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 OF 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
[JULY] 2023

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

Supervisor(s):

Supervisor : Dr. P. Jeyananthan

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

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	✓	
2.2.1 Forecasting Models	✓	
2.2.2 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	✓	✓
3.2.1 Data Selection	✓	
3.2.2 Data preprocessing		✓
3.2.3 Feature selection	✓	✓
3.2.4 Apply 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 Phenotype Data Selection	✓	
4.2.2 Dataset Selection	✓	
4.3 Database Preparation	✓	✓
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	:	Ultrahigh-Performance Liquid Chromatography Mass Spectrometry
VGG	:	Visual Geometry Group

CHAPTER 1: Introduction

1.1 Motivation and Overview

Diabetes mellitus (DM) is becoming more common in emerging and wealthy nations. Six hundred twenty-nine million people will have diabetes worldwide by 2045, according to estimates [1]. A medical disorder called diabetic retinopathy (DR) is brought on by diabetes mellitus (DM). The impairment of glucose metabolism and various micro- and macrovascular abnormalities that result in DM make it a chronic condition. One of DM's most prevalent and dangerous side effects, DR, causes severe blindness by deforming the human retina [2]. Early detection and accurate diagnosis of DR are essential for prompt intervention and efficient disease treatment. The detection and categorization of several disorders, including DR, have improved because of developments in molecular data analysis tools in recent years [2][3]. A thorough and comparative examination of all available molecular data modalities is still required to ascertain which ones provide the most dependable and accurate way of diagnosing DR [4].

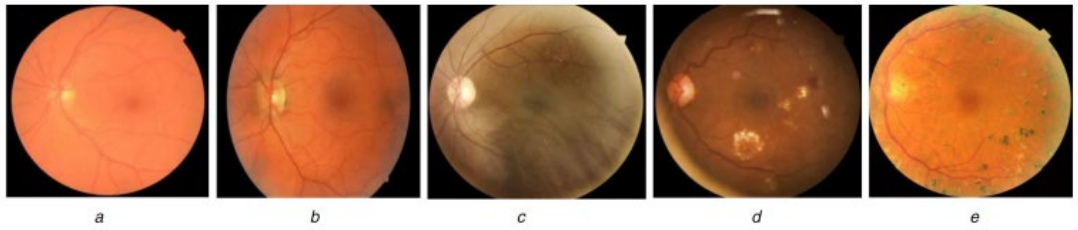


Figure 1 : Images arranged in increasing severity levels of DR [11]
(a) No DR, (b) Mild NPDR, (c) Moderate NPDR, (d) Severe NPDR, (e) PDR

Addressing this significant information gap is the driving force behind the present research issue. We hope to advance significantly in medical diagnostics and customized therapy by comparing and assessing diverse molecular data types in the context of DR identification [7]. Creating more accurate and effective DR detection systems can be facilitated by an awareness of the advantages and disadvantages of various data modalities, which may enhance patient outcomes and advance medical procedures [9].

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

The main aim of our research project is to examine and contrast the effectiveness of several molecular data sets in detecting diabetic retinopathy (DR) using machine learning techniques. Diabetic retinopathy is a significant diabetes complication that is still one of the primary causes of blindness globally. Early detection and proper diagnosis of DR are critical for timely intervention and visual loss prevention. Molecular data, such as gene expression profiles, protein biomarkers, and epigenetic markers, hold much promise for improving DR diagnostic accuracy.

Our research's main objectives are:

- To collect and arrange relevant diabetic retinopathy molecular data from publicly available databases and clinical sources.
- To investigate the utility of gene expression patterns, protein biomarkers, and epigenetic markers in detecting and distinguishing diabetic retinopathy.
- To assess and compare the performance of various machine learning methods in identifying diabetic retinopathy based on different molecular data, such as support vector machines, random forests, and deep neural networks.
- To find the best mix of molecular data and machine learning techniques for detecting diabetic retinopathy accurately and early.
- To, through data analysis and interpretation, provide insights into the underlying molecular mechanisms and pathways linked with diabetic retinopathy.

1.3 Research Scope

Our research will concentrate on collecting and preparing various molecular datasets, such as gene expression profiles, protein markers, and other pertinent molecular information on diabetic retinopathy. Machine learning methods will be assessed and implemented based on the molecular data collected to measure their accuracy and performance in detecting diabetic retinopathy. Ethical issues will be considered, and the study will be carried out within a specific timeline, considering data availability and sample size constraints. The purpose of the research is to shed the spotlight on possible biomarkers for diabetic retinopathy detection and to make suggestions for using machine learning-based diagnostic tools in clinical settings.

