Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Software Department

Prediction Stages of Type 1 Diabetes Using Machine Learning Methods

A Thesis

Submitted to the Council of the College of Information Technology for Postgraduate Studies of the University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology/Software

By

Noor Ali Fadhel Mahdi

Supervised by

Asst. Prof. Dr. Sura Zaki Alrashid

2023 A.C. **1444 A.H.**


بسم الله الرحمن الرحيم

﴿ لَئِن شَكَرْتُمْ لَأَزِيدَنَّكُمْ ﴾


صدق الله العظيم

سورة ابراهيم الأيه ٧

Supervisor Certification

I certify that this thesis is prepared under my supervision at the Department of Software / Collage of Information Technology / Babylon University, by *Noor Ali Fadhel Mahdi* as a partial fulfillment of the requirements for the degree of Master in Information Technology

Signature:
Supervisor Name: Dr. Sura Zaki Al-Rashid
Date: / / 2023

The Head of the Department Certification

In view of the available recommendations, I forward this thesis for debate by the examination committee.

Signature:
Prof. Dr. Ahmed Saleem Abbass
Head of Software Department
Date: / / 2023

Acknowledgements

In the name of ALLAH, the most compassionate, the most merciful, praise be to ALLAH and prays and peace are on his prophet Mohammed (Allah's blessing him) and peace be on his relatives. All praise be to ALLAH Almighty who enabled me to complete this task successfully.

All thanks and respect to my supervisor Dr. Sura Zaki Al-Rashid, for her guidance, and supervision during this work.

I must express my very profound gratitude to all my family especially my husband for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

I would like to express my special thank goes to the college of Information Technology staff for the great cooperation they have offered me.

Finally, I would like to thank all the kind, helpful and lovely people who helped me directly or indirectly to complete this work and apologize to them for not being able to mention them by name here, but their support will always be cherished.

Abstract

Type 1 diabetes disease (T1D) is a chronic disease when destruction occurs in the beta cells of the pancreas, pancreas releasing minimal or no insulin. Glucose may enter cells through the hormone insulin, which the body utilizes to make energy. The major problem with T1D is an incurable disease, treatment focuses on controlling blood glucose levels. T1D can be inherited through genes, thus, genetic factors greatly impact the development of T1D. According to current investigations, Biological researchers analyzed this disease using statistical analysis methods, and researchers in the bioinformatics field classified T1D using various feature selection methods and some of machine learning methods for finding the genes (features) that cause this disease might assist regulate the condition of the patient and support early treatment. Further, with the advantages of modern technology, the microarray is a helpful tool that enables scientists to concurrently quantify the expression levels of hundreds of thousands of genes.

The main aim of this thesis is to build a model that predicts T1D progression with the possible highest accuracy and the least error by selecting the most important genes (informative genes). The proposed system consists of two main stages: the select genes stage and classifying and regression stages. The process of gene selection is performed using two parts, parallel and sequential. This process is used to select a subset of the important genes and improve the prediction accuracy of the proposed model. Feature selection methods include four methods: Mutual Information, Analysis of Variance, Chi^2, and Principal Components Analysis.

Generally, the sequential feature selection part of the proposed system learns which genes are most informative at each stage and selects the next

genes based on the previously selected genes. These are then entered into the model to confirm their significance and the accuracy that would be achieved by utilizing these genes. While the second part of the proposed system performs the parallel feature selection part which selects informative genes by each method then the selected genes are entered into the model.

Furthermore, this work sought to provide a prediction model based on classification and regression models to define the difference between patients with T1D from subjects with genetically non-carriers. The available dataset has been used to achieve the objectives of the present thesis namely: the T1D dataset. The evaluation has been performed depending on measures of prediction (accuracy, recall, precision, and RMSE). The results show that the performance of the proposed system is effective, where the prediction accuracy reached (0.100) four times from the sequential part for classification models, one time from Random Forest, and three times from the Support Vector Machine. The lowest RMSE reached (0.00) from the Linear Regression model using (the without replicate data).

Declaration

I as a result of this declare that this dissertation entitled "Prediction Stages of Type 1 Diabetes Using Machine Learning Methods", submitted to the University of Babylon in partial fulfilment of requirements for the degree of Master in Information Technology \ Software, has not been proposed as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose source is appropriately cited in the references.

Signature:

Name:

Date: / / 2023

Dedication

To my husband (Dr.Aladdin Al-Sharefi) who facilitates my life.
To my family who supported me, and finally, to my supervisor Assist. Prof. Dr. Sura Zaki AL-Rashid

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Table of Contents

List of Tables

List of Figure

List of Figure

List of Abbreviations

Abbreviation	Meaning
Acc	Accuracy
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
BPSO-DT	Binary Particle Swarm Optimization With A Decision Tree
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
FN	False Negative
FP	False Positive
GEO	Gene Expression Omnibus
HC	Health Control
HLA	Human Leukocyte Antigen
KNN	K-Nearest Neighbors
L1-LR	L1-regularized LR
LASSO	Least Absolute Shrinkage and Selection
lncRNA	long non-coding ribonucleic acid
LoHR	Longitudinal High Risk
LoLR	Longitudinal Low Risk
LR	Linear Regression
LR1	Logistic Regression
MI	Mutual Information
ML	Machine Learning
MLR	Multivariate Linear Regression
mRNA	Messenger Ribonucleic Acid
NB	Naïve Bayes

List of Abbreviations

Abbreviation	Meaning
NCBI	National Center for Bioinformatics Information
NN	Neural Network
PCA	Principle Component Analysis
RF	Random Forest
RMA	Robust Multi-Array Average
RMSE	Root Mean Square Error
RNA-seq	RNA Sequence
RO	Resent Onset
SCGRNs	Single-Cell Gene Regulatory Networks
SVM	Support Vector Machines
SVR	Support Vector Regression
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
TCGA	The Cancer Genome Atlas
TN	True Negative
TP	True Positive
VAE	Variational Auto Encoder

List of Algorithms

Chapter One

General Introduction

Chapter One

General Introduction

1.1. Introduction

According to the International Diabetes Federation (IDF), an estimated 463 million people worldwide had diabetes in 2019, and that number is projected to increase to 578 million by 2030 and 700 million by 2045. Type 1 diabetes (T1D) is one of the most common chronic diseases in children and adolescents and can begin at any age. The incidence and prevalence of T1D are steadily increasing, accounting for about 5-10% of people with diabetes [1].

While its natural history is not completely known, type 1 diabetes is a chronic polygenic illness with a multifactorial genesis that involves enduring genetic, autoimmune, and environmental components. The condition is characterized by hyperglycemia and impaired insulin secretion. Consequently, insulin treatment is necessary for life for those with T1D. T1D is frequently linked to higher mortality rates, challenges, and shorter life spans [2].

A complex interplay of polygenetic predisposition and environmental factors leads to type 1 diabetes (T1D). T1D is an extremely difficult condition as a result of its many causes. Many problems have previously been linked to a variety of factors including the incidence of autoimmunity, beta cell death pathways, and genetic changes linked to either one (autoimmunity) or the other (beta cell death) or both [3]. Individuals with type 1 diabetes are often younger, with an average age under 30. Increased blood sugar levels, frequent urination, and increased appetite are typical clinical signs. Patients with this kind of diabetes need insulin treatment because oral medications alone cannot treat it. Machine learning has been

used in various areas of medical health due to its rapid progress, including T1D disease [4].

An interdisciplinary field called bioinformatics creates techniques and software tools to illustrate biological facts and structures, which are frequently represented by a large amount of information. Bioinformatics is a synthesis of several disciplines that assesses and presents biological and genetic data, including software engineering, statistics, computer science, and engineering sciences [5]. Global efforts to advance bioinformatics led to the formation of computer networks that simplified the entry of biological data and aided in the development of basic analytic algorithms. Access to several worldwide projects that offer gene and protein databases over the Internet is open to the whole scientific community [6].

A system for evaluating medical data to identify illnesses is made possible by machine learning techniques, which are becoming more and more crucial in the healthcare industry. The intention is to make it easier for medical professionals to diagnose illnesses. Machine learning is a technique that enables computers to automatically learn from experience and improve. The development of computer programs that can access data and utilize that data through a number of algorithms to decide for themselves what actions to take in response to that data is known as machine learning [7].

1.2. Thesis Motivation

The specific death of insulin-producing beta cells in the pancreas causes type 1 diabetes (T1D), a chronic autoimmune illness. The complicated processes that cause illness beginning and progression—which entail the breakdown of many tolerance networks—have not yet been completely understood [3]. The International Diabetes Federation

believes that the prevalence of type 1 diabetes among children and adolescents under the age of 15 is rising globally, with notable regional variations, at an estimated rate of roughly 3% annually. According to estimates, type 1 diabetes affects more than 96,000 children and teenagers under the age of 15 each year, rising to more than 132,600 by the age of 20. It is estimated that more than one million children and adolescents under the age of 20 suffer from type 1 diabetes worldwide [8].

1.3. Problem Statement

Gene expression data plays a crucial role in understanding the underlying molecular mechanisms and identifying potential biomarkers or therapeutic targets for T1D. The problem is to develop an accurate and reliable machine learning model that can predict gene expression patterns in individuals with Type 1 Diabetes (T1D) using a given dataset. The dataset contains genetic information and corresponding gene expression levels from T1D patients, with the goal of understanding the underlying molecular mechanisms and identifying potential biomarkers associated with T1D.

The successful completion of this project will contribute to a better understanding of T1D at the molecular level, potentially leading to the discovery of novel therapeutic targets and personalized treatment strategies for individuals with T1D, and it could have significant implications for early diagnosis and personalized treatment. The primary issue addressed in this thesis is "the prediction of T1D ".

1.4. Related Works

Gene expression offers the possibility of diagnosing different diseases, by using precise Deoxyribonucleic Acid (DNA) arrays containing

powerful gene expression data acquisition technology. In many different sectors, research has been conducted to improve the quality of care for patients with diabetes and reduce its effects, including artificial intelligence and machine learning. In several studies, ML models used to predict and classify diabetes have been mentioned.

In [2] an identification marker genes of incident T1D in PBMC of children via an ML analytic strategy attuned to the high dimensional structure of microarrays, with downstream analyses providing high biological plausibility. Data used are available on the open-source National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) platform with the unique GEO accession ID of GSE9006. Gene expression values of this curated dataset were $\log_2$ Normalized to create symmetric distributions devoid of skewness. 16 dimension-reduction algorithms had been applied belonging to three categories after removing redundancies, the collated feature set produced by all algorithms contained 1003 genes. Machine learning with hyper parameter tuning was performed on the reduced feature set using four algorithms, namely, multivariate adaptive regression splines (MARS), adaptive boosting (AdaBoost), random forests (RF), and extreme gradient boosting dropouts meet multiple additive regression trees (XGB-DART) with an accuracy of 100% for each model .

In [9], the gene expression dataset (GSE55098) for type 1 diabetes (T1D) is used to classify the immune cell-related genes that have a role in the occurrence and development of T1D and possibly contribute to finding an immunotherapy for it. Hub genes are investigated using least absolute shrinkage and selection factor (LASSO) and support vector machines (SVM). Gene prognostication in T1D is evaluated using receptor operating

characteristic curves and the highest value is area under the ROC curve (ROC) AUC = 0.918 for the immune-related genes.

In [10] , Genes have been identified as the most frequently reported and differentially expressed genes (DEGs) in diabetes using bioinformatics approaches. Text mining was used to screen 40,225 article abstracts from diabetes literature. These studies highlighted 5939 diabetes-related genes spread across 22 human chromosomes, with 112 genes mentioned in more than 50 studies. Three datasets (44 patients and 57 controls) were subjected to gene expression analysis, the analysis revealed 135 significant DEGs. Text mining and gene expression analysis results are used as attribute values for machine learning (ML) analysis. The decision tree, extra-tree repressors and random forest algorithms were utilized in ML analysis to identify unique markers that could be used as diabetes diagnosis tools. These algorithms produced prediction models with accuracy ranging from 0.6364 to 0.88. There are 39 biomarkers that could distinguish diabetic and non-diabetic patients.

In [11], three datasets are analyzed to determine the causes of beta-cell dysfunction and type 2 diabetes (T2D) deficiency. In order to distinguish between healthy cells and cells with T2D, attention and focus are given to the levels of expression of essential genes. To obtain the influencing factors, the mutual information and correlation coefficients are calculated. Five machine learning classifiers are used: Bayesian network, SVM, RF, LR1,and neural network (NN), The highest accuracy (ACC) obtained for random forest classifier is 0.907.

In [12], six long non-coding ribonucleic acid (lncRNA) expression for T2D is used with four classification machine learning models, including nearest neighbors (KNN), SVM, LR1, and artificial neural networks (ANN) to diagnose T2D. These algorithms are compared with

each other in terms of diagnostic accuracy. The best area under the curve had SVM and logistic regression among the classification methods with an average AUC of 0.95.

In [13], two datasets (GSE38642) and (GSE13760) of T2D gene expression data are used from the Gene Expression Omnibus (GEO) database, the feature selection methods used are Fisher Score and Chi-square. By these methods, a subset of genes was obtained, numbering between (1800 – 2700). Logistic regression and SVM classifiers machine learning models are applied on the subset of genes for predicting genes that cause T2D, the highest accuracy is 0.902 for the logistic regression model when it is used on the fisher score method, when using the same model on chi-square accuracy is 0.88, while SVM classification did not produce satisfactory results.

In [14], the RNA-Seq dataset (GSE164416), which contains T2D samples and healthy samples, is used to discover biomarkers that directly affect the diagnosis of T2D and early prediction of its risk. The validity of the biomarkers is validated using (SVM) machine learning model. The identification of patients with T2D is evaluated with a sensitivity and specificity of 100%.

In [15], the authors developed a brand-new deep learning (DL) model to forecast T2D. First, 224 single-cell gene regulatory networks (SCGRNs) from pancreas with T2D and healthy pancreas have been assembled into a dataset and made available to the general public. Then, the dataset is used to train each of the seven DL architectures—VGG16, VGG19, Xception, ResNet50, ResNet101, DenseNet121, and DenseNet169 and a test set is used to evaluate each architecture's predictions. The results showed that the VGG19 prediction results are

consistent, with this design attaining the highest accuracy of 0.98. Summary of the related works, with more details are listed in Table (1.1).

Table (1.1): A Summary of Related Works.

Author	Dataset type	Disease type	Preprocessing/ feature selection method	Model	Evaluation
[2]	Gene Expression omnibus GEO GSE9006	T1D	Log_2 Normalization	MARS, AdaBoost, RF, and XGB-DART	Acc: 0.100
[9]	Gene expression omnibus GEO (GSE55098)	T1D	Log_2	SVM and LASSO	AUC: 0.918
[10]	Three gene expression datasets	Diabetes	Chi^2 RF	DT, RF	Acc : 0.88
[11]	Three expression dataset of β-cell	T2D	Correlation coefficients and MI	Bayesian N. SVM, NN LR1, and RF	Acc(RF):0.907
[12]	Lnc RNA expression and demographic data	T2D	Normalization Missing value	KNN, SVM LR1, and ANN	Acc(SVM): 0.95
[13]	RNA-seq expression data (GSE38642) and (GSE13760)	T2D	Fisher score and Chi^2	SVM and LR1	Acc (LR): 0.902
[14]	RNA-seq expression data (GSE164416)	T2D	Log_2	SVM	Acc: 0.100
[15]	Single-cell data consist of genes and cells	T2D	Optimization	Deep learning	AUC: 0.861 Acc: 0.861

1.5. Thesis Aim and Objectives

This thesis aimed to identify genes that may affect T1D and provide classification and regression models that can help the medical sector predict T1D patients, prevent the progression of the disease, and support health. To achieve these aims, first, determine a subset that represents the most informative genes from gene expression datasets used for the

classification and regression tasks. Then implementing classification and regression models based on the selected genes.

1.6. Challenges of the Problem Statement

- Dimensionality's curse: There are an enormous number of genes in the gene expression data (containing thousands of genes). Not all genes are helpful; some genes are redundant and useless data in the dataset, which is a regular occurrence. Thus, it is challenging to work with this huge number of genes.
- Applying machine learning models: This provides another challenge since it must be done with the least amount of mistakes and a maximum level of accuracy.
- The replications in T1D dataset: the dataset has number of replicates, where each replicate contain a set of different samples. The manipulation of these replicates in the dataset present a difficult.

1.7. Thesis Outline

After chapter one, which presents a general introduction, the rest of the thesis is organized as follows:

- Chapter Two introduces a theoretical background of the extensive description of the main fundamental biological concepts, preprocessing technique, methods used to reduce the data dimensions, machine learning algorithms, and metrics of evaluating the models which are used later in this thesis.
- Chapter Three shows the proposed system. It explains the preprocessing phase, then the ranking genes phase for dimension reduction, the followed phases identify the important features (genes)

in the T1D dataset. Then, the models of classification and regression are created.

