features to use. It starts with all the features in the model and removes the least important one at a time based on how well it works. It aims to get rid of parts of the model that don't have much effect on how well it works, which will make the model more efficient and easier to understand. The process continues until a stopping point, like when the desired model performance is achieved.

3.2.4 Apply machine learning methods

In machine learning, prediction means using trained models to make guesses about new data that they haven't encountered before. These models learn from labelled training data to recognise patterns and then use that knowledge to make predictions or put new data into different groups. The goal is to make accurate predictions on entirely new data, showing that the model can work well in practical, real-life situations.

- Support Vector Machine (SVM)

SVM is an algorithm for classification and regression problems that uses supervised learning. It finds the best hyperplane for separating the different classes in the data to make the difference between the classes as big as possible. SVM works well with high-dimensional data and can deal with data that doesn't separate linearly by using kernel functions to move data into higher-dimensional areas. It is used extensively in bioinformatics, text classification, and picture recognition.

- K-Nearest Neighbors (K-NN)

K-NN is a simple classification and regression method based on supervised learning. It gives each data point in the feature space a class or value based on the majority class or average value of its K nearest neighbours. It is often used in suggestion systems, recognizing patterns, and finding outliers.

- Naive Bayes

Based on Bayes' theorem, Naive Bayes is a statistical way to sort things into groups. It thinks that features are independent of the class label, which makes calculations easier. Even though this is a simple assumption, Naive Bayes often does surprisingly well at jobs like classifying text and filtering spam. Compared to other algorithms, it works well with high-dimensional data and only needs a small amount of training data.

- Random Forests

Random Forests is an ensemble learning method that builds multiple decision trees and uses all their predictions to make a final choice. Each tree is trained on a random subset of the data and a random subset of the features. This prevents overfitting and makes learning from new data easier for the tree. Random Forests are reliable, work well with big data sets, and can handle high-dimensional data. They are used for many things, like classification, regression, and ranking the value of features.

3.2.5 Compare performance

We must compare different models in machine learning to find the best one for a specific task. To do this, we use various metrics to measure how well each model makes predictions. These metrics help us understand the strengths and weaknesses of each model and how effectively they work. By comparing models, researchers and practitioners can determine which is best at predicting and handling new, unseen data. This careful comparison allows us to decide based on data and choose the most accurate and dependable model for a particular problem.

- Accuracy

Accuracy is a key performance metric that counts how many instances out of all instances were correctly classified. It gives a general idea of how good the model is but can be misleading when one class is more important than the other.

- Precision

The model's accuracy is measured by how many true positive predictions it makes out of all its positive predictions. It shows how well the model can avoid false positives, which is very important when they are expensive.

- Recall

Recall, also called sensitivity, is the percentage of true positive predictions from all real positive cases in the dataset. It checks how well the model can find positive cases. This is important when you don't want to miss positive cases.

- Area Under the Curve (AUC)

AUC is a performance gauge often used to measure how well a model can distinguish between positive and negative examples. The true positive rate (recall) is shown on the y-axis of the Receiver Operating Characteristic (ROC) curve, and the fake positive rate is shown on the x-axis. The area under this curve is what AUC measures. Several 1 means a perfect model, while 0.5 means guessing at random.

- F1-score

F1-score is the harmonic mean of precision and recall. It gives a balanced measure when working with datasets that are not evenly distributed. It takes into account both false positives and false negatives. This makes it a good step for balancing precision and recall.

3.3 Timeline

Weeks / Tasks	Semester 06					Semester 07					Semester 08			
	1-3	4-6	7-9	10-12	13-15	1-3	4-6	7-9	10-12	13-15	1-2	3-4	5-6	7-8
Literature review	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Bibliography writing		■	■											
Proposal writing				■	■									
Data collection						■								
Data preparation							■							
Finalize the model								■	■					
Model implementation									■	■	■			
Report writing										■	■	■	■	■
Research paper writing													■	■

CHAPTER 4 : Experimental Framework

4.1 Introduction

To study Diabetic Retinopathy (DR), we're using the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI). GEO has various molecular datasets, and we're specifically interested in DNA methylation and two types of RNA-Seq data (total RNA and small RNA). Our strategy involves searching GEO, selecting relevant datasets related to DR, and checking the data quality. Once chosen, we download and integrate these datasets for analysis, aiming to understand how different molecular information performs in detecting DR. This systematic approach helps us make the most of the available resources at GEO and NCBI for our research.

4.2 Database Collection

4.2.1 Phenotype Data selection

So far, our study efforts have focused on choosing phenotype data. More specifically, we have used the clinical data of patients to find out if they have been diagnosed with diabetic retinopathy and, if so, what state of DR they are in.

4.2.2 Data set selection

Our project's primary goal is to find Diabetic Retinopathy (DR) using three kinds of Omic data. To do this, we got three datasets from The Gene Expression Omnibus (GEO) and ensured they were reliable and consistent for DR analysis. Most of our research work is done by preprocessing and exploring these datasets.

- DNA Methylation - GSE140842 (70 x 18602)
- smallRNA – GSE160308 (79 x 2576)
- totalRNA – GSE160306 (79 x 58051)

4.3 Database Preparation

In our research to identify diabetic retinopathy (DR) through omics data, we collected relevant datasets from the internet. Alongside, we utilized clinical data to determine whether patients had been diagnosed with DR and assess the disease’s severity. To create a comprehensive database, we processed and filtered out unwanted data. With the aid of Python code, we merged these files and incorporated the new information. Since our study involved multiple datasets for omics analysis, each dataset was downloaded independently and saved in separate files for future research and integration. Additionally, we performed normalization techniques to optimize the performance of the data.