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. This is because retinal imaging is non-invasive and can give detailed visual information about the retina [9][13]. Several studies have examined how artificial intelligence (AI) and machine learning algorithms can be used to look at retinal pictures and predict whether or not someone has diabetic retinopathy and how bad it is. These AI-based methods have successfully found diabetic retinopathy in retinal images with high accuracy and sensitivity [6]. Even though these methods work, it is still hard to tell the difference between different stages of the disease, like non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR) [11] [14].

Understanding the chemicals in the blood of people with diabetic retinopathy is essential because they may have something to do with how the disease worsens [1][21]. By figuring out what these substances are and how they contribute to the growth of DR, we can learn important things that can help us develop focused, therapeutic interventions and personalized treatment plans [19]. So, using plasma metabolites, amino acids, and other molecular markers can help researchers learn more about the disease and how it works [1] [10].

- 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's different about the substances in the blood of people with diabetic retinopathy and how these chemicals might be linked to how the disease gets worse [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 more about possible metabolite signs of 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 [5]. 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 be 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.

Also, different ways of classifying DR have been looked into. For example, [4] suggested an optimized hybrid ML classifier that combined neural networks (NN) and deep convolutional neural networks (DCNN) to accurately classify the severity level of DR using smartphone-based retinal imaging. This method showed that portable gadgets could be used to test and keep track of diabetic retinopathy, especially in places with few resources.

Also, much research has been done on the genetic causes of diabetes retinopathy. In a study by [16], researchers found genetic risk factors for different kinds of DR in people with type 2 diabetes who are of European descent. By genotyping single nucleotide polymorphisms (SNPs) all over the genome, the study found the most important genetic differences that cause severe diabetic retinopathy. These results show how vital DNA screening is in high-risk groups to find people who are likely to get sick early and take steps to protect them.

Also, [17] looked into the role of the NADPH Oxidase 4 (NOX4) gene 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 finding the most promising biomarkers and predictive models, we can give healthcare workers the tools they need to treat DR as early as possible and, in the end, improve patient outcomes.

2.3 Performance Analysis

The researchers aimed to compare how well different genetic data could be used to spot diabetic retinopathy using machine learning. Statistical methods like LASSO and logistic regression were used to examine the molecules in the plasma of people with diabetic retinopathy [1]. An AUC of 0.80 was found for the risk score that was made with logistic regression [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]. This study shows how machine learning could be used to find diabetic retinopathy and how important it is to teach machine learning to future doctors [18].

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 [11]

2.4 Research Gap

The paper "Automated diabetic retinopathy detection using radial basis function" 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. [9]

The performance comparison of different molecular data in the identification of diabetic retinopathy remains unexplored in the context of the current reliance on retinal images for diagnosis. Although the automated approach using a radial basis function demonstrated promising results, there is a need to investigate the effectiveness of incorporating diverse molecular data, such as genetic markers,

proteomic profiles, or other biomolecular information, in detecting and predicting diabetic retinopathy. Furthermore, utilizing a small sample size database and a single type of neural network in the existing research highlights the importance of exploring more extensive and diverse datasets and employing various advanced machine learning algorithms to enhance the accuracy and robustness of DR detection methods. Addressing these gaps can improve diagnostic techniques and patient outcomes in managing diabetic retinopathy.

2.5 Available Databases

Peripheral venous blood samples [1]

- including 42 DR patients and 32 T2DM patients without DR

Imbalanced dataset [2]

- 7137 images for single-lesion and multi-lesion detection.

DIARETDB1,e-ophtha datasets, MESSIDOR dataset, and Kaggle dataset[2]

- These datasets used for different kinds of ML and DL models for image analysis

APTOS-2019-Blindness-Detection, and EyePacs datasets [4]

- To generate retina images by simulating the field of view for a device using the retina images from APTOS-2019-Blindness-Detection, and EyePacs datasets.
- For both left and right eyes, there are 35,126 retina images presented in the EyePacs dataset.
- This is another dataset produced by the Asia Pacific Tele-Ophthalmology Society (APTOS) as part of the 2019 blindness finding competition. Three thousand six hundred forty-eight high-resolution fundus images were selected from the Kaggle dataset of 3661 images by different designs and cameras type in manifold health centers over a lengthy period.