- Chapter Four illustrates and discusses the implementation of the proposed system on the T1D datasets and the experimental results obtained after implementing the proposed model.
- Chapter Five describe the conclusions obtained based on the thesis outcomes and highlights possible directions for future works.

Chapter Two

Theoretical Background

Chapter Two

Theoretical Background

2.1. Introduction

This chapter describes type 1 diabetes T1D disease, then explanation for biological concepts which include, gene expression and the microarray technology used to represent data and demonstrates the used dataset. Then a literary background to the concepts of machine learning, which includes brief: data pre-processing, feature selection methods, machine learning models.

2.2. Type 1 Diabetes (T1D) Disease

Diabetes is regarded as one of the most common chronic diseases in the world, one of the diabetes types is the type 1 diabetes (T1D), and T1D is high blood glucose. Despite the severity of this disease, type 1 diabetes can influence individuals and their condition progresses to an advanced stage without them realizing it, making the disease hard to manage [1]. T1D is the result of the destruction of β-cells in the pancreas, leading to a lack of insulin such that the infected body requires daily injections of insulin to keep blood glucose under control [2].

T1D is primarily common in children and teenagers and is thought to be the primary form emerging from the interplay of autoimmune. This kind has higher mortality and health care expenses than type 2. Its precise molecular mechanism is currently unknown and ambiguous [3]. Hyperglycemia is one of the most important features of diabetes, caused by a defect in insulin secretion or action, or both. One of the complications of hyperglycemia is an imbalance of body functions and failure of some organs, and the disease becomes chronic [1]. Signs of high blood glucose include constant thirst, excessive urination, and excessive hunger [4].

2.3. Biological Concepts

In this section, explains a biological concepts that are necessary to produce the used data, including the gene expression and the microarray technology.

2.3.1. Gene Expression

The method for producing the necessary proteins through gene expression defines the physical components of living entities. Transcription and translation are the first two steps in gene expression. Enzymes are used to transmit information from DNA to RNA, and the process results in the synthesis of proteins and other biological compounds. DNA microarray is one method that may be used to measure gene expression from DNA or RNA [5]. The main genetic building blocks of living things are genes. The genetic information needed to encode certain RNA and cellular proteins is found in genes, which are parts of DNA. The basic tenet of molecular biology holds that proteins are created from DNA in a procedure that involves two key phases as illustrated in Figure (2.1).

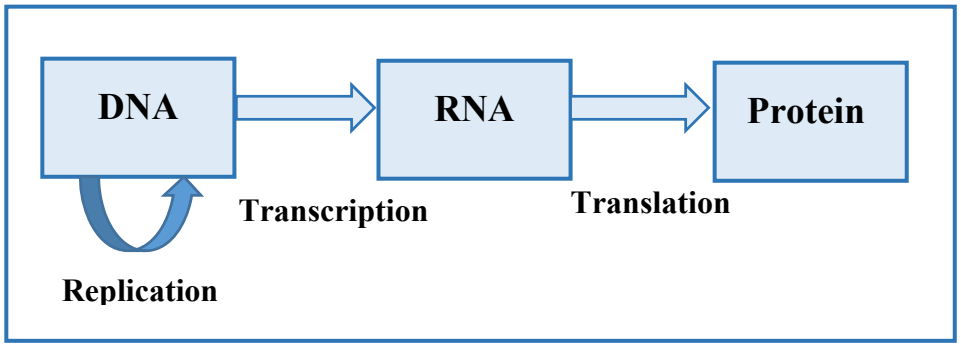

Figure (2.1): Dogma Central to Molecular Biology.

A gene in DNA is expressed in the first phase, transcription, by converting its encoded information into messenger ribonucleic acid (mRNA). When

Ribosomes have decoded the information from the mRNA, this process is known as translation and results in the production of proteins. Gene expression, which entails these two stages, enables a gene to be expressed as a protein [6].

2.3.2. Microarray Technology

Long used in global gene expression research, microarray technology (also known as DNA chip) enables the concurrent examination of hundreds or thousands of genes in a sample. It is distinguished by limited sample size and a high number of features produced by an incomplete rank and non-square matrix, which can produce many solutions in classifiers [7]. The most recent development in experimental molecular biology, the microarray technology, enables biomedical researchers to concurrently evaluate the expression levels of thousands of genes [8]. For each gene, microarray technology typically saves hundreds of expression data points. A matrix sometimes referred to as a gene expression matrix, can be used to represent the processed data from the microarray image file, containing rows expressing genes and columns expressing specific conditions. Structure of the gene expression matrix is shown Figure (2.2) [9].

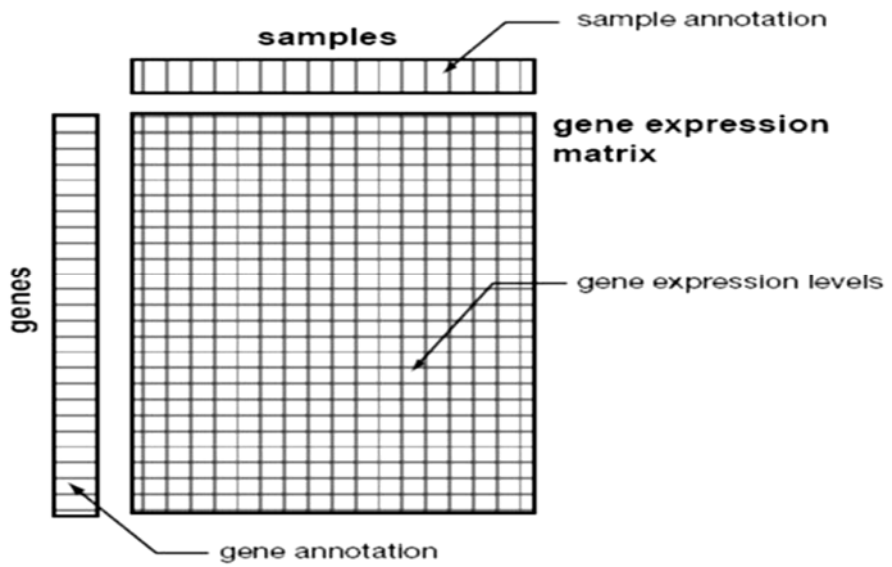

Figure (2.2): Gene Expression Matrix Structure [9]**.**

2.4. Dataset

The gene expression dataset used can be accessed Gene Expression Omnibus (GEO) database (http:// www. ncbi. nlm. nih. gov/ geo/). Gene expression data for type 1 diabetes T1D were used in the proposed system. Two datasets are used with GEO accession (GSE35725) and (GSE52724). The dataset has the code (GSE35725) published on 01 September 2012 and the dataset has the code (GSE52724) published on 13 May 2014 by the National Center for Bioinformatics Information (NCBI) [10], T1D datasets are analyzed using microarray technology to monitor the level of gene expression.

2.5. Data Preprocessing

Preparing data for the ML model is known as data preprocessing. It is the first and most crucial step in creating an ML model, and it has a significant impact on both the model's accuracy and effectiveness [11].

Reduce data size, identify relationships between data, normalize data, eliminate outliers, and extract characteristics from data are all goals of data preprocessing. There are several procedures involved, including data cleansing, integration, transformation, and reduction [12].

2.5.1. Data Cleaning

Data cleaning is the beginning of any machine learning project and is considered the most critical process in data science. This is important in ensuring the dataset is free from erroneous data. Data cleaning can be done manually with data collection tools or automatically with a computer program. Data cleaning is considered an essential first step in machine learning models, and its frameworks are increasingly being developed. It involved removing noisy or irregular data and guaranteeing no unwanted or invalid data before passing it to the machine learning model [13].

Because impure data is unreliable, users who rely on it have to expend more time verifying its accuracy, further reducing speed and productivity. Adding more manual methodologies introduces more inaccuracies and inconsistencies as the number of changed records increases. Therefore, it is important to understand the implications of these inconsistencies as helps in understanding various practical implications such as a large enough amount of incomplete information does not affect ML models. so the same amount of effort spent cleaning or retrieving specific data or missing information does not change the quality of other machine learning models [12].

2.5.2. Data Integration

The act of merging data from several sources and giving the user a consistent representation of that data is known as data integration. Building data integration systems is essential to the latest real-world applications and features a variety of theoretically intriguing topics. Establishing the correlation between the data in your assets and the data in the overall model is one of the most important elements in designing your data integration framework. In principle, many global databases

are suitable for such a source database-specific data integration system [14].

2.5.3. Missing values

Missing values are typically attributed to human errors in data processing, machine failures due to device malfunctions, respondents' refusal to answer specific questions, abandonment of surveys, and merging of unrelated data. The problem of missing values leads to various problems like slow performance, data analysis problems and skewed results caused by differences between missing and complete values [15].

In a dataset, it is typical for certain columns to contain missing values. Missing values must be taken into account since they may cause a model's feature to be eliminated. Simple interpolation techniques can fill such concerns if a respectable quantity of values is missing. The most popular approach is to utilize the mean, median, or mode values of the model features [16].

2.5.4. Data Normalization:

Data normalization is considered one of the first steps in processing a dataset. It is applied before the data is used, such as to increase or decrease a range of values. Normalization is convenient and useful in dataset problems with classification and regression by converting feature values for a specific and small range [17].

Min-Max Normalization: This is a linear transformation technique used in procedures where it is significant to maintain the relationship between the original data set. In addition, it is considered to be one of the simple techniques suitable for a dataset within certain limits [18]. Min-Max normalization is done according to Equation (2.1)

$$f\ new = \frac{f - \min(f)}{\max(f) - \min(f)} \quad (2.1)$$

Where

f: is the feature (gene) value.

$f\ new$: is the normalized f.

$\min(f)$: is the minimum amount for a feature f.

$\max(f)$: is the maximum amount for a feature f.

2.5.5. Data augmentation

Data augmentation is a series of procedures that grow and enlarge data while maintaining its label. It is considered a novel data creation approach that can create 2N new samples from N samples and has diverse data orientations. For researchers, data augmentation addresses two problems: it increases the amount of data from a little amount of data and reduces over fitting. Many data augmentation techniques are utilized to make the most of the restricted amount of data [16].

2.6. Student's t-test

The Student's t-test, also known as the t-test, is a parametric statistical test used to compare the means of two data sets. It can be used when the samples meet the conditions of normality, equal variance, and independence [19]. For a one-sample test, the value of the t-statistic is simply the distance between the sample mean and the mean of the distribution μ in units of the standard error. The calculated t-score is compared to the t-distribution to determine statistical significance. The tails of the t-distribution identify the rejection region (i.e., rejection of the null hypothesis of no difference) and are selected from tabulated t-scores [20]. The formula of the test as in the Equation (2.2):

$$t = \frac{\tilde{x} - \mu_0}{s/\sqrt{n}} \quad (2.2)$$

Where

$\tilde{x}$: is the mean of the sample.

μ_0: is the mean of the expected population.

n: is the sample size.

s^2: is the variance of the estimated population, Equation (2.3) defined as [21]:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(xi - \tilde{x})^2 \quad (2.3)$$

2.7. Feature Selection Methods

A dataset is a group of data elements that can be points, patterns, events, cases, samples, examples, or characteristics. Because of this, these data items are usually characterized by a variety of traits that provide the object's essential qualities, such as the object's mass, the time of occurrence, etc. A feature (gene) might be a distinct attribute that can be assessed or it can be a part of the phenomenon [16].

The "curse of dimensionality problem," which prevents the dataset from providing useful information, is the main flaw with microarray data [22]. Therefore, choosing a particular group of genes from data collection is difficult in the discipline of bioinformatics. ML models require a long time to run since microarrays of this sort include a small number of samples and a big number of genes [2]. The process of selecting a subset of the most significant features from a dataset to characterize the target variable is

known as feature selection. Feature selection approaches are categories in three categories, wrapper, filter, and embedded [23].

The two main sources of feature selection's popularity in microarray data analysis are biology and computer science/statistics. Finding significant genes that are impacted or even responsible for disease may be accomplished by choosing genes from microarray data that allow for good differentiation between healthy and afflicted patients. This is a crucial step in comprehending a basic biological procedure. Nevertheless, because of sample sizes for each feature, computational and statistical experts are primarily focused on finding solutions to the issues of over fitting, redundancy, and data noise. Trait selection has therefore been used to decrease the number of features (genes) by eliminating illogical features like categorical data in gene expression data and considerably improving the categorization accuracy [24].

2.7.1. Mutual Information (MI) Method

Calculating the amount of information for one variable in respect to other related variables is the process of determining mutual information. The idea of mutual information may be used for feature selection since it provides a technique for determining how well a subset of features covers the output vector [25].

Assume X and Y stand for a gene and a class label, respectively. First, the entropy H(X) is used to calculate the entropy of (X), and the entropy H(Y) is used to calculate the entropy of (Y). The equations (2.4) and (2.5) represent H(X) and H(Y) respectively [3].

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \qquad (2.4)$$

$$H(Y) = -\sum_{y \in Y} p(y) \log p(y) \qquad (2.5)$$

The probability of (X) and (Y) are represented, respectively, by P(X) and P(Y). Second, the joint entropy H (X, Y) calculation measures the degree of uncertainty between (X) and (Y). It is calculated using Equation (2.6):

$$H(X;Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y) log\, p(x,y) \qquad (2.6)$$

P (X, Y) denotes the joint probability that (X, Y) values will occur simultaneously, and the conditional entropy is defined in Equation (2.7) [26]:

$$H(X|Y) = -\sum_{x\in X}\sum_{y\in Y} p(x|y) log\, p(x|y) \qquad (2.7)$$

Lastly, it calculates the mutual information MI (X, Y) between two random variables (X, Y) according to the equation (2.8).

$$I(X;Y) = -\sum_{x\in X}\sum_{y\in Y} p(x|y) \log \frac{p(x,y)}{p(x)\, p(y)} \qquad (2.8)$$

From Equation (2.8), the relation between MI and entropy may be derived in Equation (2.9):

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(X|Y) \qquad (2.9)$$

In light of this, MI may be used to select a subset of characteristics that enhance the relevant data and X, Y that minimize H (X, Y). In the end, this aids in understanding how the original data's underlying relationships are represented by machine learning models [27]. The use of mutual information to select features is done by selecting a subset of features (genes) from the data set that includes all the features [28].

2.7.2. Analysis of Variance (ANOVA) Method

ANOVA is known as the F statistic; It is used for reducing the size of datasets containing a huge number of features in order to acquire a new dataset that can be defined using the fewest number of variables [29]. ANOVA is generally used in a data set with many classes to check if the means have a big contrast between them [22]:

(a) Equations (2.10) and (2.11) are used to compute the variation between the group:

$$Between\ sum\ of\ squares\ (BSS) = \sum ni(\bar{x}i - \bar{x})^2 \quad (2.10)$$

$$Between\ mean\ squares\ (BMS) = BSS/df \quad (2.11)$$

(b) Equations (2.12) and (2.13) are used to compute the variation within the groups:

$$Within\ sum\ of\ squares\ (WSS) = (xij - \bar{x}i)\sigma^2 \quad (2.12)$$

$$Within\ mean\ squares\ (WMS) = WSS/d\ fw \quad (2.13)$$

Where:

df: is the degree of freedom.

$df_w = (N - k)$

σ: is the standard deviation.

N: is the number of samples.

K: is the number of groups.

n_k: is the number of samples in group k.

(c) F-test statistic is calculated as in Equation (2.14):

$$F = BMS/WMS \quad (2.14)$$

2.7.3. Chi-square (χ^2) Method

One of the statistical analysis techniques that are non-parametric. The Chi-square test is a useful tool for comparing experimentally obtained findings with those expected theoretically based on a hypothesis.The Chi-square formula calculates the actual discrepancy between the observed and expected frequencies. In sampling studies, where it is constantly important to look at the gap between theory and reality, it is obvious that the relevance of such a measure would be significant [30]. The value of the Chi-square is calculated using the Equation (2.15)

$$x^2 = \sum_{i=1}^{k} \frac{(Oi-Ei)^2}{Ei} \qquad (2.15)$$

Where

Oi: is the observed value in the ith class.

Ei: is the expected value.

k: is the number of classes [23].

2.7.4. Principle Component Analysis (PCA) Method

PCA is regarded as a linear transformation method that performs an orthogonal translation of bound input data into unbound features. It is a frequently used technique for feature extraction, data compression, size reduction [32]. PCA creates base components, which are input features converted from related to unrelated because the output data contains fewer unrelated features [33]. PCA method applied using applying the equations of the calculated the covariance matrix, this matrix represents the variance of the data and also shows the covariance between the variables. Provides an empirical description of the data [34]. The covariance matrix is formulated using the formula in the Equation (2.16):

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{(n-1)} \quad (2.16)$$

The result would be a square matrix of d ×d dimensions.

Diagonal elements store the variance of the principal components. When the off-diagonal entries of the covariance matrix are positive, x and y grow together. Otherwise, x decreases as y increases.