4.3.1 Data Integration

In our study of Diabetic Retinopathy (DR), we encountered separate files for various molecular datasets. To consolidate this information, we employed Python code to download and combine each dataset, integrating them with their corresponding phenotype data. The target variables for our implementations include patient-specific details, such as genomic data linked to the DR status and other relevant attributes. For instance, we associated genomic data with information about the severity of DR or specific characteristics of patients with Diabetic Retinopathy. The patient ID serves as the common field, facilitating the seamless merging of these datasets. This approach allows us to create a unified dataset that combines molecular information with pertinent patient attributes, enabling a more comprehensive analysis of Diabetic Retinopathy.

4.3.2 Missing values handling

In our dataset analysis, we meticulously examined each variable, and to our delight, no missing values were detected. Consequently, there was no need to employ any missing value-handling techniques. This absence of missing data ensures the robustness of our findings and contributes to the overall data integrity.

4.3.3 Encoding

Ensuring compatibility with machine learning algorithms, we undertake the task of translating object-type values in the dataset into numeric types. This conversion is imperative for subsequent modelling and analysis processes, allowing for accurate computations and predictions.

4.4 Feature Selection

In our study of Diabetic Retinopathy (DR), we implemented four feature selection methods, each contributing to the identification and extraction of relevant features from the datasets. Using each method we create datasets feature counts from 1 to Count of Patients.

- Information Gain (Mutual Info):

Information Gain, measured through Mutual Information, is crucial for understanding the relevance of features. In this method, we leveraged the scikit-learn library to compute the mutual information score for each feature, identifying those with the highest information gain.

- Correlation Coefficient:

Assessing the correlation between features and the target variable is vital for effective feature selection. We utilized the Pearson Correlation Coefficient through Pandas dataframe.corr() to examine pairwise correlations. Features demonstrating a strong correlation with the target column were retained for further analysis.

- Chi-Square:

The chi-square test is employed to evaluate the independence between categorical variables. We utilized this statistical method to measure the significance of the association between each feature and the target variable. Features showing a significant relationship were selected as part of our feature set.

- Feature Importance:

Feature Importance, determined through permutation importance in the scikit-learn library, provides insights into the contribution of each feature to the model's predictive performance. We used the resulting feature importance scores to select the most impactful features for our analysis.

4.5 Prediction Method

In our Diabetic Retinopathy (DR) analysis, we implemented various neural and non-neural network-based prediction models using the scikit-learn library, tailoring specific parameters to meet the requirements of each model.

- Support Vector Machine (SVM):

SVM models were implemented with four different kernels: linear, poly, rbf, and sigmoid. The accuracy of each model was assessed, and the most suitable model was selected based on performance. The hyperparameter 'k' value, crucial for SVM, was chosen considering the available data instances in the relevant dataset.

- Logistic Regression:

Logistic Regression, a linear model suitable for binary classification tasks, was implemented to predict Diabetic Retinopathy. Relevant parameters were adjusted to optimize the model's performance.

- Artificial Neural Network (ANN):

An Artificial Neural Network, a neural network-based model, was employed for prediction tasks. The architecture and hyperparameters of the ANN were adjusted to enhance its ability to capture complex patterns within the data.

- Naive Bayes:

Naive Bayes, a probabilistic classifier, was implemented for its simplicity and efficiency. The model parameters were adjusted to fit the characteristics of the Diabetic Retinopathy dataset.

- Random Forest:

Random Forest, an ensemble learning method, was utilized with a thousand estimators to enhance predictive accuracy. The model was fine-tuned to capture the nuances of the DR dataset.

4.6 Evaluation Method

We checked how well our model works using accuracy and the ROC curve. By doing cross-validation on each dataset, we figured out the average values and how much they vary from the average. This helps us understand how reliable our model is.

4.7 Chapter Summary

In this section, we briefly introduce the experimental framework using a flowchart to illustrate the key steps.