IDRID and ILSVRC datasets [20]

- The first dataset prepared for the Indian population for detecting eye disease is the IDRID dataset.
- Out of thousands of images, experts verified 516 images to form a dataset based on adequate quality, clinical relevance, and no duplication of images.
- The dataset includes typical DR lesions and some normal retinal structures.
- The dataset highlights facts related to the stage of the disease along with the severity of the disease.
- The dataset contains eighty-one color fundus dataset images with the DR trace capability.
- The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset trains the VGG16 model, with over 122 color fundus images, fine-tuned on the IDRID dataset using transfer learning.

Clinical imaging, EHRs, genomics, and wearable device datasets.[8]

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

Dataset DIARETDB0 and DIARETDB1 [9]

- with 130 and 89 images examined in this experimental work to detect early diabetic retinopathy.
- Dataset DIARETDB0 & DIARETDB1 gets accessible by a site www.it.lut.fi/project/imageret [9].

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

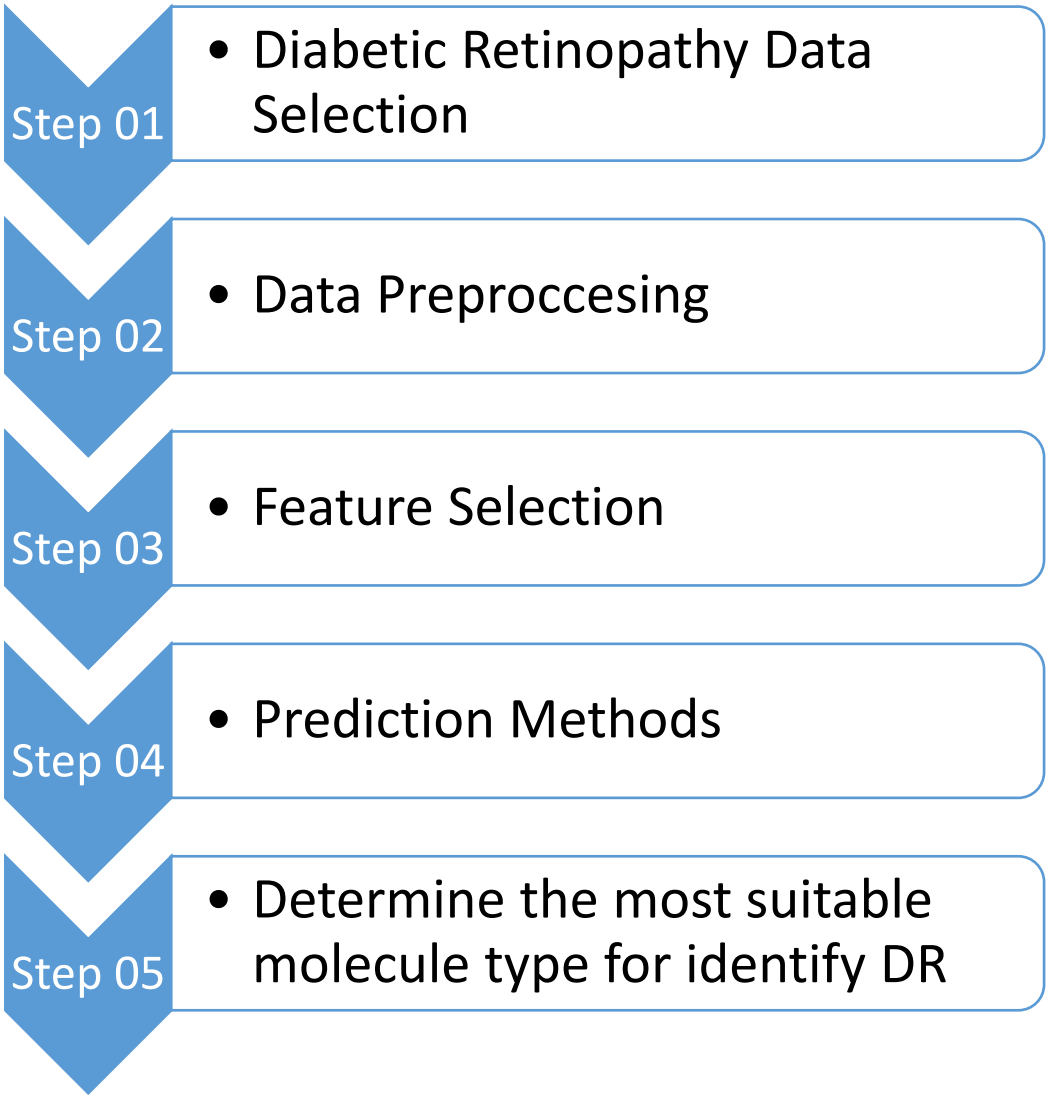
- In the paper, they showed all datasets in Table 1