Using the covariance matrix, eigenvalues and eigenvectors are computed for each predictor. The eigenvalues are the solutions to the characteristic Equation (2.17):

$$|A - \lambda I| = 0 \quad (2.17)$$

Where:

A: is a matrix.

I: is the identity matrix.

λ: is a scalar value.

Once the eigenvalues are found, the eigenvectors can be found by solving the equation (2.18) for each eigenvalue:

$$Av = \lambda v \quad (2.18)$$

Where:

v: is a non-zero vector.

The following step is to organize the eigenvalues in descending order as shown in equation (2.19), where the feature with the largest eigenvalue is the main component of the data set.

$$\lambda 1 \geqslant \lambda 2 \geqslant \cdots \lambda n \quad (2.19)$$

A new feature vector is formed that contains all the basic components of the dataset. The size of the dataset is diminished to include only those

features that capture the greatest amount of knowledge and are not associated with each other.

2.8. Machine Learning Models

One of the branches of artificial intelligence that deals with many issues is machine learning. To address these challenges, several algorithms have been created [33]. In the field of medicine, machine learning models are becoming more useful, with the development of machine learning, complicated and high-capacity biomedical data can now be computationally analyzed [35]. Models with excellent predictive capabilities are used, which are created expressly for issues with big volumes of data and noise. ML models are utilized for a variety of issues, including classification and regression issues.

Using regression analysis, two ideas are supported. First and foremost, regression analysis is frequently used for forecasting and prediction, two tasks for which machine learning and its application are relatively comparable. Second, regression analysis may be utilized in specific circumstances to establish the causal links between the independent and dependent variables [36].

2.8.1. Random Forest (RF) Method

A supervised method with several prediction trees is called random forest [37]. Random forests are one of the useful methods used recently in biological research because to their adaptability and simplicity. The outcomes demonstrate the relative importance of each characteristic when there are several variables since each one serves a distinct purpose in the prediction process [38].

Random forest combines the results from several trees, adding additional layer for bagging, as opposed to determining a solution. By bootstrapping dataset samples, bagging produces n predictors utilizing independent sequential trees. An estimation or classification problem is handled by averaging the n predictors. By averaging the forecasts for each tree, decision trees that make up a random forest structure are created [39][40]. While the average of numerous different tree predictions is used in regression, the majority of trees suggest answers to classification problems [41]. Here is a brief description of the random forest method to construct number of trees, select bootstrap samples from the original dataset. For each bootstrap sample, build a classification or regression tree. At each of the tree's nodes a predictor variables (or a subset of features) are randomly chosen from all of the predictor variables and placed at each node of the tree (random subspace). The binary split is carried out on the node with the predictor variable that offers the best split. The previous phase is carried out by the next node, which then randomly chooses another set of m variables from among all predictor variables. If a new dataset has to be categorized, use the sub trees overall majority vote.

2.8.2. Support Vector Machine (SVM) Method

Support vector machines are supervised learning methods for data analysis and pattern identification. The three major uses of SVMs are regression analysis, classification, and novelty detection. Using a two-class learning task that splits subsequent observations between the two classes on each side of a hyperplane shown in Figure (2.3), an SVM training technique builds a model or classification function from a collection of training data, SVM is a non-probabilistic binary linear classifier. The observations are represented as points in space by an SVM model, and as

a result, they are split into various partitions based on the maximum margin to the closest observation data point of each class. Following that, further observations are anticipated to fall on one of the two sides of the partition, corresponding to a class. When a training dataset of *n* points of the form $(p_1, q_1)\ldots (p_n, q_n)$, where the q_i are either 1 or −1, each indicating the class to which the point p_i belongs and each p_i is a p-dimensional vector. The maximum margin hyperplane that divides the group of points p_i for which q_i =1 from the group of points for which q_i = -1, which is defined in Equation (2.20), the distance between the hyperplane and the nearest point p_i from either group is maximized. Any hyperplane can be written as the set of points p_i satisfying

$$w * p - b = 0 \qquad (2.20)$$

Where w is the normal vector to the hyperplane.

The parameter $\frac{b}{||w||}$ determines the offset of the hyperplane from the origin along the normal vector w [39].

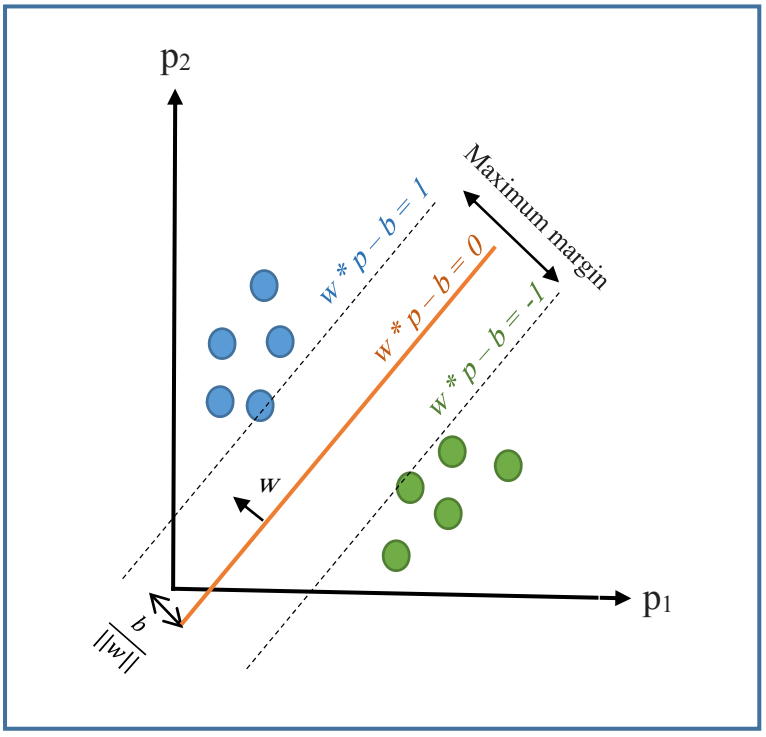

Figure (2.3): The SVM Method Find The Hyperplane.

SVM is used in applications involving binary classification to minimize the error by locating an acceptable hyperplane that divides the two categories. The kernel function may be calculated by choosing data from a lot of different dimensions. This step could improve in separating the data linearly because most problems do not have a linear relationship [35]. In the realm of medicine, this method is frequently used to identify diseases. The SVM method attempts margin maximization, which implies that it seeks to establish a maximum distance between various classes since the dataset may contain several such hyperplanes [34]. The fundamental model is a classifier that identifies the largest margin in the region. SVM is used to categorize two types of data.

Regression problems are resolved with support vector regression (SVR) utilizing a support vector machine (SVM). The strip is used by the SVR to fit the data, there is a setting that controls and adjusts the width of this strip. SVR will find an appropriate line (or hyperplane in higher dimensions) to fit the data depending on how much model error is tolerable,

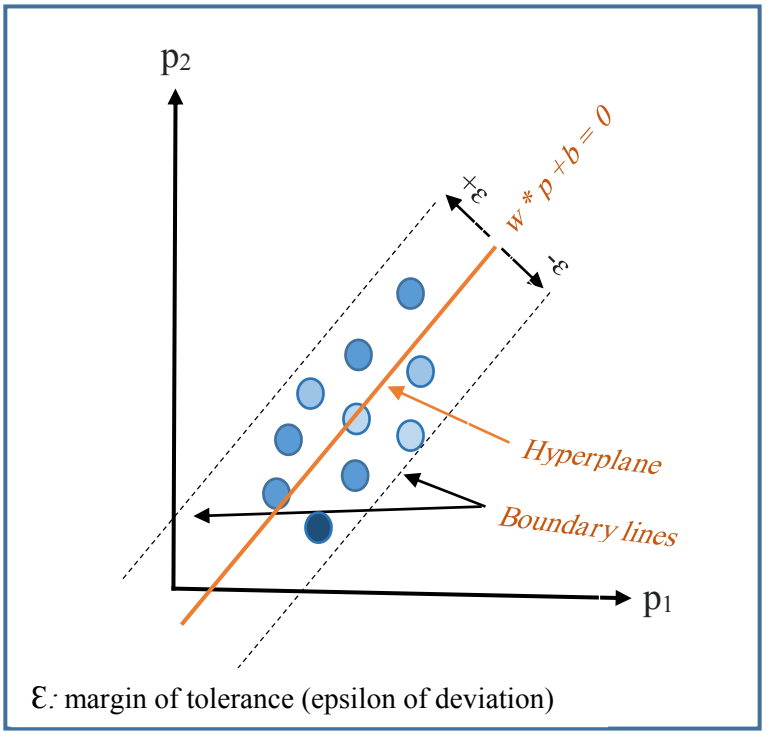

Figure (2.4): The SVR Method Find The Hyperplane.

according to the Equation (2.21). Reduce the discrepancy between the function's anticipated cost and the actual output for a certain input [42].

$$y = w * p + b \quad (2.21)$$

2.8.3. Linear Regression (LR) Model

The most often used prediction model for finding out how variables relate to one another is linear regression. There are two forms of linear regression: simple linear regression and multiple linear regression. [43] Equation (2.22) describe the linear regression model:

$$y = x\beta + \varepsilon \quad (2.22)$$

Where:

y : is the dependent (response) variable.

x: is the independent (predictor) variable.

β: is the estimated slope.

ε: is the estimated intercept.

A statistical technique called multivariate linear regression (MLR) makes use of a number of explanatory factors to predict the results of a response variable. The goal of (MLR) is to model the linear connection that will be investigated between the independent variables x and the dependent variable y [44]. Equation (2.23) and Figure (2.5) show the fundamental model for MLR:

$$y = \sum xi\ \beta i + \varepsilon \quad (2.23)$$

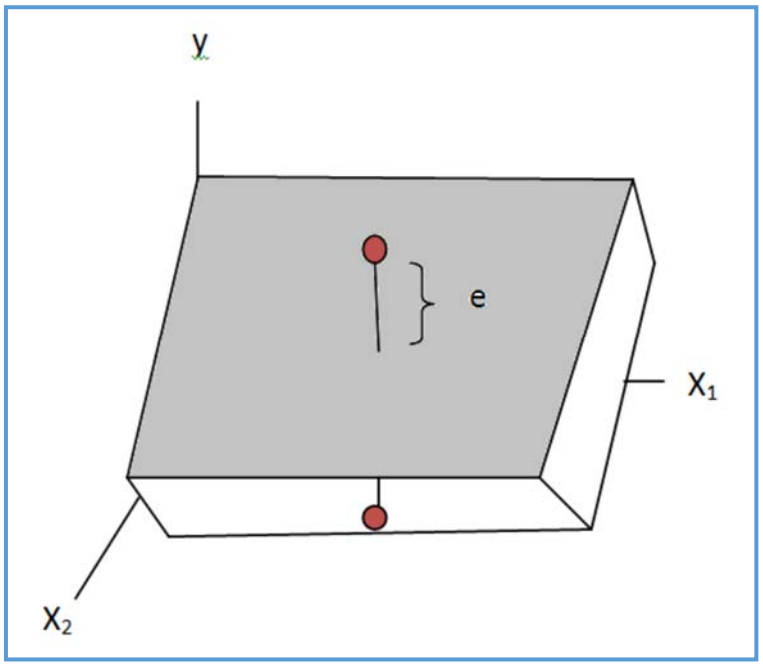

Figure (2.5): Multiple Linear Regression [43]**.**

2.9. Evaluation Metrics

In order to train a machine learning model to its full potential, evaluation metrics are crucial. As such, choosing appropriate assessment criteria is a crucial step in differentiating and achieving the best model [45].

2.9.1. Accuracy (Acc): The accuracy metric calculates the proportion of accurate predictions to all occurrences that are analyzed [45]. Equation (2.24) can be used to determine accuracy:

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \qquad (2.24)$$

Where:

TN: true negative.

TP: true positive.

FP: false negative.

FN: false negative.

2.9.2. Precision: is employed to calculate the percentage of accurately predicted positive patterns inside a positive class [45]. Equation (2.25) contains the following formula that may be used to calculate precision:

$$Precision = \frac{TP}{TP+FP} \qquad (2.25)$$

2.9.3. Recall: is the proportion of Positive samples to all Positive samples that have been correctly classified as Positive. How effectively the model can distinguish Positive samples is measured by recall. When the recall is higher, there are more positive samples discovered [23]. The recall is calculated using Equation (2.26).

$$Recall = \frac{TP}{TP+FN} \qquad (2.26)$$

2.9.4. Root Mean Square Error (RMSE): is a metric that is frequently used to calculate the prediction scores or standard deviation of the residuals. Equation (2.27) illustrates the departure of predictions from the regression line [5].

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(xi-\tilde{x}i)^2}{n}} \qquad (2.27)$$

Chapter Three

The Proposed System

Chapter Three

The Proposed System

3.1. Introduction

In this chapter, the main objectives of this thesis are explained. This includes a proposed system for classifying and predicting gene expression data on type 1 diabetes using machine learning methods. The block diagram of the proposed system is clarified first and then discussed in the pre-processing to prepare the data for the following stages, the total number of genes is reduced by the ranking of genes, then a subset of the total genes is selected to be the stage of feature selection, then the classification models and regression models, finally the evaluation methods for the proposed system.

3.2. The Proposed System Design

The design of the proposed system consists of two main parts, the first part includes five main stages (pre-processing, ranking, feature selection, classification models, and evaluation) to achieve the goal of this thesis. The stages are clarified here, the first stage is the preprocessing of the data, which includes data cleaning, integration, handling missing values , and normalization. The second stage is gene ranking using the student's t-test to select a subset of the original genes. Feature selection is the third stage for reducing the original genes to the genes most related to the disease and these methods include Mutual Information, Chi-Square, Analysis of Variance, and Principle Component Analysis. Two paths are used to apply the feature selection methods, single feature selection and multiple feature selection. The fourth stage is machine learning classification models, including Random Forests and Support Vector Machines. The fifth stage evaluates each model using the performance metrics.

The second part of the proposed system includes four stages (pre-processing, ranking, regression models, and evaluation), these parts are used for this thesis's goals. The stages are explained here, the first stage is the preprocessing of the data, which includes handling missing values and normalization. The second stage is the ranking of genes using the student's t-test. The third stage is the machine learning regression models, including (Linear Regression, Random Forest Regression, and Support Vector Regression). The fourth stage is the evaluation of the model performance. Figure (3.1) shows the general design of the proposed system.

3.3. Preprocessing Stage of the T1D Dataset

Preprocessing is a necessary step to prepare the data to fit machine learning models (classification models and regression models), raw data may contain missing values that must be handled, as well as transforming the data into a simple and efficient format, in the chapter four the details of the T1D dataset are mentioned and applied the preprocessing methods.

3.3.1. Data cleaning: The two datasets contain unwanted samples, and therefore specific samples are taken from both sets to be suitable for working, and the take depends on the stage of the T1D disease, a description of the dataset in the appendix A.

3.3.2. Integration: In this step, data is taken from several sources and concatenated in one dataset to work on it permanently. Samples for different stages of the disease are combined from two datasets that have the same features (genes). This merge aims to assist in the classification of the type 1 diabetes dataset. Figure (3.2) illustrates the concatenation of two datasets.