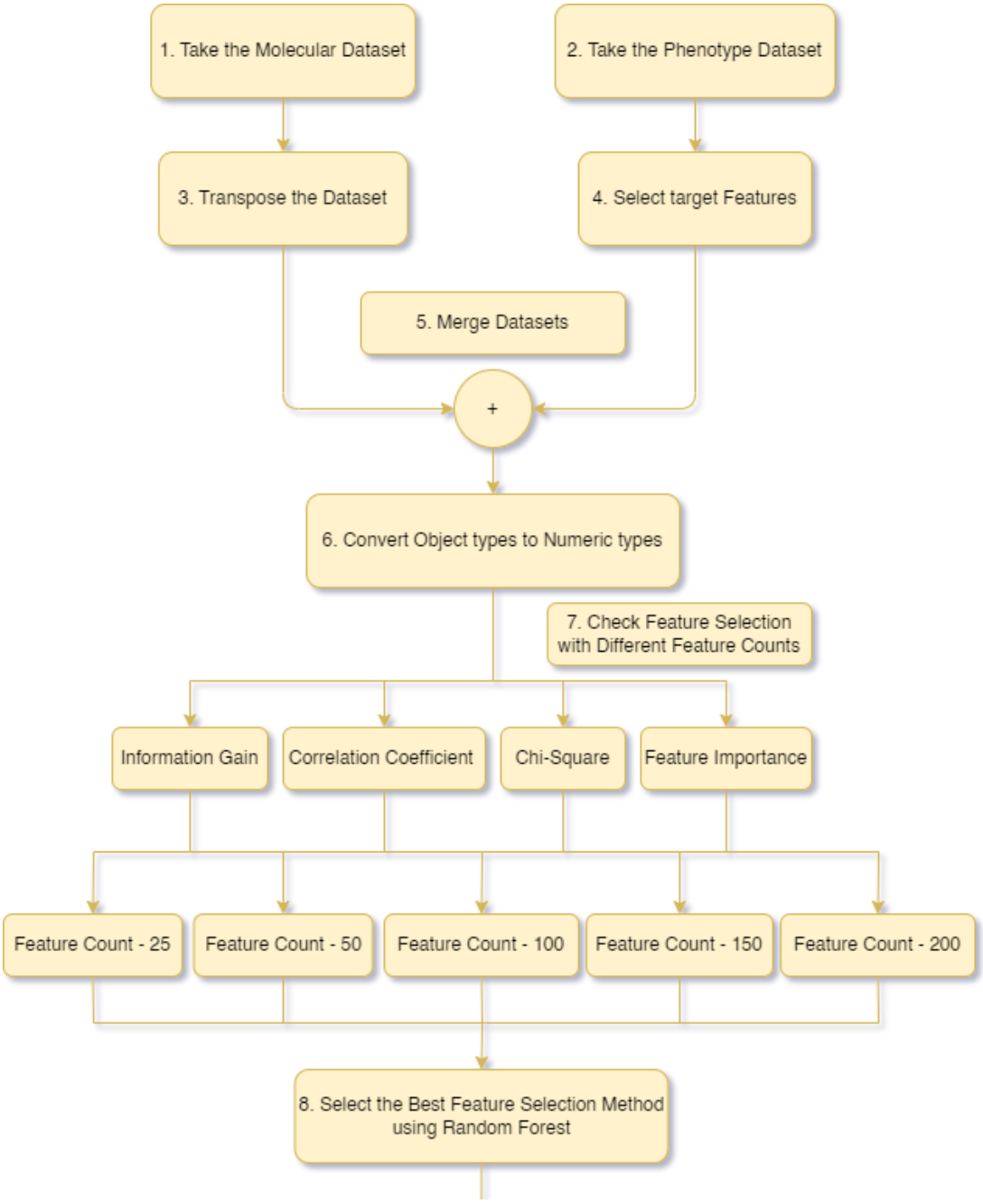


Figure 2 : Flow Chart Representation for Experimental Representation Part 1

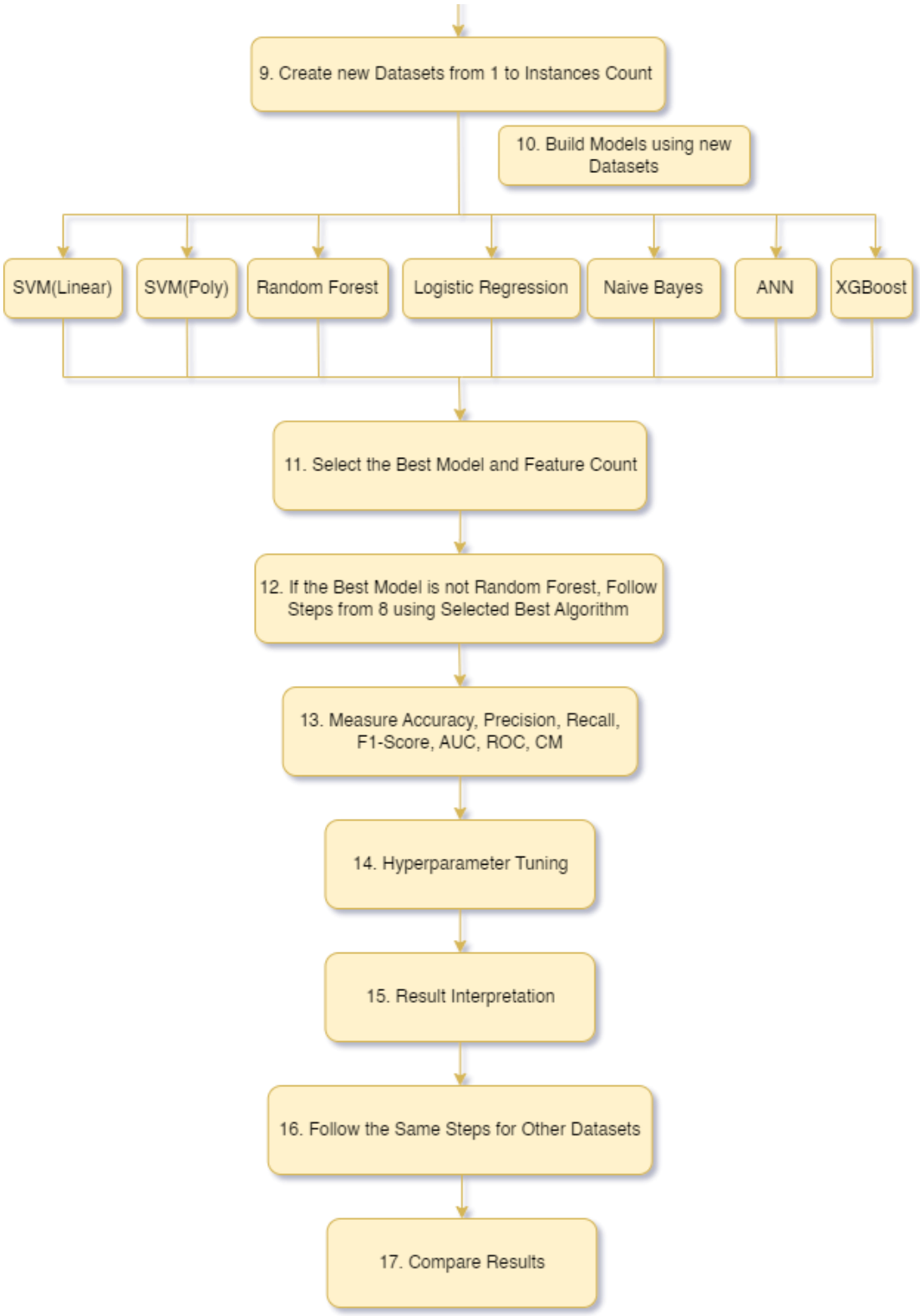


Figure 3 : Flow Chart Representation for Experimental Representation Part 2

CHAPTER 5 : Experimental Result

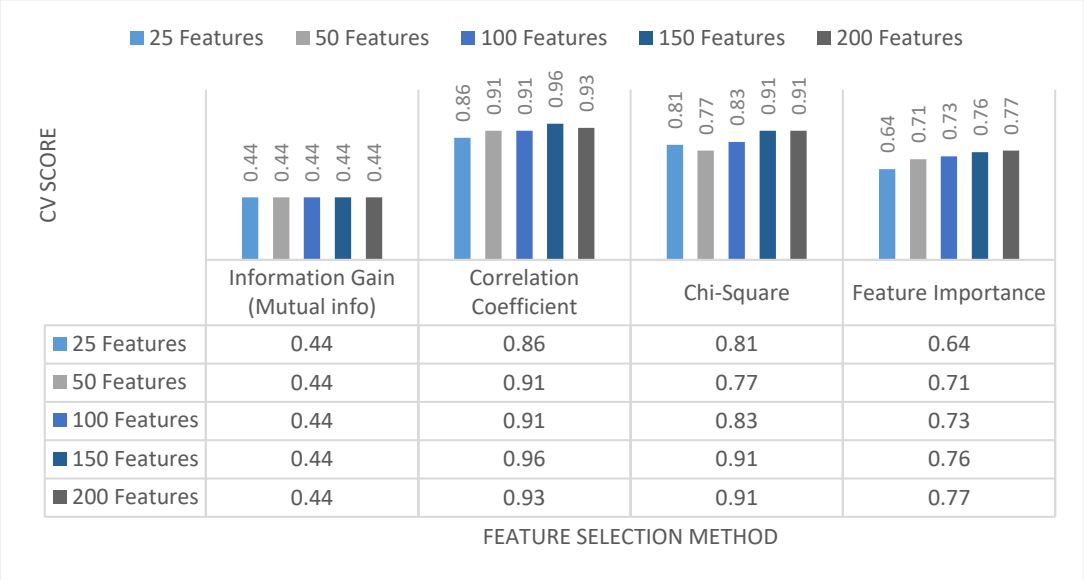
5.1 Introduction

Initially, we observed cross-validation values for the Random Forest model trained using various feature selection methods (Information Gain, Correlation Coefficient, Feature Importance, Chi-square). We assessed the performance of each algorithm by selecting feature counts of 25, 50, 100, 150, and 200, determining which one yielded the highest accuracy for each count.

Then we recorded cross-validation scores for multiple models (Random Forest, Logistic Regression, SVM, ANN, XGBoost, Naive Bayes) across diverse datasets, ranging from feature count = 1 to the total number of instances in each dataset. Subsequently, we identified and presented the top 10 results for each dataset based on their cross-validation values.

5.2 Experimental Result

DNA Methylation Dataset



	Information Gain (Mutual info)	Correlation Coefficient	Chi-Square	Feature Importance
25 Features	0.44	0.86	0.81	0.64
50 Features	0.44	0.91	0.77	0.71
100 Features	0.44	0.91	0.83	0.73
150 Features	0.44	0.96	0.91	0.76
200 Features	0.44	0.93	0.91	0.77

Table 2 : Feature Selection Results Chart for DNA Methylation Dataset

Model Name	Feature Count	CV Score (Mean +/- Std)
Logistic Regression	48	0.9571 +/- 0.0350
SVM(linear)	50	0.9571 +/- 0.0350
SVM(linear)	51	0.9571 +/- 0.0350
SVM(linear)	52	0.9571 +/- 0.0350
SVM(linear)	53	0.9571 +/- 0.0350
SVM(linear)	54	0.9571 +/- 0.0350
SVM(linear)	61	0.9571 +/- 0.0350
SVM(linear)	48	0.9429 +/- 0.0535
ANN	49	0.9429 +/- 0.0535
SVM(linear)	60	0.9429 +/- 0.0286