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



3.2 Detailed Methodology

3.2.1 Data Selection of Diabetic Retinopathy

Detecting and characterizing diabetic retinopathy (DR) by molecular data analysis is critical for understanding the underlying processes and creating appropriate diagnostic and therapeutic techniques. 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) hosts GEO, which has a large and diversified collection of publicly available gene expression data. GEO is a storehouse of genomic information relating to diverse biological processes, illnesses, and experimental circumstances that researchers worldwide contribute to.

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 are decided to get these types of datasets related to DR

- DNA methylation
- RNA-Seq
    - totalRNA
    - smallRNA

We are taken 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.

- totalRNA
- smallRNA

3.2.2 Data preprocessing

Data preprocessing is essential to preparing molecular data for machine learning (ML) analyses, as mentioned in the sources [1, 2, 3]. UPLC-MS finds chemical substances like amino acids in plasma products [1]. Normalization 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.

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 Importances. 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. The model will be helpful and applicable if the features are chosen well.

- Information Gain (mutual_info_classif)

Information Gain measures how much information a target variable gains when a specific feature is present. 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. Mutual_info_classif is a version of Information Gain that is used for jobs that need to be sorted.

- 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. Pearson correlation is often used to measure the strength and direction of a linear relationship between two factors when the data are continuous.

- 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 Importances

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

Forward Feature Selection is a way to gradually choose features that add features to the model. It 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 uses trained models to make predictions or choices about data that has not yet been seen. The models learn patterns from training data that has been labeled, and then they use those patterns to guess what will happen or put new data into specific categories. The idea is to make accurate predictions on data that has never been seen before. This shows that the model can generalize and work well in real-world 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 closest neighbors. K-NN is easy to understand and doesn't use parameters, but it can be sensitive to noisy data and needs K to be tuned carefully. 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.

- Convolutional Neural Network (CNN)

CNN is a design for deep learning that is mainly used for tasks like recognizing images and videos. Using convolutional layers, it uses images to learn hierarchical traits automatically. These layers find edges, patterns, and shapes, which are then used by later layers to spot more complicated objects. CNN is very good at classifying images, finding objects, and recognizing faces because it can learn hierarchical representations.

- Recurrent Neural Network (RNN)

An RNN is a deep learning model for handling sequential data, like time series and natural language data. RNNs use feedback loops to keep track of their hidden states, which lets them find dependencies and patterns in a series of data sets. Traditional RNNs, on the other hand, have problems with gradients that disappear or explode during training. This led to the creation of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) structures. RNNs are often used for speech recognition, computer translation, and figuring out how people feel about something.

- 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.

- K-Means Clustering

K-Means Clustering is an unsupervised learning method that groups data into K different groups. It gives each data point, one at a time, to the cluster whose center (mean) is closest to it. The goal of the method is to make the sum of the squared distances between data points and the centers of their clusters as small as possible. K-Means are often used to reduce data, divide images into groups, and divide customers into groups in marketing. It needs the number of clusters (K) to be set up front, and its performance can change based on where the cluster centers are put in the beginning.

3.2.5 Compare performance

In machine learning, it's important to compare how well different models do to choose the best one for a given job. Different metrics are used to compare to measure how good each model is at making predictions. It helps determine each model's strengths and flaws and how well it works. By comparing models, researchers and practitioners can evaluate which works best in predicting and applying data they haven't seen yet. Careful comparison lets people make decisions based on data, which leads to using the most accurate and reliable model for a given problem. This helps machine learning users in many fields move forward and succeed.

- 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.

- 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.

- 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 : Progress To Date

4.1 Literature Review

We've spent much time looking at research articles, books, and educational websites, all of which relate to our study topic. More than 21 pieces have been carefully looked at and reviewed. The literature review will remain an essential part of our study.

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.

4.3 Database Preparation

In our work on identifying diabetic retinopathy (DR) through omics data, we started by getting the data sets we needed from the web. Then, we used clinical data to determine if the patients had been identified with DR and how far along they were in the disease. With the help of Python code, we combined these files and added the new information to make a complete database. Since our study identifies DRs using omics data, we have worked with more than one data set. Each dataset was downloaded independently and saved in its file to be analyzed and put together in the future.

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 edema 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