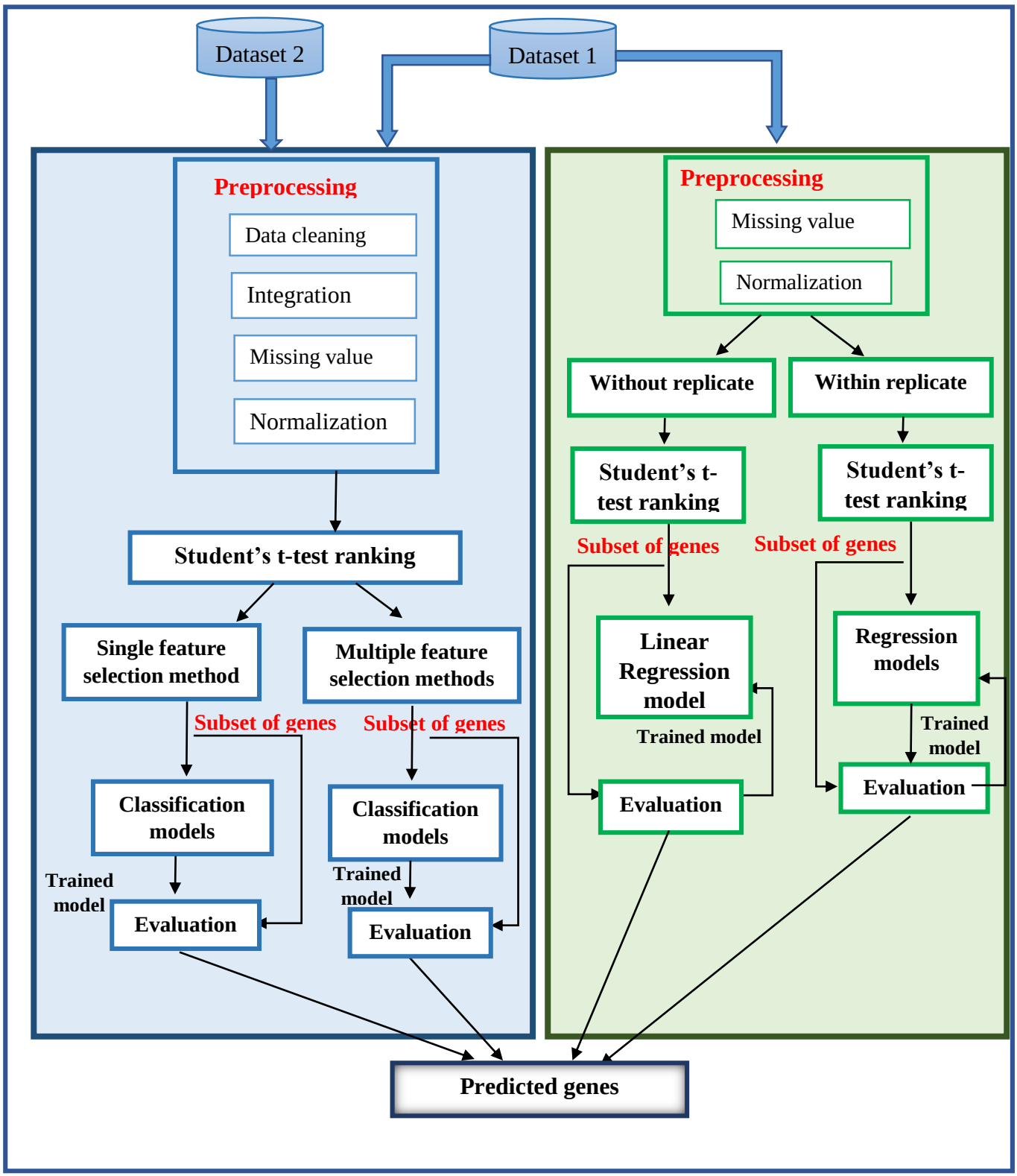

Figure (3.1): Block Diagram of the Proposed System.

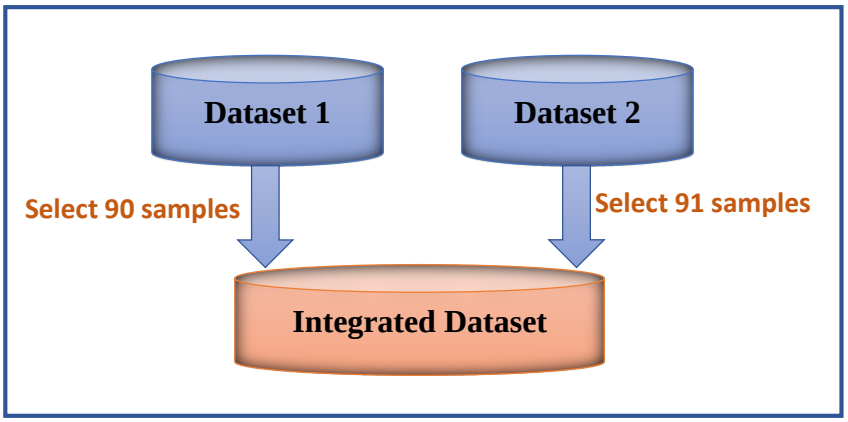

Figure (3.2): The Integration Step of the Data.

3.3.3. Handling missing values: the missing values were handled using the method of substitution by a new value obtained by calculating the mean of the gene column containing the missing value.

3.3.4. The normalization: the min-max normalization method is applied to all genes for normalizing the values to avoid high values that might affect the calculation of the results, according to the Equation (2.1) mentioned in the chapter two in the section (2.5.4). All the numeric values of the genes in the T1D dataset that serve as input to the machine learning models are normalized to be in the range zero and one.

3.3.5. Data augmentation: the number of samples in the data is little. Therefore, the augmentation technique used for increasing the number of data samples from N samples to 2N samples.

3.4. Ranking Stage

The T1D dataset is characterized by high dimensions, it contains a huge number of genes, and perhaps not all of them are related to the disease. For this reason, the ranking of the genes is performed to determine a certain number of genes that are used as an input to machine learning models, and to facilitate dealing with data, the student's t-test is applied to all genes using the Equation (2.2) in the chapter two in the section (2.6).

3.5. Feature Selection Stage

The high dimensions of the dataset necessitated that feature selection be used to reduce the dimensions, select the features most relevant to the disease, and obtain a satisfactory classification model performance. Two paths are used to apply the feature selection methods, single feature selection and multiple feature selection. Figure (3.3) illustrates the feature selection paths.

In the single feature selection path, feature selection methods are applied separately to obtain a subset of the data and serve as an input to the classification model. While in the multiple feature selection path, one method of feature selection is applied and a subset of data is obtained represent the most relevant features (genes) of T1D, it will be an input to another method of feature selection as illustrate in the Figure (3.4). The implementation is carried out in several different sequences and with a different number of feature selection methods, and the result of each sequence at the end serves as input to the classification models.

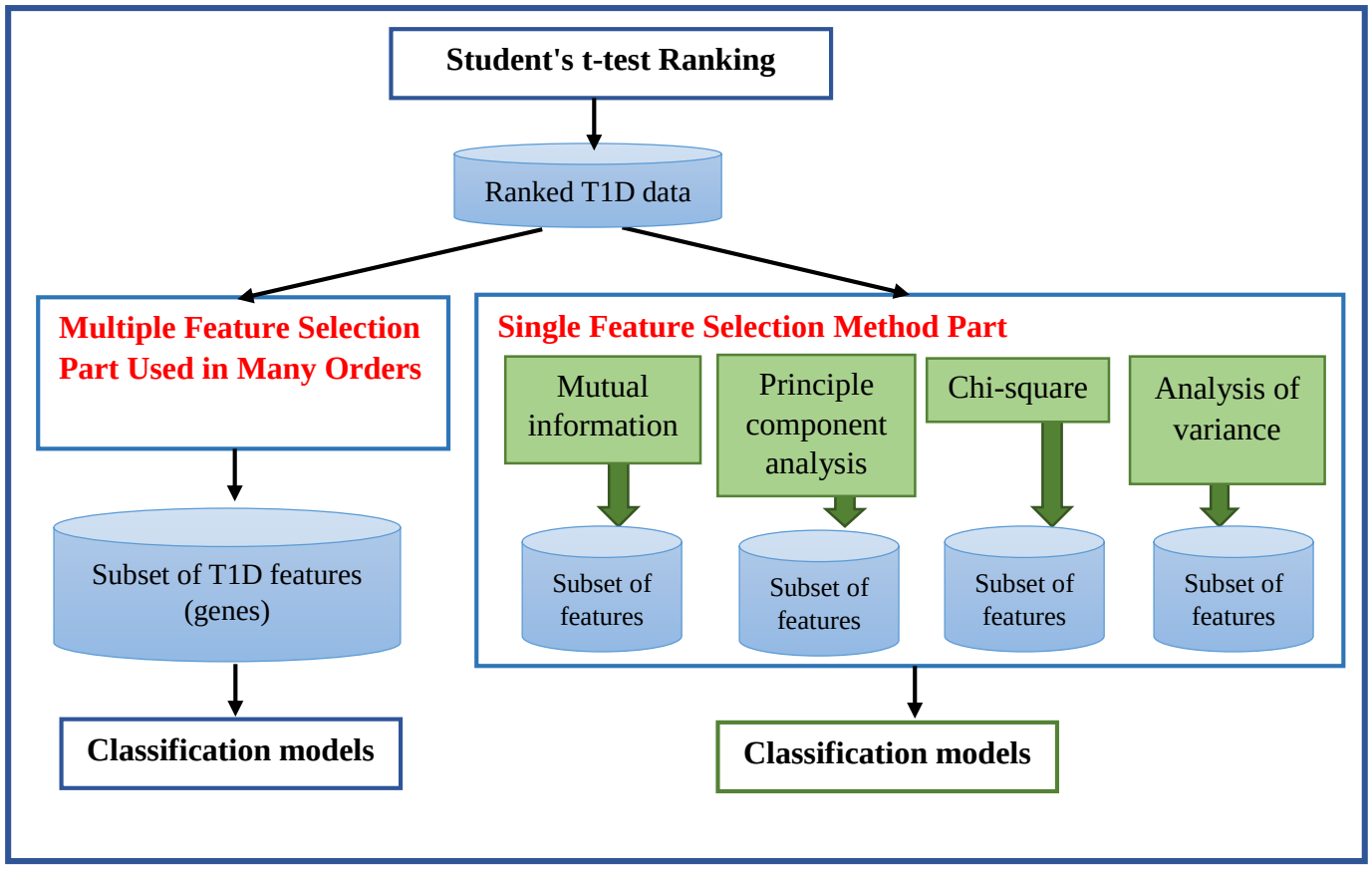

Figure (3.3): The Feature Selection Methods.

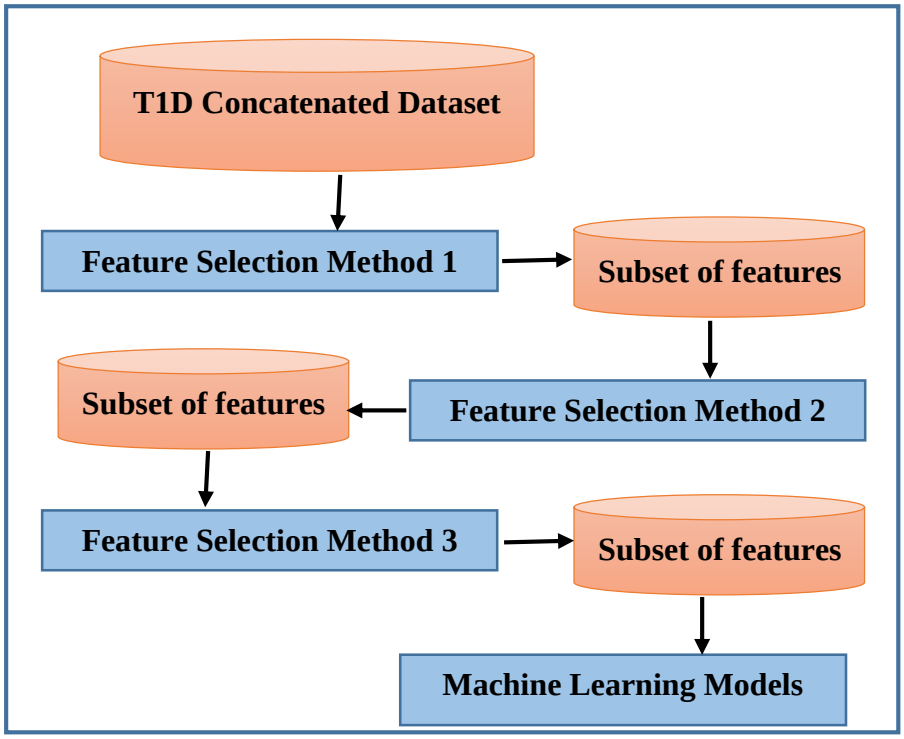

Figure (3.4): The Multiple Feature Selection Methods Paths.

3.5.1. Mutual Information Method

The mutual information method is one of the filtering methods, implemented to select the relevant and important features. Figure (3.5) shows the process of selecting features according to the method of mutual information. The mutual information between the features (genes) and the class label is calculated according to the Equation (2.8) mentioned in the chapter two in the section (2.7.1).

The gene that has a higher weight that obtained by mutual information method is selected and, conversely, the gene that has a lower weight is neglected, and thus the genes are arranged in descending order according to their weights resulting from the mutual information. The resulting subset is either an input to another feature selection method or a classification model. Algorithm (3-1) shows more details about mutual information method.

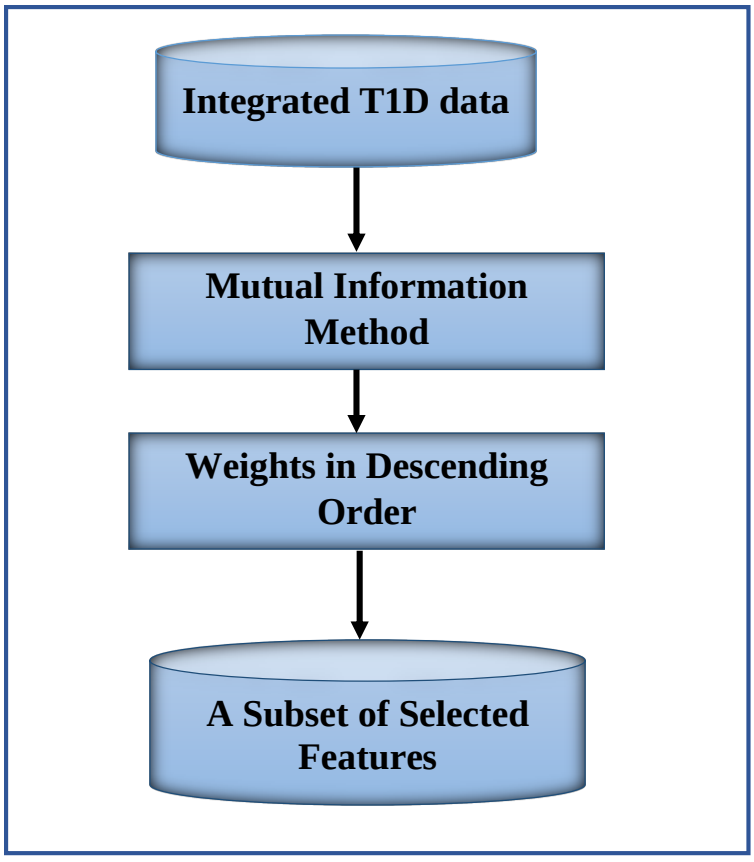

Figure (3.5): The Mutual Information Feature Selection Process.

Algorithm (3-1): Mutual Information Method
Input: *Two-dimensional array RankedD [S * F] where S is the number of samples and F is the number of features.*
Output: *Subset of Data (SD): the subset contains features have the highest of Mutual Information values.*
Begin
1. Set features_subset to NULL. // ***Initialize the feature_ subset***
2. for i = 1 to F
3. for j = 1 to C_label // ***C_label : is the number of classes***
4. Calculate the Mutual Information between feature F_i *and class label* C_label_i *based on the equations in section (2.8.2)*
5. end for j
6. end for i
7. Select the features F_i *with maximum MI (*F_i *,*C_label_i*)*
8. The selected features are sorted a the mutual information values in descending order
End

3.5.2. Analysis of Variance Method

The Analysis of Variance approach is used to choose the T1D dataset most pertinent characteristics. The p-value considered as the measure of the effectiveness of ANOVA, which is applied to each character and the set of genes chosen in accordance with the predetermined threshold serves as indicators of how well an ANOVA performed. To choose the subset of features from the original data, the genes are organized in descending order of p-value, and the feature with a value greater than the threshold is cancelled. A block diagram of the ANOVA method is shown in Figure (3.6). Equation (2.14) is used to apply ANOVA, which mentioned in section (2.7.2) in the chapter two, for more details, Algorithm (3-2) describes it.

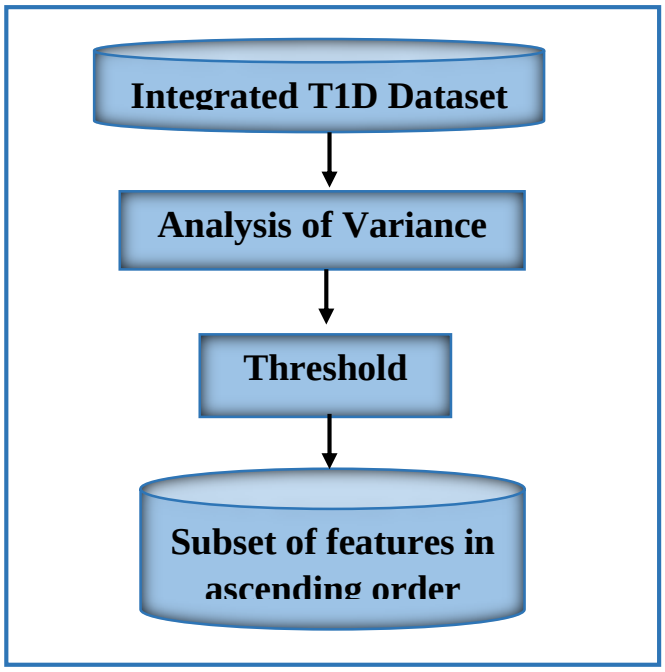

Figure (3.6): Block Diagram of Analysis of Variance Method.