Table 3 : Model Selection Results Chart for DNA Methylation Dataset

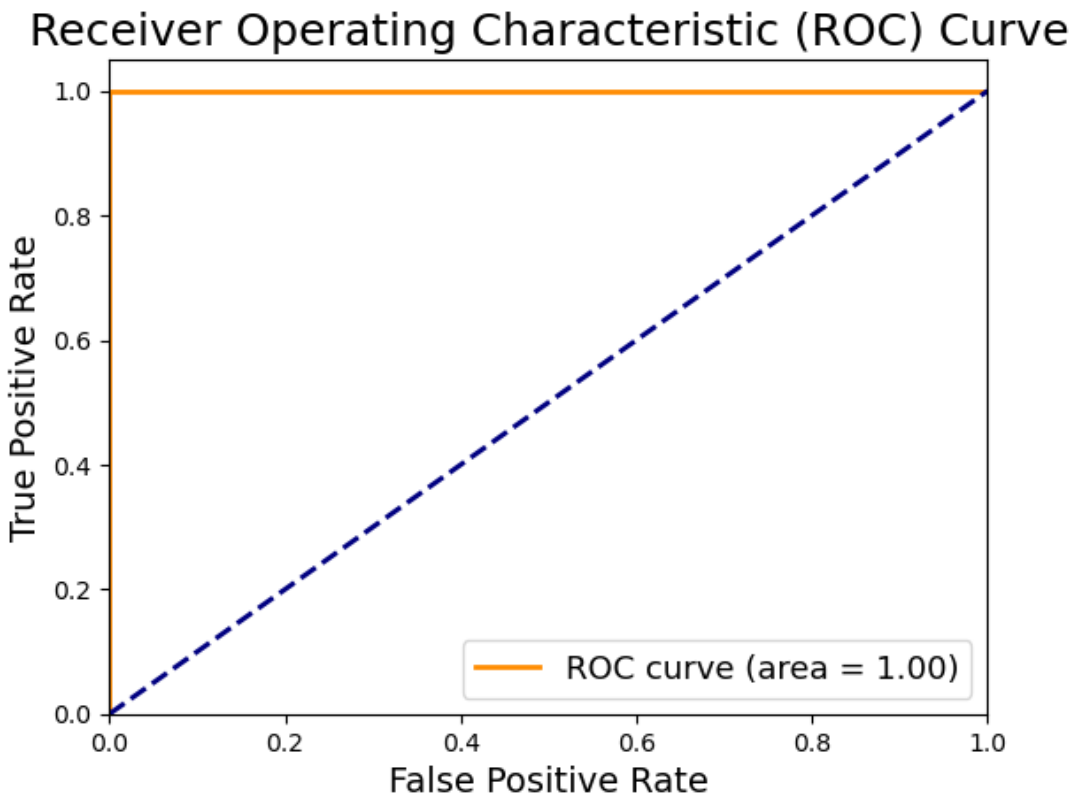


Figure 4 : ROC Curve for Best Model on DNA Methylation Dataset

RNA-SEQ(smallRNA) Dataset

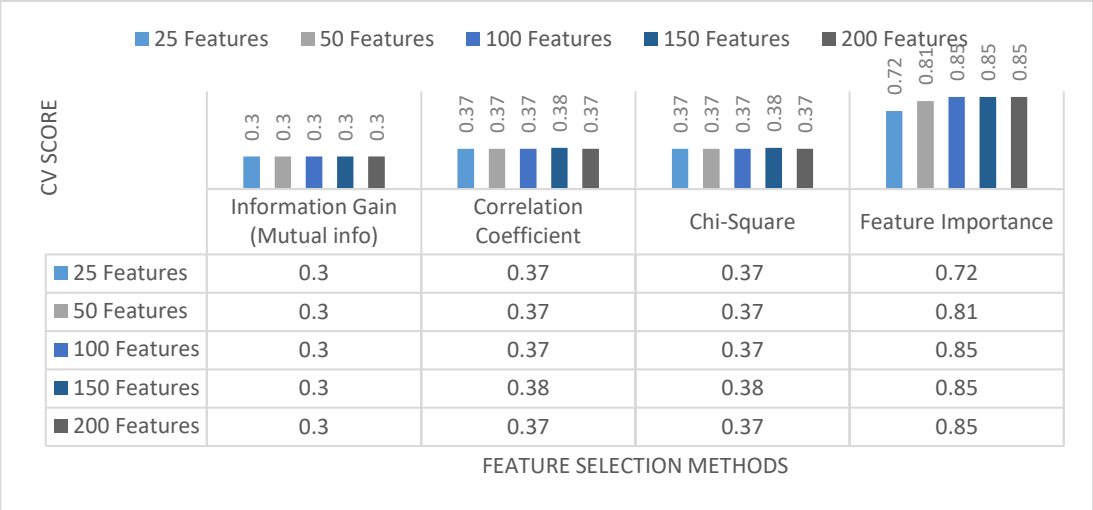


Table 4 : Feature Selection Results Chart for smallRNA Dataset

Model Name	Feature Count	CV Score (Mean +/- Std)
ANN	59	0.9500 +/- 0.0468
ANN	72	0.9500 +/- 0.0468
ANN	73	0.9500 +/- 0.0468
ANN	74	0.9500 +/- 0.0468
ANN	75	0.9500 +/- 0.0468
ANN	76	0.9500 +/- 0.0468
ANN	77	0.9500 +/- 0.0468
ANN	78	0.9500 +/- 0.0468
ANN	79	0.9375 +/- 0.0559
SVM(poly)	43	0.9375 +/- 0.0559