Algorithm (3-2): Analysis of Variance Method
Input: *Two-dimensional array RankedD [S * F] where S is the number of samples and F is the number of features.*
Output: *Subset of Data (SD)*
Begin
1. for each feature $\mathbf{f_i}$ *do*
2. Calculate the BMS values by using the equation in sub-section (2.7.2)
3. Calculate the WMS values by using the equation in sub-section (2.7.2)
4. Calculate the F-value ($\boldsymbol{F_i}$ *= BMS / WMS)*
5. Find the p-value (pi) corresponding to each F-value
6. end for
7. For each feature set $\mathbf{f_i}$ *do*
8. if $\boldsymbol{p_i} < 0.05$ *then*
9. choose the feature, called f $\mathbf{c_i}$
10. Else
11. Cancel the feature
12. end if
13. end for
End

3.5.3. Chi-Square Method

This statistical test has been used to select the most responsive features to reduce the huge number of T1D datasets. Equation (2.15) mentioned in the section (2.7.3) in chapter two shows the chi-square formula; the higher chi-square value indicates that the feature is more dependent on the response. For this reason, the subset of features (genes) is selected based on a threshold predefined, for more details see the algorithm (3-3). The result of this test is passed as input to another feature selection method or as input to a classification model.

Algorithm (3-3): Chi-square Method
Input: *Two-dimensional array RankedD [S * F] where S is the number of samples and F is the number of features.*
Output: *Subset of Data (SD): the subset contains features have the highest of* Chi^2 *values.*
Begin
1. for i = 1 to S
2. for j = 1 to F
3. Compute the Chi^2 *values based on the equation in section (2.7.3)*
4. end for j
5. end for i
6. Select the features F_i *with maximum Chi (*F_i ,C_label$_i$*)*
8. The selected features are sorted according to the Chi^2 *values in descending order*
End

3.5.4. Principle Component Analysis Method

Principle component analysis method is used for reducing the high dimensionality of the dataset, whereby a new set of features is created while preserving as much as possible the original features. Principle components (PCs) are

formed when correlated variables are converted to linearly non-correlated variables. Calculation of the eigenvalues and eigenvectors of the covariance matrix produces the PCs. The dimensions of the dataset are reduced by eliminating the principle components with the lowest variance, considered weak components. And, high-variance components are arranged in descending order and saved as input to the method for selecting another feature or for the classification model. Figure (3.7) shows more about the PCA method.

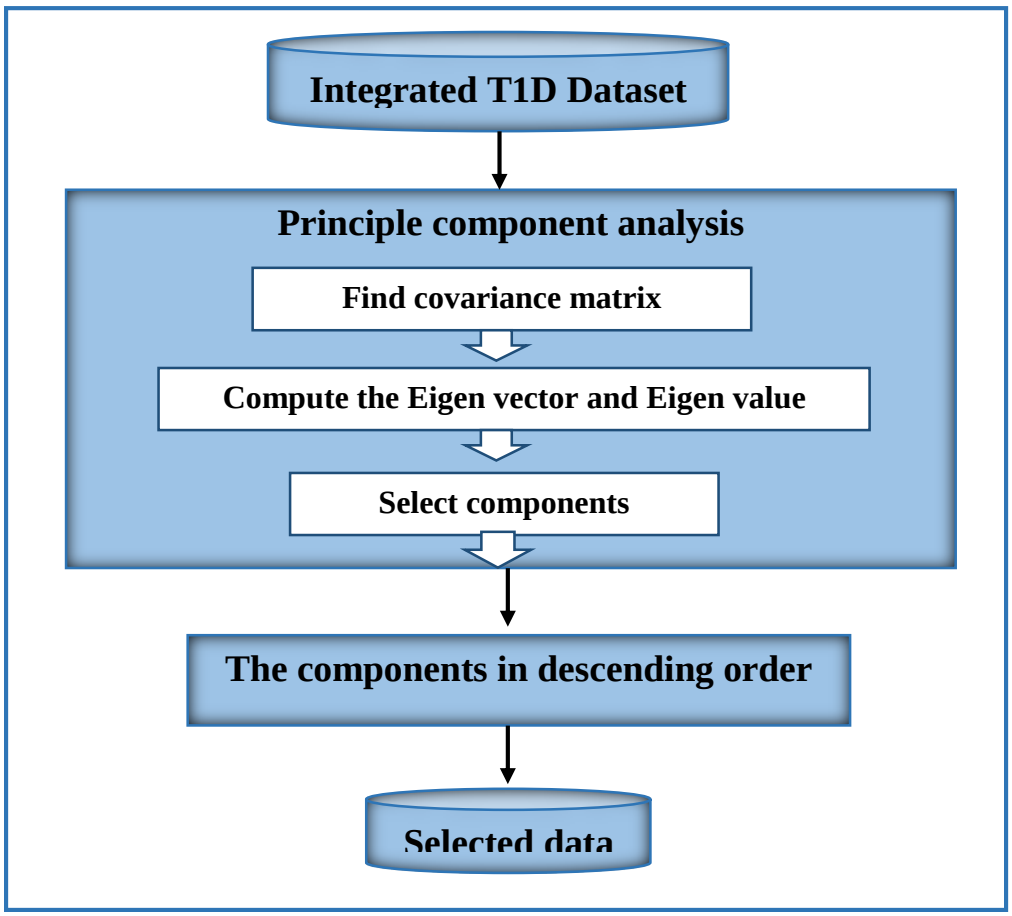

Figure (3.7): Block Diagram of Principle Component Analysis Method.

3.6. Machine Learning Models

In this work, the machine learning models are used for two purposes, classification and regression, to find the most relevant genes for the T1D disease.

3.6.1. Classification Models

Two of the machine learning models are used to achieve the classification of the T1D dataset, including a random forest model and a support vector machine model, Figure (3.8) illustrates the block diagram for machine learning classification models.

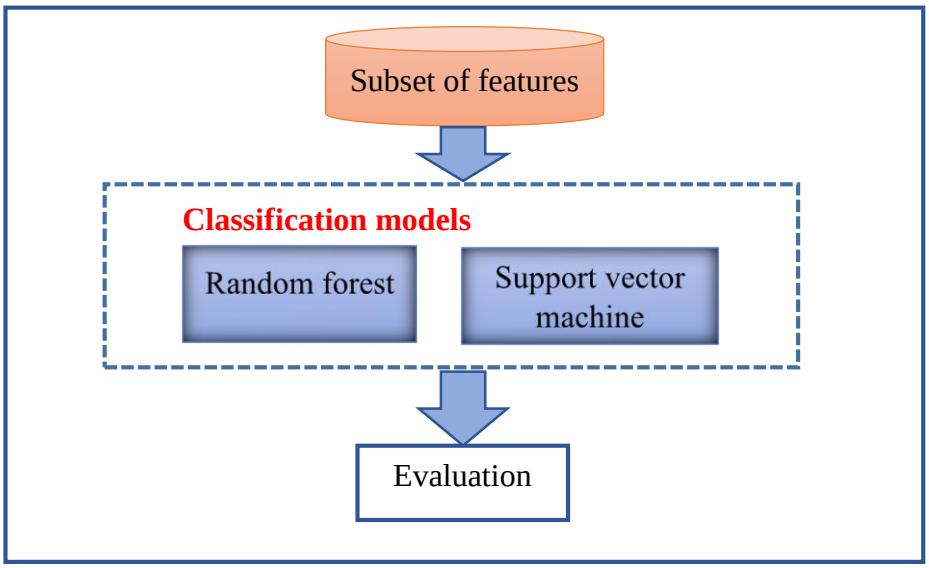

Figure (3.8): Block Diagram of ML Classification Models.

3.6.1.1. Random Forest Model (RF)

Before applying the random forest model, the data obtained from the feature selection stage is divided into two parts, the training set, and the test set. The separated data is distributed to be the largest part of the training data and the smallest part of the test data. The random forest model is implemented and trained using the training set of data, and then the model is tested using the test set. The RF classification model obtains the result of each tree based on the majority voting of class prediction. Figure (3.9) illustrate the random forest classification model and the algorithm (3-4) shows more details of random forest.

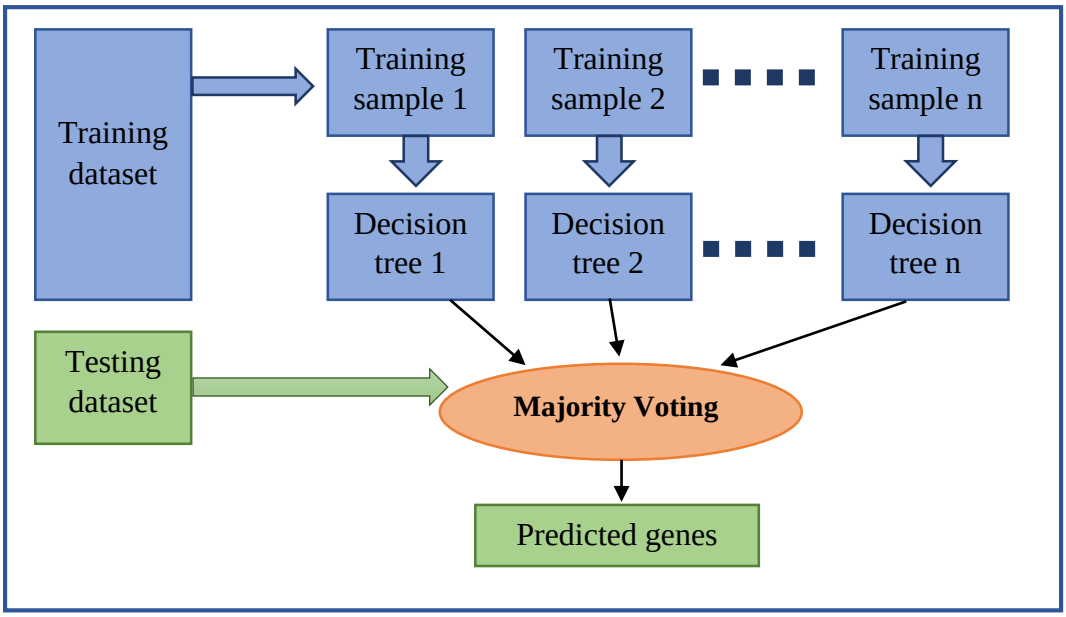

Figure (3.9): Block Diagram of Random Forest Classification Model.

Algorithm (3-4): Random Forest Algorithm
Input: *Two-dimensional array SD [S * F] where S is the number of samples and F is the number of selected features.*
Output: *Subset of Data (SD): the subset contains features have the highest of Random Forest values.*
Begin
1. Split SD into two set: training set TR, and testing set TS.
2. for i = 1 to S
Select random K data points from TR.
Build the decision tree associated with the selected data points (Subsets).
Choose the number N for decision trees that want to build.
3. Predict and store the outcome of each decision tree on TS.
4. Compute the total vote for individual class.
5. Declare the majority class as the final outcome class.
End

3.6.1.2. Support Vector Machine Model (SVM)

The SVM is used for classifying the T1D dataset according to the Equation (2.16) mentioned in the chapter two in the section (2.8.2), where the linear kernel is used in the implementation of the model. The reduced dataset obtained using the feature selection methods is divided into training set and testing set before applying the model. The SVM model is trained by the training set and then tested by the test set. The SVM linear model separate the training data points by finding the optimal hyper plane and the maximum margin between the data points.

3.6.2. Regression Models

Three regression models are used on T1D datasets, including linear regression model, random forest regression model, and support vector regression model as shown in Figure (3.10) for predicting the genes that are most relevant to the T1D dataset.

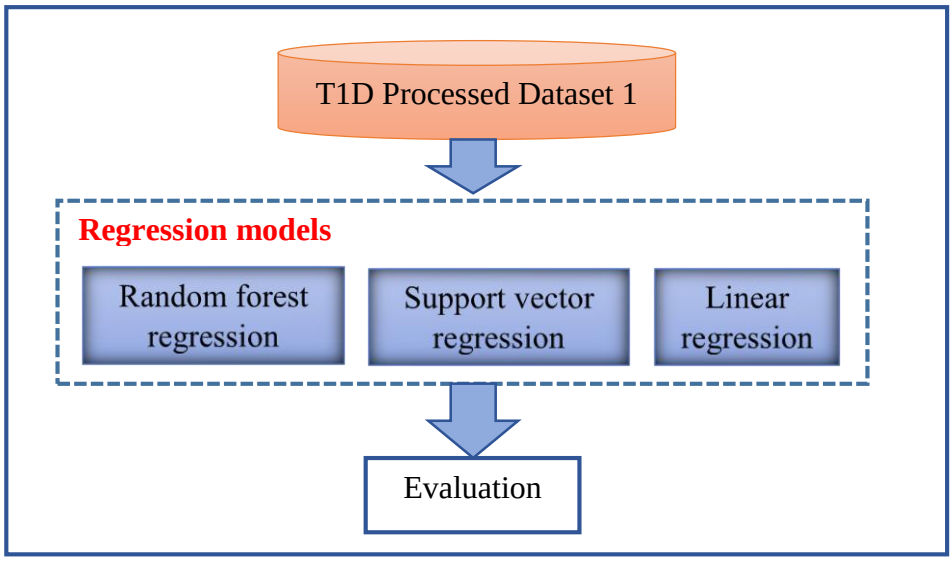

Figure (3.10): Machine Learning Regression Models.

3.7. Evaluation Models

In this stage, the performance of ML models is evaluated using the metrics for the classification and regression models. The classification models are evaluated by calculating each model's accuracy, recall, and precision using training data and test data as input to the model according to the Equations (2.24), (2.25), and (2.26) respectively mentioned in the chapter two in the section (2.9). The accuracy metric evaluates performance for ease of understanding and simplicity. The root means square error (RMSE) metric is used to assess the regression models by using the Equation (2.27) that mentioned in the chapter two in the section (2.9.4).

Chapter Four

Results and Discussions

Chapter Four

Results and Discussions

4.1. Introduction

The proposed system, illustrated in chapter three, has been implemented to cover the goals of this thesis The goals are to classify and predict the T1D dataset using machine learning models and to identify the genes that most influence disease progression through samples from different cases of the disease. First, the requirements of the research are mentioned in a summary of the T1D dataset. Then, the results for each stage are described and discussed in this chapter. To evaluate the results, performance evaluation metrics are used such as accuracy, recall, precision, and RMSE.

4.2. The Proposed System Requirement

The proposed system is implemented using a Python development environment, the version used is Python 3.9 with anaconda software with Jupiter package that support Python. The operating system is windows 10 environments, 64 bits. The processor manufacturer is Intel with Core i5_8250U running at 1.8 GHz. The capacity of RAM is 4 GB. The capacity of storage is 1 TB.

4.3. T1D Dataset Description

The Gene expression dataset used can be accessed from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. Gene expression data for type 1 diabetes T1D is used in the proposed system. The values are obtained by mapping the RMA algorithm using the Limma package in the R language. Two datasets are used, the first dataset with code GSE35725 is

downloaded in a raw file and has 114 samples, the sample identification (ID) between GSM874033 to GSM874146. UPN727 cells are produced with unrelated healthy control plasma (n=44), autologous plasma (n=7), longstanding T1D plasma (n=11), recent onset T1D plasma (n=46), or longitudinal series plasma (n=6 time points), Table (4.1) and Table (4.2) illustrate the details of the first T1D datasets.

Title of dataset	**T1D dataset**
Number of classes	5
Number of instances	114
Number of genes	54675

Table (4.1): Details of the First T1D Dataset.

The second dataset with code GSE52724 has 286 samples, and the sample identification (ID) is between GSM1274585 to GSM1274870. UPN727 cells are produced with Auto-antibody-negative (AA-) Low HLA Risk Siblings plasma

Table (4.2): Details of the First T1D Dataset with the Replicate Number.

Replicate number	Experiment	No. of samples
1	Autologous Unrelated health control Recent onset	1 3 7
2	Autologous Unrelated health control Recent onset	1 2 2
3	Autologous Unrelated health control Recent onset	1 4 6
4	Autologous Unrelated health control Recent onset	1 5 6
5	Autologous	1

	Unrelated health control Recent onset	16 19
6	Autologous Unrelated health control Recent onset	1 6 9
7	Autologous Unrelated health control Recent onset Longstanding Longitudinal series	1 2 3 11 6

(n=42), Auto-antibody-negative (AA-) High HLA Risk Siblings plasma (n=30), Recent onset T1DM plasma cultured with IL1RA (n=42), or longitudinal series plasma on T1DM progressor (n=27), on auto-antibody-negative (AA-) High HLA Risk Siblings (n=60), on auto-antibody-positive (AA+) High HLA Risk Siblings (n=54), on auto-antibody-negative (AA-) Low HLA Risk Siblings (n=31). Table (4.3) illustrates the details of the second T1D dataset.

Table (4.3): Details of the Second T1D Dataset.