Table 5 : Model Selection Results Chart for smallRNA Dataset

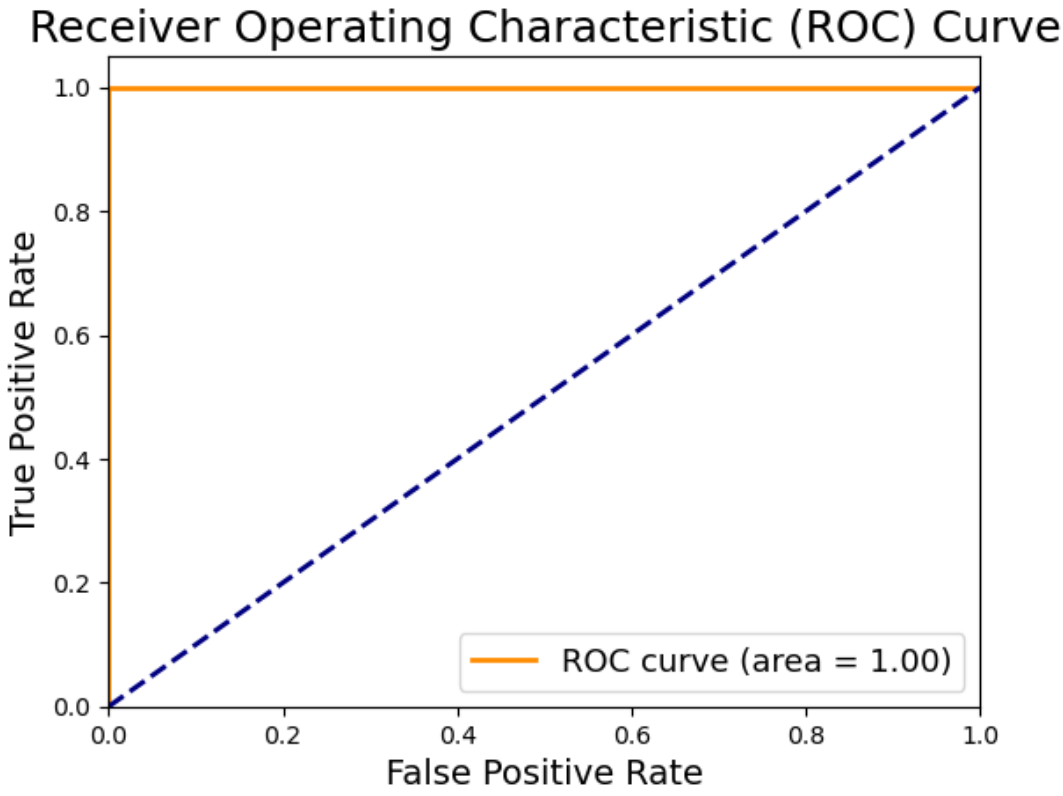
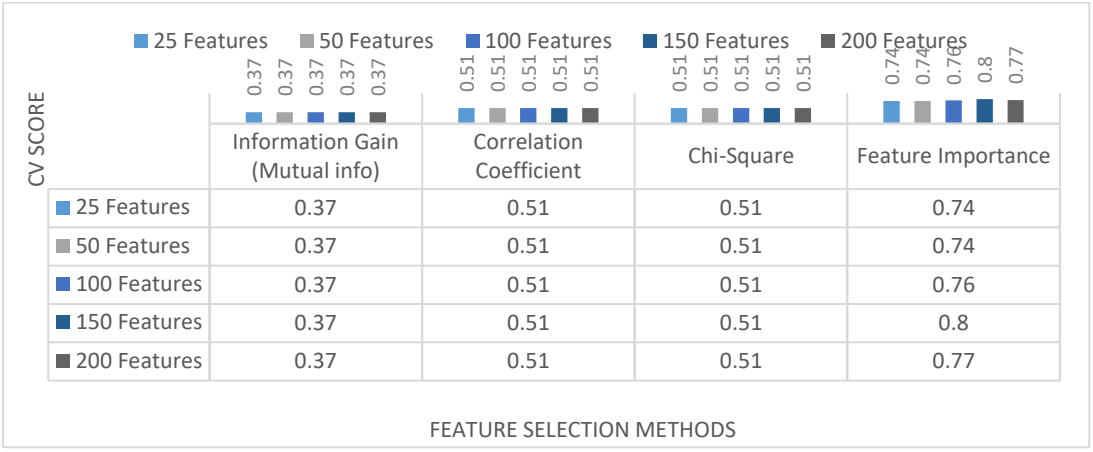


Figure 5 : ROC Curve for Best Model on smallRNA Dataset

RNA-SEQ(totalRNA) Dataset



	Information Gain (Mutual info)	Correlation Coefficient	Chi-Square	Feature Importance
25 Features	0.37	0.51	0.51	0.74
50 Features	0.37	0.51	0.51	0.74
100 Features	0.37	0.51	0.51	0.76
150 Features	0.37	0.51	0.51	0.8
200 Features	0.37	0.51	0.51	0.77

Table 6 : Feature Selection Results Chart for totalRNA Dataset

Model Name	Feature Count	CV Score (Mean +/- Std)
Naive Bayes	14	0.9625 +/- 0.0500
Naive Bayes	15	0.9625 +/- 0.0306
Naive Bayes	16	0.9625 +/- 0.0306
Random Forest	43	0.9500 +/- 0.0250
ANN	78	0.9375 +/- 0.0559
ANN	52	0.9375 +/- 0.0395
SVM(poly)	20	0.9375 +/- 0.0395
ANN	55	0.9375 +/- 0.0395
ANN	61	0.9375 +/- 0.0395
ANN	64	0.9375 +/- 0.0395