Title of dataset	T1D dataset
Number of classes	7
Number of instances	286
Number of genes	54675

4.4. Results of Preprocessing Stage

There is a need to make four steps as pre-processing to facilitate the process of applying the classification models to T1D data.

4.4.1. Data cleaning: The T1D datasets include unwanted samples and therefore a certain number of samples has been determined for each dataset. The first dataset includes 114 samples, 90 samples (healthy control 44 and

resent onset 46) are selected, and the second dataset includes 286 samples, 91 samples (auto-antibody-negative (AA-) High HLA Risk Siblings 60 and auto-antibody-negative (AA-) Low HLA Risk Siblings 31) are selected, as illustrate in Figure (4.1).

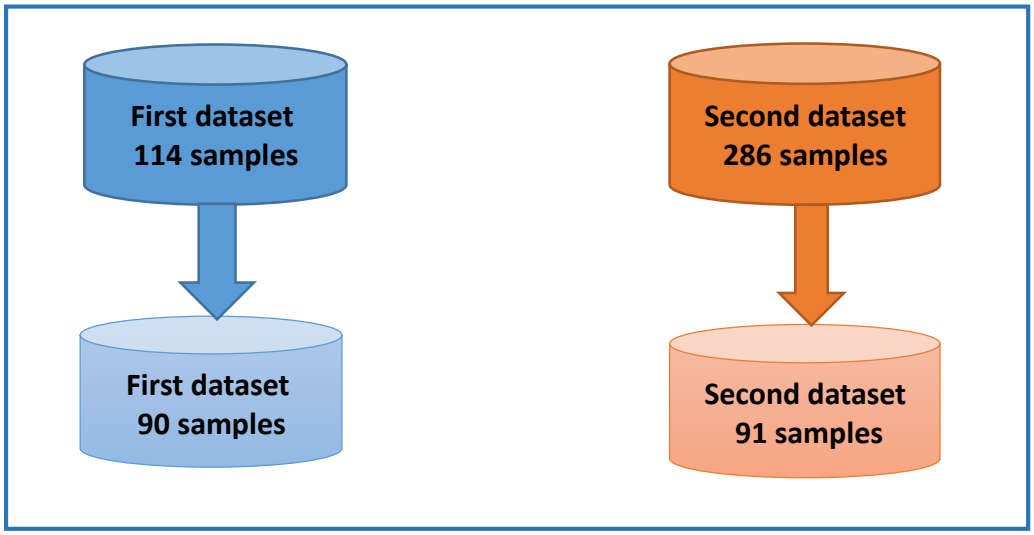

Figure (4.1): Summarization of Data Cleaning Process.

4.4.2. Data integration: after data cleaning of the two T1D datasets that have the codes (GSE35725) and (GSE52724), these datasets are integrated to be a new dataset used to work on. Table (4.4) illustrates the details of the resulting data, and Figure (4.2) shows the description of class numbers in the data. The categorical label is inserted into the resulting dataset, based on each sample type. Figure (4.3) shows the data after inserting the class label.

Table (4.4): Details of the Concatenated T1D Dataset.

Title of dataset	T1D dataset
Number of classes	4
Number of instances	181
Number of genes	54675

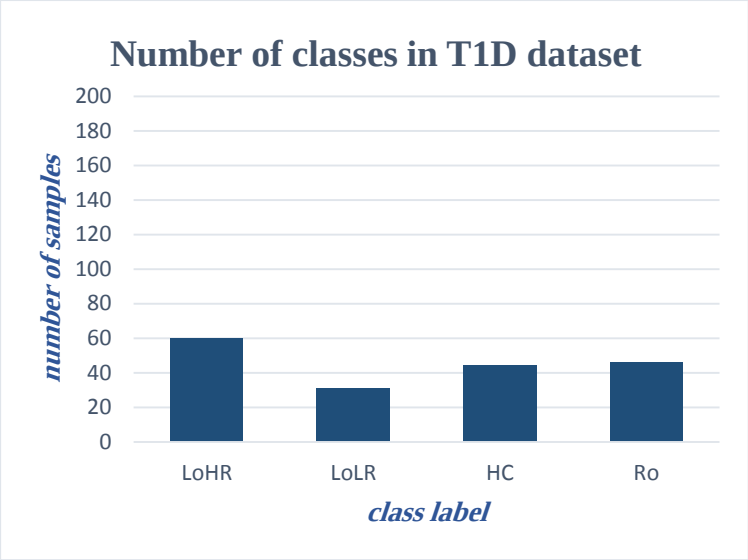

Figure (4.2): Number of Classes in the Concatenated T1D Dataset.

ıt	AFFX.r2.Ec.bioB.M_at	AFFX.r2.Ec.bioC.3_at	AFFX.r2.Ec.bioC.5_at	AFFX.r2.Ec.bioD.3_at	AFFX.r2.Ec.bioD.5_at	AFFX.r2.P1.cre.3_at	AFFX.r2.P1.cre.5_at	class_ID
2	8.886943	10.562086	10.037961	12.315261	12.018307	13.286399	13.068286	LOLR
9	8.891760	10.605547	10.051510	12.359409	12.139325	13.388932	13.178899	LOLR
7	8.804130	10.454272	10.243066	12.599541	12.271811	13.538746	13.439012	LOHR
5	8.278778	9.965606	9.670166	12.249959	11.863494	13.329613	13.119069	LOHR
9	8.467234	10.016323	9.955562	12.163728	11.731864	13.690298	13.469791	Rsent Onset
5	8.590368	9.973739	9.932847	12.250042	11.848898	13.688295	13.450672	Rsent Onset
7	8.756533	10.373639	10.095476	12.461384	12.066123	13.747875	13.491958	Healthy Controled
1	8.267198	10.094002	9.582314	12.023198	11.767945	13.372343	13.152104	LOHR
4	8.929728	10.645979	10.269578	12.310587	12.158987	13.487115	13.369457	LOHR
9	9.007457	10.418938	10.383533	12.625428	12.273643	13.775891	13.607125	Healthy Controled

Figure (4.3): A Sample of the T1D Dataset after Inserting the Class Labels.

4.4.3. Handling missing values: the raw data contained a set of genes that did not contain values or NaN values. Therefore, the mean is used to be the estimated value of the missing values of the corresponding column that contains values for the same gene for all remaining samples.

4.4.4. Normalization: Min-Max normalization is applied according to the equation (2.1) as mentioned in the chapter two section (2.5.4.1), it is convenient and useful in dataset problems that depend on classification, and the conversion of feature values for a specific and small range from 0 to 1. Figure (4.4) (a) shows the dataset before normalization, and (b) shows the dataset after normalization.

	X1007_s_at	X1053_at	X117_at	X121_at	X1255_g_at	X1294_at	X1316_at	X1320_at	X1405_i_at	X1431_at	.
0	6.473938	6.429887	4.553297	6.728591	2.162856	6.986925	5.445699	3.007922	10.614062	2.620390	.
1	6.532863	6.279546	4.654606	6.895716	1.984436	6.989370	5.481289	3.056686	10.850557	2.705002	.
2	6.228897	6.572721	4.525556	6.644601	1.875057	6.846834	5.494930	3.049621	10.504282	2.637369	.
3	6.376891	6.571219	4.482034	6.751540	2.222694	7.025842	5.313009	2.990400	10.502850	2.724437	.
4	6.350954	6.414588	5.254237	6.679166	1.969671	7.116201	5.967898	2.837656	10.738380	2.429475	.

(a)

	X1007_s_at	X1053_at	X117_at	X121_at	X1255_g_at	X1294_at	X1316_at	X1320_at	X1405_i_at	X1431_at	.
0	0.684004	0.572427	0.086847	0.455171	0.632080	0.387749	0.256399	0.665962	0.342334	0.395792	.
1	0.749740	0.430668	0.103515	0.576122	0.404602	0.390257	0.272440	0.736276	0.523848	0.489809	.
2	0.410642	0.707106	0.082283	0.394386	0.265148	0.244038	0.278588	0.726088	0.258076	0.414659	.
3	0.575741	0.705690	0.075122	0.471780	0.708372	0.427672	0.196592	0.640697	0.256977	0.511405	.
4	0.546806	0.558001	0.202170	0.419402	0.385776	0.520366	0.491766	0.420455	0.437750	0.183654	.

(b)

Figure (4.4): Data Min-Max Normalization (A) Before Normalization Step (B) After Normalization Step.

4.4.5. Data augmentation: the increasing of samples number used for the new data which has 181 samples, increased to 362. While it is used for the data without replicate that consist of seven subsets with different samples sizes but same number of features, Table (4.5) shows the samples number before and after applying the augmentation.

Table (4.5): The Number of Samples before Augmentation and after Augmentation.

Number of subset	Samples before augmentation	Samples after augmentation
1	11	22
2	5	10
3	11	22
4	12	24
5	36	72
6	16	32
7	24	46

4.5. The Ranked Features Stage's Result

The student's t-test has been used for ranking genes, and the number of genes required for all samples is determined to be 10,000 because it was the suitable number after trying many numbers, as the number of genes is reduced from 54,675 to 10,000 as illustrate in Figure (4.5). The resulting dataset after applying the ranking contained 10,000 genes and 181 samples.

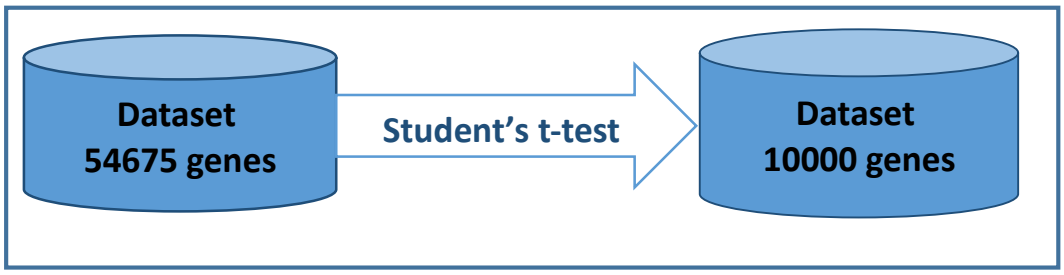

Figure (4.5): Summarization of Ranked Features Stage.

4.6. Classification Stage Results

The gene expression data are characterized by high dimensions due to a large number of genes. The situation requires selecting the most important genes related to the disease and eliminating the redundant genes. Four feature selection methods

have been used to select a subset of the dataset that includes the most important genes associated with T1D disease. Two parts have been implemented to achieve these methods, single feature selection and multiple feature selection.

4.6.1. Single Feature Selection Part

In this part, feature selection methods are used for each method separately to obtain a subset of the features, and the result of each method serves as an input for classification models.

4.6.1.1. The Result of the Mutual Information Method

Mutual information has been used to reduce the high dimensionality of the T1D gene expression dataset and select this disease's most relevant and influential features. The mutual information method is calculated between the gene and the class. Genes with a value of mutual information greater than 0.05 are selected, where 7542 genes out of 10,000 genes are selected as informational and important features, Figure (4.6) shows the mutual information of T1D data.

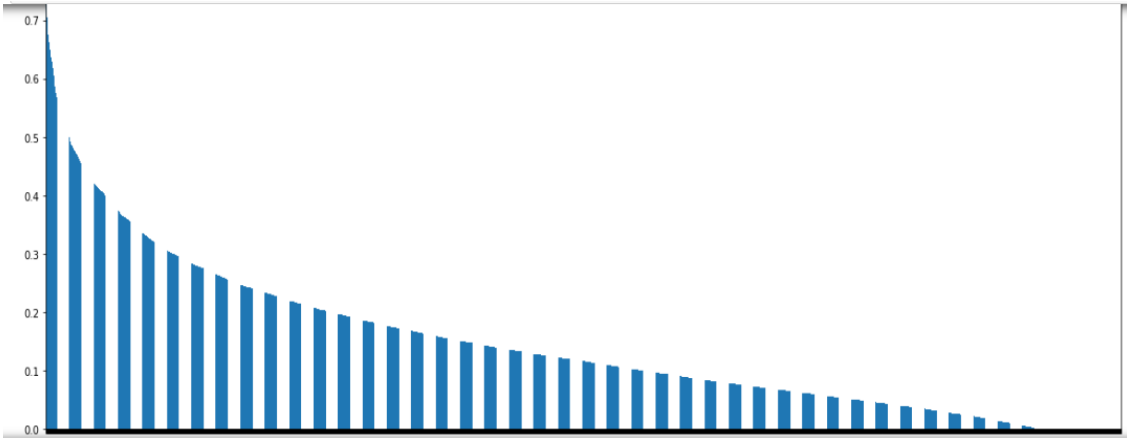

Figure (4.6): Mutual Information of the T1D Data.

4.6.1.2. The Result of the Analysis of Variance Method

The ANOVA method has been used to reduce the high dimensionality of the type 1 diabetes dataset and selects a set of features that most affect the disease. The number of features (genes) is reduced from 10,000 to 8583 after using ANOVA, where the p-value is set at 0.05. The result is passed into another feature selection method or classification model.

4.6.1.3. The Result of the Chi-Square Method

In this step, the Chi-square test has been used for selecting features (genes) and it is applied to all datasets with several genes of 10,000 genes, 8415 genes are selected according to the predefined threshold of 0.5. The result serves as input into another feature selection method or classification model.

4.6.1.4. The Result of the Principle Component Analysis Method

The PCA method reduces the dimensionality of the data by making the principle components and ordering them according to which has the highest variance. The covariance matrix provides the eigenvectors and eigenvalues as the selection of features in the PCA method depends on them, and using the threshold the number of eigenvectors is determined. The ranked T1D dataset is entered in the form of an array with dimensions of 10,000 genes and 181 samples. Before applying PCA, the augmentation process is implemented and the samples increased to 362 samples. When applying PCA to the dataset, the dimensions are reduced to 362 genes and 362 samples.

4.6.1.5. Random Forest Classification Model

In this step, a random forest is used to classify the T1D dataset after selecting a subset of features representing the most important features. In RF the dataset is divided into a training dataset and a test dataset, with a ratio of 80% to 20%. The number of trees used is determined by 100 trees of the classification model.

4.6.1.6. Support Vector Machine Classification Model

The linear kernel SVM has been used, and the T1D dataset obtained from the feature selection phase has been divided into two sets, the training set, and the test set. The ratio division is eighty present for train data and twenty present for test data.

4.6.1.7. Performance Evaluation

The classification models applied to the reduced data that resulted from the feature selection stage are evaluated by calculating the accuracy for each subset serving as an input into classification models. The highest accuracy of 94.5% is obtained by applying RF and SVM with the reduced data that resulted from the PCA feature selection method. Table (4.6) shows the accuracy of classification models.

Table (4.6): Accuracy of Classification Models for Single Part.

Feature selection method	Number of genes	Accuracy	
		RF	SVM
MI	7542	0.837	0.864
ANOVA	8583	0.837	0.81
Chi2	3967	0.756	0.891
PCA	362	**0.945**	**0.945**

4.6.2. Multiple Feature Selection Part

The feature selection methods have been implemented sequentially in many arrangements, where the subset obtained by one method is passed to other methods, the details of each path are described in Table (4.7). The arrangements are as follows:

a) ***MI-ANOVA****:* the subset of genes resulting from the MI method is 7542 genes and the result is passed into the ANOVA method, the resulting genes are 6819 genes. The selected subset of the genes served as input to the machine-learning classification models.

b) ***MI-PCA****:* the reduced data by the MI method is 7542 genes, then passed to the PCA method reduced the data to 362 genes and the selected subset of the genes be as input to the classification models.

c) ***ANOVA-PCA:*** the reduced data resulting from ANOVA is 8583 genes, used as input into the PCA method to obtain a subset of 362 genes. And the selected subset of the genes will be as input to the classification models.

d) ***MI- Chi2:*** the result of applying the MI method is 7542 genes, which passed as input into Chi2 for reducing it into 6550 genes. The selected subset of the genes served as input to the machine-learning classification models.

e) ***ANOVA-MI:*** the number of genes that resulted after implementing the ANOVA method is 8583 genes, then the result passed into the MI method to obtain 7103 genes. The selected subset of the genes is used as input to the classification models.

f) ***ANOVA-MI-PCA:*** 8583 genes are the result after applying the ANOVA feature selection method on the T1D dataset, then the result is used as input into the MI method to obtain 7103 genes, and the resulting subset of genes serves as input to the PCA method, the number of genes of reduced data is 362 genes. The selected subset of the genes served as input to the machine-learning classification models.

g) ***ANOVA-Chi2:*** the reduced data using ANOVA is 8583 genes, then the result passed as input into the Chi2 method to obtain 3967 genes. The selected subset of the genes served as input to the machine learning classification.

Table (4.7): Number of Genes Selected using Multiple Feature Selection.

	Feature Selection methods	**The Number of Selected Genes**			
1.	***MI-ANOVA***	MI:7542	ANOVA:6819		
2.	***MI-PCA***	MI:7542	PCA: 362		
3.	***ANOVA-PCA***	ANOVA:8583	PCA: 362		
4.	***MI-Chi2***	MI:7542	Chi2: 6550		
5.	***ANOVA-MI***	ANOVA:8583	MI: 7103		
6.	***ANOVA-MI-PCA***	ANOVA:8583	MI: 7103	PCA: 362	
7.	***ANOVA-Chi2***	ANOVA:8583	Chi2: 3967		

4.6.3. Random Forest Classification Model

In this step, a random forest is used to classify the T1D dataset after selecting a subset of features representing the most important features. In RF the dataset is divided into a training dataset and a test dataset, with a ratio of 80% to 20%. The number of trees used is determined by 100 trees of the classification model. The important features are selected by determining a threshold of 0.005 producing a result of 55 genes when the accuracy is 96.3% from the feature selection sequence the MI and then the PCA methods, Figure (4.7) illustrates the features (genes) and their values of feature importance where the vertical side represent the genes number in the dataset and the horizontal side represent the value of the importance feature, according to this value the features arranged in descending order.

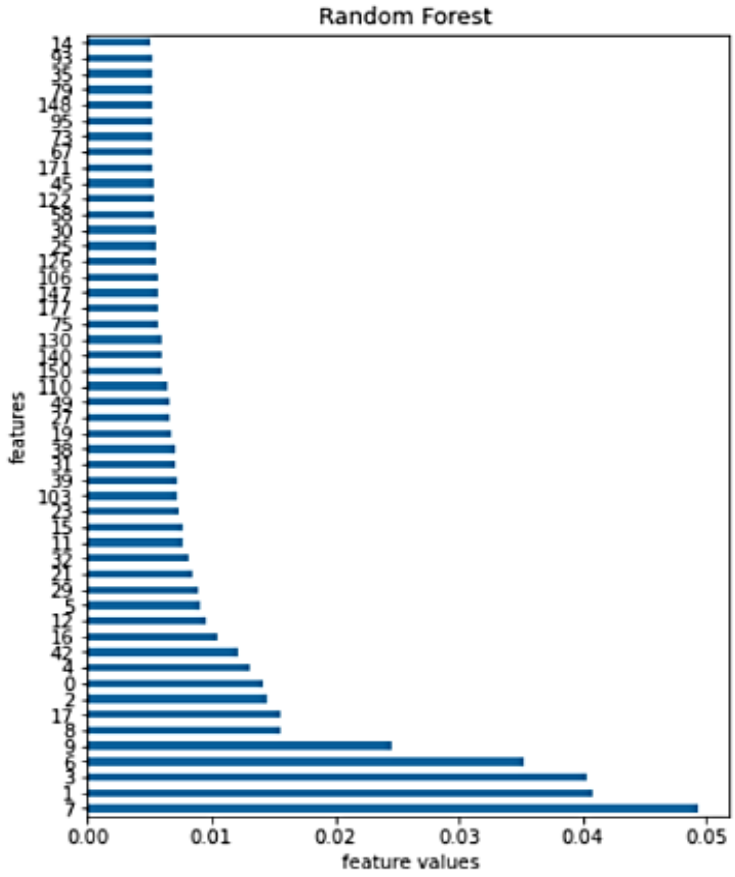

Figure (4.7): Random Forest Result for the Sequence (ANOVA-PCA).

4.6.3.1. Evaluation of Random Forest Model

In this step, the random forest model is implemented many times according to the reduced data by a sequence of feature selection methods. The highest accuracy of **96.3** % is obtained when applying the RF to the data that resulted from the following sequence, which is first the MI method, and then the PCA method. The percent of dividing is 30 % for the test set and 70 % for the train set. Table (4.8) shows the accuracy of the random forest model.

Table (4.8): The Accuracy of the Random Forest Model for the Multiple Part.

RF 10000	Feature Selection	Accuracy 20%	30%
1.	MI-ANOVA	0.836	0.836
2.	MI-PCA	0.917	**0.963**
3.	ANOVA-PCA	0.945	0.889
4.	MI-chi2	0.243	0.272
5.	ANOVA-MI	0.837	0.854
6.	ANOVA-MI-PCA	0.945	0.908
7.	ANOVA-chi2	0.756	0.836

4.6.4. Support Vector Machine Classification Model

The linear kernel SVM has been used, the T1D dataset obtained from the feature selection stage has been divided into two sets, the train set and the test set. The ratio division is 80% for train set and 20% for test set. Different values of thresholds are used to select the important features, Table (4.9) illustrate the threshold when the accuracy is the best two times.

Table (4.9): SVM Threshold of Multiple Feature Selection Methods.

Feature selection method sequence	Threshold	Number of Selected genes
ANOVA- PCA	0.03	44
MI-PCA	0.005	31
ANOVA-MI-PCA	0.005	25

4.6.4.1. Evaluation of Support Vector Machine

The support vector machine model is applied for different sequences of feature selection methods. The reduced data resulted from feature selection methods using two different arrangements that are passed as input into the SVM model. The model produced the highest accuracy of 97.2 % two times, in which the percent of dividing is 80 % for the train test and 20 % for the test set. Table (4.10) shows the accuracy of the support vector machine for the multiple part.

Table (4.10): The Accuracy of SVM for Multiple Feature Selection Part.

SVM 10000	Feature Selection	Accuracy 20%	30%
1.	MI-ANOVA	0.864	0.872
2.	MI-PCA	**0.972**	0.963
3.	ANOVA-PCA	0.945	0.963
4.	MI-chi2	0.864	0.872
5.	ANOVA-MI	0.864	0.872
6.	ANOVA-MI-PCA	**0.972**	0.963
7.	ANOVA-chi2	0.947	0.891

4.7. Regression Stage Results

The three models of machine learning regression are used for the first T1D datasets (114 samples) only, these models are implemented for the dataset in two states, data within the replicate and the data without the replicate. Before applying the regression models, the class labels of the dataset are converted from categorical values to continuous values. Values range from 1.0 to 4.0, Table (4.11) illustrate each class label corresponding to continuous values.

Table (4.11): The Continuous Values According to Class Label Values.

Class label value	Continuous value
Recent onset	1.0
Healthy control	2.0
High HLA Risk	3.0
Low HLA Risk	4.0

4.7.1. Data Without Replicate

The data departed into seven subsets according to the number of replicates that exists in the original data, these replicates are described in Table (4.2), called without replicate data. Each subset contains a different number of samples but the same number of genes.

4.7.1.1. Ranking

Student's t-test is applied to all features of the data, many attempts were applied for determining the suitable ranking number like (1000, 10000, 54675) , the chosen number for ranking these features is 10000, the genes ranked of 10000 genes from the total of 54675 genes. The resulting genes from this test are passed as input to the regression model for made the T1D prediction.

4.7.1.2. Linear Regression Model

In this step, the multiple linear regression MLR model is implemented for predicting the T1D dataset, due to the multiple features input to the model. Two splitting percent are used, the first ratio is 20% for the test set and 80% for the train set, and the second ratio is 30% for the test set and 70% for the train set. The model is implemented on continuous features to predict labels of T1D disease.

4.7.1.3. Performance Evaluation

The linear regression model is evaluated by calculating the root means square error RMSE, the best value obtained of RMSE is 0.00. Through the seven subsets and the two percent dividing, many times obtained the best value of RMSE. Table (4.12) shows the RMSE of linear regression applied the without replicating data.

Table (4.12): The RMSE of Linear Regression for the Data without Replicate.

Set number	Number of samples	RMSE 20%	30%
1.	22	0.00	0.24
2.	10	0.00	0.00
3.	22	0.00	0.16
4.	24	0.12	0.39
5.	72	0.00	0.32
6.	32	0.00	0.01
7.	46	0.00	0.23

4.7.2. Data Within Replicate

All three models of ML regression are applied to the data within replicate. The data is used in two cases, in the first case all features served as input into the regression models, the number of all genes in the data is 54675 genes. The second

case is after applying the student's t-test, where the features (genes) are reduced to 10000 genes. There are two division ratios used: 20% for the testing set and 80% for the training set, and 30% for the testing set and 70% for the training set.

4.7.2.1. Linear Regression Model

In this step, the multiple linear regression MLR model is used for predicting the T1D dataset due to the multiple features input to the model. Before applying the model, the data are split into a train set and a test set as mentioned in the previous subsection. The model is implemented on continuous features to predict labels of T1D.

4.7.2.2. Random Forest Regression Model

The random forest regression model is used to predict the T1D dataset genes. After dividing data into a train set and a test set, the number of trees were determined using many of trying numbers such as (50, 100, 300, 600) trees, the selected number of trees used in the models in this thesis is 100 trees, each tree predicts labels and the final decision depends on the average of the tree's result from the prediction. The model trained using the train set of data then the test set used for the predicting.

4.7.2.3. Support Vector Regression Model

The support vector regression model has been used for predicting the genes of the T1D dataset, the linear kernel is used for implementing the model onto the selected features. Before implementing the model, the data are spilt into two sets, a train set and a test set, the ratio is two different values indicated previously.

4.7.2.4. Performance Evaluation

The three of machine learning regression models that are used for predicting, included linear regression, random forest regression, and support vector regression. The root means square error metric is used to evaluate the performance of regression models. The best lowest RMSE result of 1.20 is obtained from the linear regression model that is implemented after applying the ranking data. The data is divided by 20% for the test set and 80% for the train set, as shown in Table (4.13).

Table (4.13): The RMSE of Regression Models for Ranked Data within Replicate.

RMSE		
114(ranked)	**20%**	**30%**
Linear regression	**1.20**	1.39
SVR	1.207	1.388
RF regression	1.341	1.5

While all three models of regression are applied to T1D data without implementing the ranking stage, where all number of genes 54675 are used for predicting. The lowest RMSE result of 1.22 by the linear regression model, in which the ratio of dividing is 20% of the test set and 80% of the train test, Table (4.14) shows the result of RMSE for regression models.

Table (4.14): The RMSE of Regression Models for Unranked Data within Replicate.

RMSE		
114(unranked)	**20%**	**30%**
Linear regression	**1.22**	1.40

SVR	1.225	1.395
RF regression	1.451	1.481

4.8. Name of predicted genes

The prediction model of T1D resulted a genes that represented the most related and affective genes of this disease, where these genes obtained after applying the multiple feature selection methods first the MI then PCA, and the obtained subset of features are used as input for the classification machine learning models, the names this genes are shown in Figure (4.8).

Figure (4.8): The Names of Predicted Genes for T1D

		gene names	1
X1552703_s_at	15	X1558869_at	2
X1552445_a_at	16	X1559911_at	3
X1553344_at	17	X1557113_at	4
X1552952_at	18	X1557124_at	5
X1552944_a_at	19	X1561860_at	6
X1555786_s_at	20	X1561976_at	7
X1554919_s_at	21	X1561487_at	8
X1555487_a_at	22	X1561546_at	9
X1566695_at	23	X1562420_at	10
X1566658_at	24	X1560746_at	11
X1566897_at	25	X1560115_a_at	12
X1562935_at	26	X1553567_s_at	13
X1563572_at	27	X1563662_at	14

4.9. Summary

This section presented a summary of results obtained by the prediction model, the best results and the worst. In the classification part, the best accuracy of machine learning models is obtained four times, when passed the reduced data, using sequence feature selection methods, as input to predict the T1D dataset. In addition

to accuracy, recall and precision are used to evaluate the performance of ML models. Table (4.15) shows a summary of the best result of the classification part.

Table (4.15): A Summary of the Best Accuracy of Machine Learning Models.

ML model	Feature selection	Accuracy	Recall	Precision
RF	MI-PCA	96.3%	96%	96%
SVM	MI-PCA	97.2%	97%	97%
SVM	ANOVA-MI-PCA	97.2%	97%	97%

In the same part, the worst accuracy ranged (from 24%-27%), which resulted when using the reduced data obtained from Multiple feature selection, first the mutual information method and then the chi-square method, and served as input for the random forest classification model.

In the regression part, the lowest RMSE value of 0.00 is obtained more than one time, when using the seven subsets of the data called (without replicate data) as input for the linear regression model.

Chapter Five

Conclusions and Future Works

Chapter Five

Conclusions and Future Works

5.1. Conclusions

The most essential characteristics revealed during the implementation of the proposed methodology and the discussion of its results are the following:

1. The proposed system has effectively proved in identifying the relevant genes (the best genes) and removing irrelevant one. In addition, this proposed system gives promising results in the prediction models according to all common evaluation measures.

2. The proposed system has effectively proved in multiple feature selection methods according to the T1D dataset nature to such an extent that the results are satisfactory in prediction models.

3. According feature selection methods, the genes are significantly reduced, which had a positive effect on the results.

i. The first usage for the feature selection methods is single feature selection, where every method applied separately and reduced the genes to serve as input for the classification models.

ii. The second usage for the feature selection methods is multiple feature selection, where the reduced genes that obtained by one method served as input to another method. Seven arrangements (paths) are implemented, at last every path reduced the genes to significant genes.

4. The results indicate that the accuracy of the classification models for the MI, ANOVA, Chi^2, and PCA methods in single feature selection method part. Whereas by using the genes the best accuracy from the multiple feature selection part from two different arrangements, because of these multiple feature selection methods reduced the genes into more related genes to serve as input into machine learning methods, these paths are:
   i. MI-PCA: the genes have been reduced firstly by mutual information method, the obtained result passed as input for principle component analysis method.
   ii. ANOVA-MI-PCA: ANOVA is the first method applied for reducing the genes, the result passed as input to the MI method for reducing these genes, finally the reduced genes are served as input to the PCA method.

5. The worst accuracy is obtained when using RF model with the genes selected from the path MI- Chi^2 due to the Chi^2 genes obtained are not the most related for T1D.
6. The best RMSE has been obtained when using the dataset after departed it into seven subsets form the linear regression model .

7. The worst RMSE is obtained from the RF regression model when using the data within replicate after ranked it by the student's t-test.

5.2. The Future Works

Some future directions can be highlighted here:

1. Applying the proposed system on other datasets such as Alzheimer dataset, prostate dataset, and colon cancer dataset, etc.
2. Implementing the deep learning techniques on the dataset for trying to minimize the error of prediction.
3. Using the bagging and boosting with the regression part for enhancing the gene prediction.
4. Choosing a specific method to determine the thresholds for all models and feature selection methods.

References

[1] F. Wang, J. Liang, D. Zhu, P. Xiang, L. Zhou, and C. Yang, "Characteristic gene prognostic model of type 1 diabetes mellitus via machine learning strategy," pp. 1–14, 2019. doi: 10.1507/endocrj.EJ22-0178.

[2] J. Enticott, "children : A machine learning strategy for large-p , small-n scenarios," pp. 43–51, 2022.

[3] S. Geravandi, H. Liu, and K. Maedler, "Enteroviruses and t1d: Is it the virus, the genes or both which cause t1d," *Microorganisms*, vol. 8, no. 7. MDPI AG, pp. 1–20, Jul. 01, 2020. doi: 10.3390/microorganisms8071017.

[4] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/fgene.2018.00515.

[5] V. Majhi, S. Paul, and R. Jain, "Bioinformatics for Healthcare Applications Bioinformatics for Healthcare Applications," *2019 Amity Int. Conf. Artif. Intell.*, no. February, pp. 204–207, 2019, doi: 10.1109/AICAI.2019.8701277.

[6] A. Bayat, "Clinical review," vol. 324, no. April, pp. 1018–1022, 2002.

[7] A. U. Haq *et al.*, "Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data," *Sensors (Switzerland)*, vol. 20, no. 9, May 2020, doi: 10.3390/s20092649.

[8] A. Czmil, S. Czmil, and D. Mazur, “applied sciences A Method to Detect Type 1 Diabetes Based on Physical Activity Measurements Using a Mobile Device,” pp. 1–16, 2019, doi: 10.3390/app9122555.

[9] J. Lin, Y. Lu, B. Wang, P. Jiao, and J. Ma, “Analysis of immune cell components and immune-related gene expression profiles in peripheral blood of patients with type 1 diabetes mellitus,” *J. Transl. Med.*, vol. 19, no. 1, pp. 1–16, 2021, doi: 10.1186/s12967-021-02991-3.

[10] A. M. Elsherbini *et al.*, “Decoding Diabetes Biomarkers and Related Molecular Mechanisms by Using Machine Learning, Text Mining, and Gene Expression Analysis,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 21, pp. 1–18, 2022, doi: 10.3390/ijerph192113890.

[11] L. M. and J. Zheng2*, “Single-cell gene expression analysis reveals β-cell dysfunction and deficit mechanisms in type 2 diabetes,” in *29th International Conference on Genome Informatics*, 2018. doi: 10.1186/s12859-018-2519-1.

[12] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, and Z. Mansoori, “Type2 diabetes mellitus prediction using data mining algorithms based on the long- noncoding RNAs expression : a comparison of four data mining approaches,” pp. 1–13, 2020.