Table 7 : Model Selection Results Chart for totalRNA Dataset

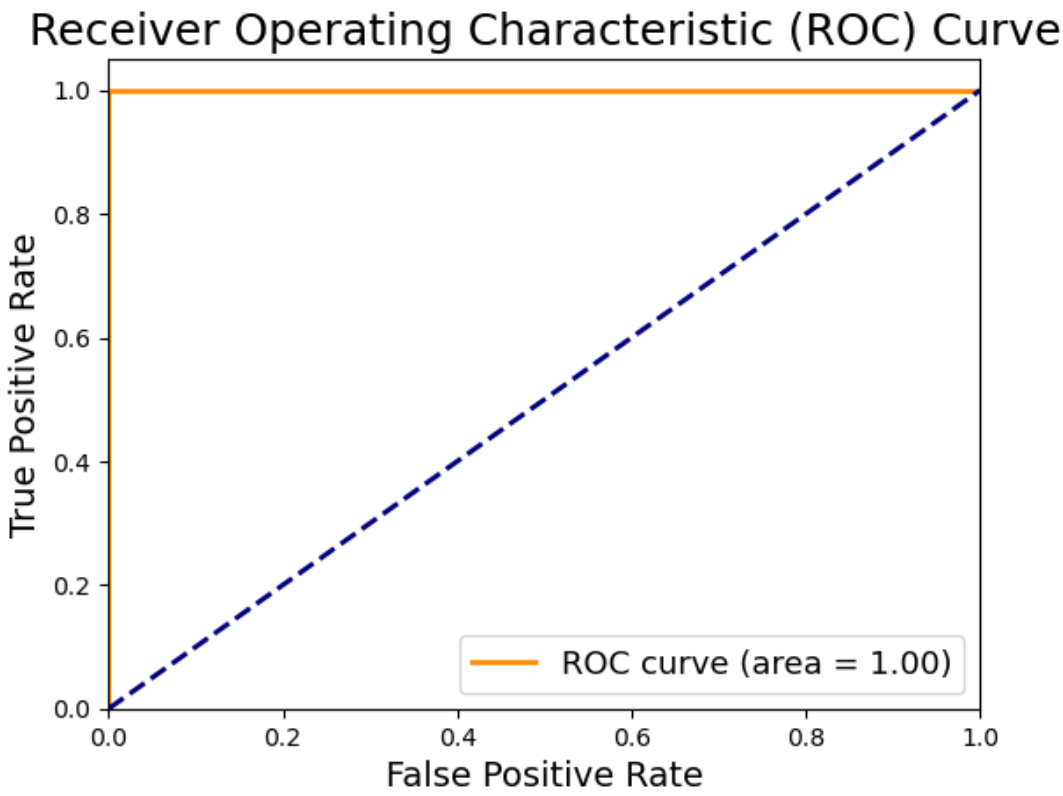


Figure 6 : ROC Curve for Best Model on totalRNA Dataset

CHAPTER 6 : Result Analysis

6.1 Introduction

To determine the optimal omic data type, a thorough comparison of numerical values generated by each model is essential. Simultaneously, it's crucial to discern the patterns inherent in the raw results, providing valuable insights into the distinctive characteristics of each data type.

6.2 Result and Discussion

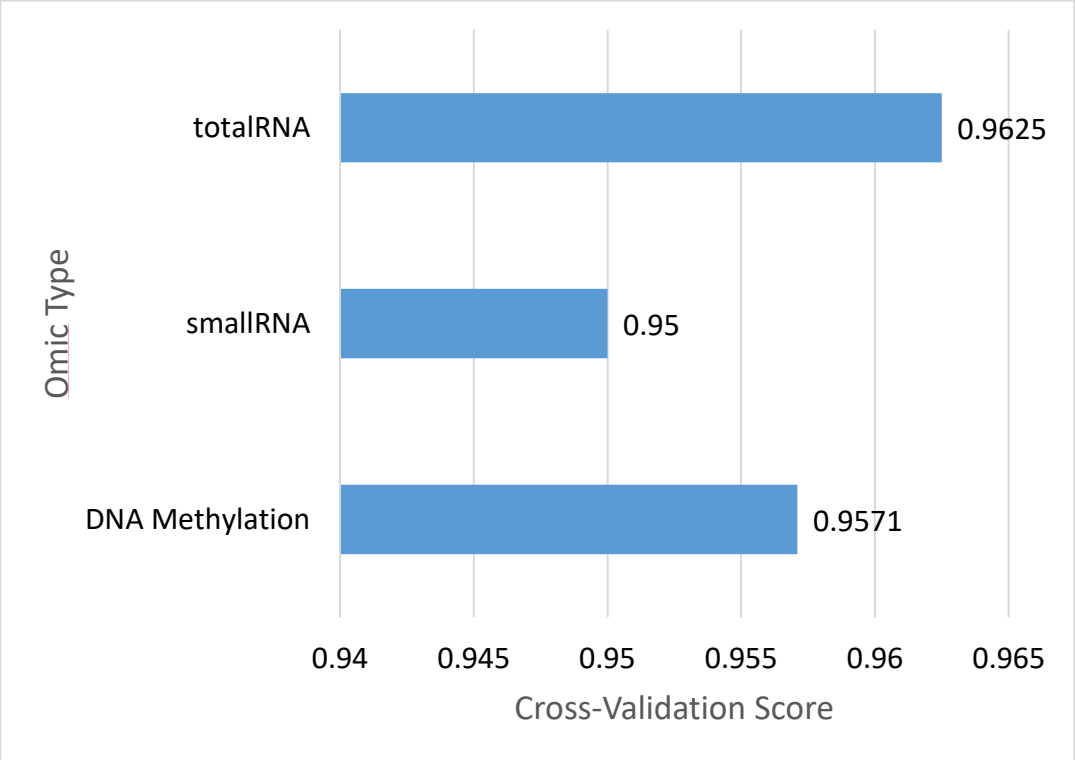


Figure 7 : Final Result Analysis of Omic Types

In the presented figure, the final result analysis encompasses three omic types: DNA methylation, small RNA, and total RNA. Notably, the highest cross-validation (CV) value is observed for total RNA, specifically 0.9625 +/- 0.05. Based on the cumulative findings, it is evident that total RNA stands out as the most effective omic type for the identification of Diabetic Retinopathy.

CHAPTER 7 : Conclusion Future Directions

7.1 Conclusion

In our research, we utilized three types of single omic datasets to build and evaluate six machine learning models. The objective was to determine which model, coupled with feature selection methods, yields the highest cross-validation (CV) scores, providing valuable insights into the accuracy of predictions. This exploration holds significant implications for advancing treatments for Diabetic Retinopathy, a condition affecting millions of individuals globally.

Future researchers can draw valuable insights from our study, particularly in understanding the selection of machine learning techniques and their impact on results. Moreover, as the field advances, exploring multi-omic datasets and comparing them with single omic datasets could offer a more comprehensive understanding of Diabetic Retinopathy, potentially influencing improved treatment strategies for patients worldwide. Our research contributes to the ongoing efforts to enhance the efficacy of treatments for this prevalent and impactful medical condition.

7.2 Future Work

we plan to focus on multi-omic data analysis to gain a more comprehensive understanding of Diabetic Retinopathy. While our current research utilized single-omic datasets such as totalRNA-seq, smallRNA-seq, the next step involves integrating these omic layers by collecting them from the same set of patients. This approach will enable us to perform detailed multi-omic analyses, allowing us to explore the interactions between these different data types. By doing so, we aim to uncover new insights into the molecular mechanisms underlying Diabetic Retinopathy, potentially leading to more accurate predictive models and improved treatment strategies.