[13] H. Alshamlan and H. Bin Taleb, “A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression,” pp. 38–41, 2020, doi: 10.1109/ICICS49469.2020.239549.

[14] J. Li, J. Ding, D. U. Zhi, K. Gu, and H. Wang, “Identification of Type 2 Diabetes Based on a Ten-Gene Biomarker Prediction Model Constructed Using a Support Vector Machine Algorithm,” *Biomed Res. Int.*, vol. 2022, 2022, doi: 10.1155/2022/1230761.

[15] T. Turki and Y. Taguchi, “Discriminating the single-cell gene regulatory networks of human pancreatic islets : A novel deep learning application,” *Comput. Biol. Med.*, vol. 132, no. September 2020, p. 104257, 2021, doi: 10.1016/j.compbiomed.2021.104257.

[16] J. Lin, Y. Lu, B. Wang, P. Jiao, and J. Ma, “Analysis of immune cell components and immune-related gene expression profiles in peripheral blood of patients with type 1 diabetes mellitus,” *J. Transl. Med.*, vol. 19, no. 1, pp. 1–16, 2021, doi: 10.1186/s12967-021-02991-3.

[17] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” in *Procedia Computer Science*, 2020, vol. 167, pp. 706–716. doi: 10.1016/j.procs.2020.03.336.

[18] N. Mahendran, P. M. Durai Raj Vincent, K. Srinivasan, and C. Y. Chang, “Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions,” *Frontiers in Genetics*, 2020. https://courses.lumenlearning.com/boundless-statistics/chapter/tests-for-ranked-data/

[19] H. Muthanna, “Gene Expression Classification of Alzeheimar Disease Stages Using Machine Learning,” 2021.

[20] F. Morais-rodrigues *et al.*, “Analysis of the microarray gene expression for breast cancer progression after the application modi fi ed logistic regression,” vol. 726, no. July 2019, 2020, doi: 10.1016/j.gene.2019.144168.

[21] S. Guo, M. R. Lyu, and T. Lok, “Gene Selection Based on Mutual Information for the Classification of Multi-class Cancer,” pp. 454–463.

[22] A. K. Al-Mashanji and S. Z. Al-Rashi, “Computational Methods for Preprocessing and Classifying Gene Expression Data- Survey,” *4th Sci. Int. Conf. Najaf, SICN 2019*, pp. 121–126, 2019, doi: 10.1109/SICN47020.2019.9019349.

[23] S. Gao, N. Wolanyk, Y. Chen, S. Jia, M. J. Hessner, and X. Wang, “Investigation of coordination and order in transcription regulation of innate and adaptive immunity genes in type 1 diabetes,” *BMC Med. Genomics*, vol. 10, no. 1, pp. 1–14, Jan. 2017, doi: 10.1186/s12920-017-0243-8.

[24] C. Sen Seah *et al.*, “An effective pre-processing phase for gene expression classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, 2018. https://courses.lumenlearning.com/boundless-statistics/chapter/tests-for-ranked-data/

[25] Suad A. Alasadi and Wesam S. Bhaya, “Review of Data Preprocessing Techniques in Data Mining,” *J. Eng. Appl. Sci.*, 2017.

[26] G. Y. Lee and L. Alzamil, “Model Performance,” no. September, pp. 0–6, 2021.

[27] F. A. Aziz, “A Deep Learning-based Prediction Transcription Factor Binding Sites of DNA Sequences,” University of Babylon\ College of Information Technology, 2022.

[28] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, *A survey on missing data in machine learning*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00516-9.

[29] K. Maharana, S. Mondal, and B. Nemade, “A review : Data pre-

processing and data augmentation techniques," vol. 3, no. April, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.

[30] S. G. K. Patro and K. Kumar, "Normalization : A Preprocessing Stage," no. March, 2015, doi: 10.17148/IARJSET.2015.2305.

[31] Z. S. and B. K. Luai Al Shalabi, "Data Mining : A Preprocessing Engine Luai Al Shalabi , Zyad Shaaban and Basel Kasasbeh Applied Science University , Amman , Jordan," J. Comput. Sci., vol. 2, no. 9, pp. 735–739, 2006.

[32] T. K. Kim, "T test as a parametric statistic," no. Table 2, 2015.

[33] A. Al-achi, "The Student ' s t-Test : A Brief Description," J. Hosp. Clin. Pharm., vol. 5, no. 1, pp. 4–6, 2019.

[34] J. Englund, "Another Student ' s T -test Proposal and evaluation of a modified T-test," 2014.

[35] M. Kumar, N. K. Rath, A. Swain, and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," *Procedia - Procedia Comput. Sci.*, vol. 54, pp. 301–310, 2015, doi: 10.1016/j.procs.2015.06.035.

[36] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, Mar. 2021, doi: 10.3390/ijerph18063317.

[37] F. K. Ahmad and K. Kerian, "A Review of Feature Selection Techniques via Gene Expression Profiles," 2008.

[38] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," *Expert Syst. Appl.*, vol. 140, p. 112873, 2020, doi: 10.1016/j.eswa.2019.112873.

[39] N. Barraza, S. Moro, M. Ferreyra, and A. de la Peña, “Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study,” *J. Inf. Sci.*, vol. 45, no. 1, pp. 53–67, 2019, doi: 10.1177/0165551518770967.

[40] W. Zhongxin, S. Gang, Z. Jing, and Z. Jia, “Feature Selection Algorithm Based on Mutual Information and Lasso for Microarray Data,” no. 193, pp. 278–286, 2016, doi: 10.2174/1874070701610010278.

[41] H. Zhou, X. Wang, and R. Zhu, “Feature selection based on mutual information with correlation coefficient,” Appl. Intell. Springer, 2021, doi: 10.1007/s.

[42] E. Odhiambo, G. Onyango, and M. Waema, “Feature Selection for Classification using Principal Component Analysis and Information Gain,” *Expert Syst. Appl.*, vol. 174, no. November 2020, p. 114765, 2021, doi: 10.1016/j.eswa.2021.114765.

[43] S. O. S. P. Science, “TOPIC NAME : Chi-Square Test : Uses | Statistics Meaning , Applications and,” vol. 2.

[44] M. O. Arowolo, M. Adebiyi, A. Adebiyi, and O. Okesola, “PCA Model for RNA-Seq Malaria Vector Data Classification Using KNN and Decision Tree Algorithm,” *2020 Int. Conf. Math. Comput. Eng. Comput. Sci. ICMCECS 2020*, 2020, doi: 10.1109/ICMCECS47690.2020.240881.

[45] S. Karthik and M. Sudha, “A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases,” *International Journal of Engineering and Advanced Technology*, 2018. https://courses.lumenlearning.com/boundless-

statistics/chapter/tests-for-ranked-data/

[46] I. Journal, C. Science, and I. T. Ijsrcseit, “Breast Cancer Prediction using SVM with PCA Feature Selection Method”.

[47] D. Fernández-Edreira, J. Liñares-Blanco, and C. Fernandez-Lozano, “Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes,” *Expert Syst. Appl.*, vol. 185, Dec. 2021, doi: 10.1016/j.eswa.2021.115648.

[48] D. H. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” vol. 01, no. 04, pp. 140–147, 2020, doi: 10.38094/jastt1457.

[49] A. C. Study, “applied sciences Data Mining Techniques for Early Diagnosis of Diabetes :,” pp. 1–12, 2021.

[50] M. Ram, A. Najafi, and M. T. Shakeri, “Classification and biomarker genes selection for cancer gene expression data using random forest,” *Iran. J. Pathol.*, vol. 12, no. 4, pp. 339–347, 2017, doi: 10.30699/ijp.2017.27990.

[51] R.Khanna and M.Awad, *Efficient Machine Learning*.

[52] O. Okun and H. Priisalu, “Random forest for gene expression based cancer classification: Overlooked issues,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4478 LNCS, no. PART 2, pp. 483–490, 2007, doi: 10.1007/978-3-540-72849-8_61.

[53] J. C. C. Vladimir Svetnik,*,† Andy Liaw,† Christopher Tong and and B. P. F. Robert P. Sheridan, “Random Forest A Classification and Regression Tool for Compound Classification and QSAR Modeling.pdf.” 2003.

[54] D. Miao, H. Tang, and B. Wang, “Support Vector Regression with Gaussian kernel for Housing Prices Prediction,” *J. Phys. Conf. Ser.*, vol. 1994, no. 1, 2021, doi: 10.1088/1742-6596/1994/1/012023.

[55] C. Science, C. Science, and C. Science, “A Comparative Analysis on Linear Regression and Support Vector Regression,” 2016.

[56] I. B. Copy, C. N. Network, and C. Hevc, “MULTIPLE LINEAR REGRESSION FOR HIGH EFFICIENCY VIDEO INTRA CODING Zhaobin Zhang University of Missouri Kansas City University of Science and Technology of China Tencent America”.

[57] I. Journal, D. Mining, and K. M. Process, “A REVIEW ON EVALUATION METRICS FOR DATA,” no. April, 2020.

Appendix A

Data Set Specification

When the two datasets used together, further the age is restricted at the time of study to between 6 and 20 years old. The human leukocyte antigen (HLA) system is a gene complex encoding the major histocompatibility complex (MHC) proteins in humans. HLA types are inherited and some of them are linked to autoimmune disorders and/or other diseases, including T1D diabetes. The data contains transcription profiles of cryopreserved peripheral blood mononuclear cells PBMC responding to the sera from 148 human subjects belonging to one of the following cohorts:

1. Unrelated Healthy Controls (uHC) with no familial history of any autoimmune/auto inflammatory disorder.
2. Recent Onset T1D (RO-T1D) patients. Blood samples were collected after stabilization on exogenous insulin 2–7 months after diagnosis.
3. Autoantibody-negative siblings of T1D patients that are of high genetic risk (with DR3/4 haplotypes of HLA) for T1D (HRS).
4. Autoantibody-negative siblings of T1D patients of low genetic risk (with non-DR3/4 HLA genotypes) for T1D (LRS).

Appendix B

The Published Paper

Gene Expression Dataset Classification Using Machine Learning Methods: A Survey

Publisher: IEEE Cite This PDF

Noor AlRefaai ; Sura Z. Al-Rashid All Authors

Abstract:
The process of converting the instructions in the DNA of human beings into a functional product such as proteins is called gene expression. Gene expression is an organized process that allows the cell to adapt to its changing environment after converting the information stored in DNA into instructions, forming the basis for making protein and other molecules. It is a tightly regulated process that allows a cell to respond to its changing environment. Gene expression classification is important because it helps a lot in the prediction, prevention, and early control of some genetic diseases. Bioinformatics is a broad field that specializes in developing software methods and techniques for analyzing and understanding biological data, especially large ones that are difficult to analyze without software assistance. Bioinformatics is not limited to biology and computer sciences only, but chemistry, physics, mathematics, statistics, and information engineering participate in this discipline to help analyze and interpret large-scale biological data. Machine learning methods are used extensively in bioinformatics on the biological data sets and provide good results, such as Random Forest and Support Vector Machine. In this paper, a comprehensive survey is made of the recently used machine learning methods for gene expression classification.

Published in: 2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)

Date of Conference: 31 August 2022 - 01 September 2022

INSPEC Accession Number: 22572763

Date Added to IEEE *Xplore*: 03 February 2023

DOI: 10.1109/ICCITM56309.2022.10031279

ISBN Information:

Publisher: IEEE

Conference Location: Mosul, Iraq

Appendix C

The Accepted Paper

B **Bulletin of Electrical En...** Nov 28

to me

-- Paper ID# 4322
-- Please strictly use and follow the template manuscripts: http://iaescore.com/gfa/beei.docx (including biographies of the authors)

Dear Prof/Dr/Mr/Mrs: Noor Ali AlRefaai,

It is my great pleasure to inform you that your paper entitled "Classification of Gene Expression Dataset for Type 1 Diabetes using Machine Learning Methods" has been initially ACCEPTED and will be published on the Bulletin of Electrical Engineering and Informatics (BEEI), ISSN: 2089-3191, e-ISSN: 2302-9285, http://beei.org. This journal is indexed by Scopus (https://www.scopus.com/sourceid/21100826382) and ScimagoJR (https://www.scimagojr.com/journalsearch.php?q=21100826382&tip=sid&clean=0). Congratulations!

Please prepare your final camera-ready paper (in MS Word file format), adhere to every detail of the final checklist at https://beei.org/index.php/EEI/about/editorialPolicies#custom-4, and check it for spelling and grammatical mistakes. For your information, according to international regulations, the similarity score of camera-ready paper should be less than 25%. A single author is NOT acceptable, and will never publish in this journal. The editor(s) will decide whether or not your camera-ready paper is complete and meets the final checklist. Please present each reference as completely as possible while adhering to IEEE style (including volume, number, pages, and DOIs). Failing to make a proper revision may lead to the delay of your paper for publication, even may lead to the rejection of your paper.

الخلاصة

مرض السكري من النوع الأول (T1D) من الامراض المزمنة تحدث بسبب تدمير في خلايا بيتا الموجودة في البنكرياس، حيث إن البنكرياس يفرز كَمّيَّة قليلة من الأنسولين أو لا يفرز الأنسولين على الإطلاق. ان جسم الانسان يستخدم سكر الدَّم (الغلوكوز) لتوليد الطاقة ويدخل خلايا الجسم بواسطة هرمون الأنسولين. ان المشكلة الرئيسة في T1D انه من الامراض العضال، لذلك يركز العلاج على التحكم في مستويات الغلوكوز في الدَّم. مرض T1D يورث عن طريق الجينات لهذا يكون تأثير العوامل الوراثية بشكل كبير على تطوره، قد يساعد العثور على الجينات (السمات) التي تسبب هذا المرض في تنظيم حالة المريض و البدأ بالعلاج المبكر. من التقنيات الحديثة التي تمكن العلماء من تحديد مستويات التعبير لمئات الآلاف من الجينات بشكل متزامن هي المصفوفة الدقيقة التي تعتبر أداة مفيدة لهذه الاغراض.

الهدف الرئيسي من هذه الأطروحة هو بناء نموذج يتنبئ بمرض T1D وبأعلى دِقَّة ممكنة وأقل خطأ نتيجة لاختيار الجينات الأكثر أهمية واكثر ارتباط بهذا المرض(الجينات المعلوماتية). يتكون النظام المقترح من مرحلتين رئيستين: مرحلة اختيار الجينات ومرحلة التصنيف والانحدار. تتم عملية اختيار الجينات باستخدام جزأين جزء طرق اختيار الميزة المنفردة و جزء طرق اختيار الميزة المتعددة، تستخدم هذه العملية لاختيار مجموعة فرعية من الجينات المهمة وتحسين دِقَّة التنبؤ للنموذج المقترح، تتضمن طرق اختيار المِيزة أربع طرق: المعلومات المتبادلة، تحليل التباين، Chi^2، وتحليل المكونات الرئيسية.

يتعرف جزء اختيار الميزة المتعددة للنظام المقترح على الجينات الأكثر إفادة في كل مرحلة ويختار الجينات التالية بناءً على الجينات المختارة بواسطه الطريقة السابقة ثم يتم إدخالها في النموذج لتأكيد أهميتها وإيجاد الدِّقَّة التي يمكن تحقيقها باستخدام هذه الجينات، بينما يؤدي النظام المقترح جزء طرق اختيار المِيزة المنفردة الذي يختار الجينات بكل طريقة بشكل منفصل، ثم يتم إدخال الجينات المختاره في النموذج. إضافةً إلى ذلك، سعى هذا العمل إلى تقديم نموذج تنبؤ يعتمد على نماذج التصنيف والانحدار لتحديد الفرق بين المرضى الذين يعانون مرض السكري من النوع الأول من الأشخاص الذين ليس لديهم ناقل جيني. تم استخدام مجموعة البيانات المتاحة لتحقيق أهداف الأطروحة الحالية وهي: مجموعة البيانات T1D. تم إجراء التقييم اعتمادًا على مقاييس التنبؤ (الدِّقَّة، الاسترجاع، و جذر متوسط الخطأ التربيعي RMSE). بينت النتائج أن أداء النظام المقترح فعال، حيث بلغت دِقَّة التنبؤ (٠,٩٧٢) من جزء طرق اختيار الميزة المتعددة لنماذج التصنيف. بلغ أدنى RMSE (٠,٠٠) من نموذج الانحدار الخطي باستخدام البيانات دون تكرار.

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة باب لكلية تكنولوجيا المعلومات
قسم البرمجيات

مراحل التنبؤ بمرض السكري من النوع الأول بأستخدام طرق تعلم الآله

رسالة مقدمة إلى

مجلس كلية تكنولوجيا المعلومات - جامعة بابل كجزء من متطلبات

نيل درجة الماجستير في تكنولوجيا المعلومات / البرمجيات

من قبل

نور علي فاضل مهدي

بأشـراف

الاستاذ المساعد الدكتوره سرى زكي

1444 هـ

2023 م