REFERENCES

[1] Y. Sun, H. Zou, X. Li, S. Xu, and C. Liu, "Plasma Metabolomics Reveals Metabolic Profiling For Diabetic Retinopathy and Disease Progression," *Front Endocrinol (Lausanne)*, vol. 12, Oct. 2021, doi: 10.3389/fendo.2021.757088.

[2] D. Das, S. K. Biswas, and S. Bandyopadhyay, "A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning," *Multimed Tools Appl*, vol. 81, no. 18, pp. 25613–25655, Jul. 2022, doi: 10.1007/s11042-022-12642-4.

[3] M. Bader Alazzam, F. Alassery, and A. Almulihi, "Identification of Diabetic Retinopathy through Machine Learning," *Mobile Information Systems*, vol. 2021, 2021, doi: 10.1155/2021/1155116.

[4] S. Gupta, S. Thakur, and A. Gupta, "Optimized hybrid machine learning approach for smartphone based diabetic retinopathy detection," *Multimed Tools Appl*, vol. 81, no. 10, pp. 14475–14501, Apr. 2022, doi: 10.1007/s11042-022-12103-y.

[5] G. L. D'Adamo, J. T. Widdop, and E. M. Giles, "The future is now? Clinical and translational aspects of 'Omics' technologies," *Immunology and Cell Biology*, vol. 99, no. 2. John Wiley and Sons Inc, pp. 168–176, Feb. 01, 2021. doi: 10.1111/imcb.12404.

[6] A. Nomura, M. Noguchi, M. Kometani, K. Furukawa, and T. Yoneda, "Artificial Intelligence in Current Diabetes Management and Prediction," *Current Diabetes Reports*, vol. 21, no. 12. Springer, Dec. 01, 2021. doi: 10.1007/s11892-021-01423-2.

[7] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, "Machine learning in clinical decision making," *Med*, vol. 2, no. 6. Cell Press, pp. 642–665, Jun. 11, 2021. doi: 10.1016/j.medj.2021.04.006.

[8] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief Bioinform*, vol. 19, no. 6, pp. 1236–1246, May 2017, doi: 10.1093/bib/bbx044.

[9] V. V. Kamble and R. D. Kokate, "Automated diabetic retinopathy detection using radial basis function," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 799–808. doi: 10.1016/j.procs.2020.03.429.

[10] Z. W. Yu *et al.*, "High serum neuron-specific enolase level is associated with mild cognitive impairment in patients with diabetic retinopathy," *Diabetes, Metabolic Syndrome and Obesity*, vol. 13, pp. 1359–1365, 2020, doi: 10.2147/DMSO.S249126.

[11] M. Leeza and H. Farooq, "Detection of severity level of diabetic retinopathy using Bag of features model," *IET Computer Vision*, vol. 13, no. 5, pp. 523–530, Aug. 2019, doi: 10.1049/iet-cvi.2018.5263.

[12] J. H. Yun, J. M. Kim, H. J. Jeon, T. Oh, H. J. Choi, and B. J. Kim, "Metabolomics profiles associated with diabetic retinopathy in type 2 diabetes patients," *PLoS One*, vol. 15, no. 10 October 2020, Oct. 2020, doi: 10.1371/journal.pone.0241365.

[13] L. Math and R. Fatima, "Adaptive machine learning classification for diabetic retinopathy," *Multimed Tools Appl*, vol. 80, no. 4, pp. 5173–5186, Feb. 2021, doi: 10.1007/s11042-020-09793-7.

[14] S. Deuchler *et al.*, "Vitreous expression of cytokines and growth factors in patients with diabetic retinopathy- An investigation of their expression based on clinical diabetic retinopathy grade," *PLoS One*, vol. 16, no. 5 May, May 2021, doi: 10.1371/journal.pone.0248439.

[15] H. Y. Zhang, J. Y. Wang, G. S. Ying, L. P. Shen, and Z. Zhang, "Serum lipids and other risk factors for diabetic retinopathy in Chinese type 2 diabetic patients," *J Zhejiang Univ Sci B*, vol. 14, no. 5, pp. 392–399, May 2013, doi: 10.1631/jzus.B1200237.

[16] P. S. Graham *et al.*, "Genome-wide association studies for diabetic macular oedema and proliferative diabetic retinopathy," *BMC Med Genet*, vol. 19, no. 1, May 2018, doi: 10.1186/s12881-018-0587-8.

[17] W. Meng *et al.*, "A genome-wide association study suggests new evidence for an association of the NADPH Oxidase 4 (NOX4) gene with severe diabetic retinopathy in type 2 diabetes," *Acta Ophthalmol*, vol. 96, no. 7, pp. e811–e819, Nov. 2018, doi: 10.1111/aos.13769.

[18] V. B. Kolachalama, "Machine learning and pre-medical education," *Artif Intell Med*, vol. 129, Jul. 2022, doi: 10.1016/j.artmed.2022.102313.

[19] B. A. Mateen, J. Liley, A. K. Denniston, C. C. Holmes, and S. J. Vollmer, "Improving the quality of machine learning in health applications and clinical research," *Nature Machine Intelligence*, vol. 2, no. 10. Nature Research, pp. 554–556, Oct. 01, 2020. doi: 10.1038/s42256-020-00239-1.

[20] S. Goel *et al.*, "Deep Learning Approach for Stages of Severity Classification in Diabetic Retinopathy Using Color Fundus Retinal Images," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/7627566.

[21] S. Scholarship@western and S. Biswas, "Implications of long non-coding RNAs in the pathogenesis of Implications of long non-coding RNAs in the pathogenesis of diabetic retinopathy: a novel epigenetic paradigm. diabetic retinopathy: a novel epigenetic paradigm," 2020. [Online]. Available: https://ir.lib.uwo.ca/etdhttps://ir.lib.uwo.ca/etd/7116