2021

Comparison of Performances of Machine Learning Techniques in Healthcare Data

Maniruzzaman, Md.

University of Rajshahi, Rajshahi

Comparison of Performances of Machine Learning Techniques in Healthcare Data

This thesis submitted in fulfillment of the requirements for the Degree of Master of Philosophy in Statistics

By

Md. Maniruzzaman
Reg. no. 1810624503
Session: 2017-2018

Department of Statistics
University of Rajshahi

February, 2021

Dedicated to
Beloved Parents

Declaration

I hereby declare that the research work embodied in this thesis entitled "**Comparison of Performances of Machine Learning Techniques in Healthcare Data"** has been carried out by me for the degree of Master of Philosophy under the direct supervision of **Dr. Md. Jahanur Rahman**, Professor, Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh. I also declare that the results presented in this dissertation are my own investigation and any part of this thesis work has not submitted to elsewhere for any degree or diploma or similar purpose.

Md. Maniruzzaman
MPhil Fellow
Reg. No. 1810624503
Session: 2017-2018
Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh

UNIVERSITY OF RAJSHAHI
RAJSHAHI, BANGLADESH

Certificate of Approval

I have the pleasure to certify that the thesis entitled "**Comparison of Performances of Machine Learning Techniques in Healthcare Data"** is an original research done by **Md. Maniruzzaman.** He has completed the research work under my supervision. As far as I know, the thesis has not been previously submitted to any other university/institute for any kind of degree or diploma.

I also certify that I have perused the thesis and found it satisfactory for submission to Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh for the degree of Master of Philosophy.

Dr. Md. Jahanur Raman
Principal Supervisor
and
Professor
Department of Statistics
University of Rajshahi
Rajshahi-6205
Bangladesh.

Table of Contents

Acknowledgements

First of all I would like to express my sincere gratitude to my honorable supervisor, **Professor Dr. Md. Jahanur Rahman** for his constant supervision, inspiring guidance, enthusiastic encouragement, wise advice and affectionate throughout the entire period of my research work. I am grateful to my mentor **Dr. Jasjit S Suri**, Professor, Department of Electrical Engineering, Idaho State University (Affl.), Idaho, USA who also inspired me a lot to accomplish the research by providing his valuable comments, suggestions and edits. I am also grateful to all the faculty members of Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh who helped me to require knowledge on the various aspects of research. I am also grateful to **Dr. Md. Al Mehedi Hasan**, Professor, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi-6204, Bangladesh, **Md. Menhazul Abedin**, **Benojir Ahammed**, and **Mohammad Ali**, Assistant Professor, Statistics Discipline, Khulna University, Khulna-9208, Bangladesh for their cordial help and suggestion in different stages of my research. I am thankful to all employees of this Department for their various kinds of help during this research work. Thanks to **Naznin Haque Bithi**, my wife, who has endured this work, and provided support and sacrifice to keep me fed and on track. Without her, I would not have started this work, nor would I have finished. Finally, I would like to thank my family members especially my parents for their encouragement and supports. I would also like to thank all of my beloved friends.

Abstract

Due to the increasing prevalence of diabetes and cancer, it is an urgent need to develop automated system that helps to detect disease using one of the modern technologies. Nowadays, Machine Learning (ML)-based methods have become very popular as an automatically model building techniques. Despite of the rapid development of theories for computational intelligence, application of ML-based classifiers to diabetes and cancer diagnosis remains a challenging issue. Still these ML-based classifiers did not give a satisfactory accuracy and therfore cannot correctly classify healthcare data like diabetes and cancer patients. Because most of the diabetes and cancer dataset are complex in nature and contains missing values, unusual observations, multi-collinearity problems and so on. In most of the existing research, the researcher did not use feature selection (FS) techniques to identify the risk factors of cancer and diabetes disease. They applied limited classifiers to classify and predict the diabetes and cancer status but they did not tune the hyper parameter of the classifiers, as a result, their accuracy and AUC were low. Thus, an attempt has been made in this study to increase the accuracy of the classifiers in diabetes and cancer data by considering the above factors in ML-based algorithm. The main objective of this study is to comparison the performances of ML-based methods in healthcare data and suggests the best model with better performance compared to the models published in the existing research.

To fulfill the objectives, we have used 3 healthcare datasets, among them two datasets on diabetes (NHANES and Pima Indian), and another one on colon cancer. The NHANES diabetes data has been extracted from 2009-12 National Health and Nutrition Examination Survey having 14 factors and 6561 respondents with 10.01% diabetes whereas, the Pima diabetes data, extracted from UCI Repository having 768 female respondents with 34.89%~35% diabetics patients. Another one on colon cancer dataset has been extracted from Kent ridge biomedical data repository, comprising of 2000 genes and 62 patients with 40 cancerous patients. Furthermore, different feature selection (FS) methods, namely logistic regression (LR), random forest (RF), t-test, Kruskal-Wallis (KW), and so on have been implemented to extract high-risk significant factors and also implemented different ML-based classifiers like *n*aïve Bayes (NB), decision tree (DT), Adaboost (AB), and random forest (RF), and so on for prediction and classification. The performances of these ML-based classifiers were assessed by accuracy (ACC), sensitivity (SE), specificity (SP), positive predictive value (PPV),negative predictive value (NPV), F-measure (FM), and area under the curve (AUC).

In case of 2009-12 NHANES diabetes data, risk factors were extracted using LR and implemented 4 ML-based classifiers as NB, DT, AB, and RF were implemented to predict diabetic disease. LR result demonstrates that out of 14 factors, 7 risk factors were extracted for diabetes. The highest ACC of 90.62% and AUC of 0.95 for K10 were obtained by a ML-based system combining of LR-based FS and RF-based classifier. It was confirmed that LR-RF based combination performs better for prediction of diabetes compared to others.

In case of Pima diabetes data, we proposed a robust ML-based system for predicting diabetes. ML-based system was designed, optimized, and evaluated, where (i) the features were extracted and optimized from 6 FS techniques (RF, LR, MI, PCA, ANOVA, and FDR), and combined with 10 classifiers (LDA, QDA, NB, GPC, SVM, ANN, AB, LR, DT, and RF) under the assumptions that both missing values and outliers were replaced by group median and median will be improved the ACC. Our findings demonstrate that missing values were replaced with a group median and outliers with a median values, and further using the combination of RF-RF-based system yields ACC, SE, SP, PPV, NPV, and AUC as 92.26%, 95.96%, 79.72%, 91.14%, 91.20% and 0.93, respectively. It was also confirmed that RF-based classifier can be accurately classifies of diabetes patients and performs better ACC in case of both presence and absence of outliers.

Another contribution of this research was the application of different ML-based classifiers on colon cancer microarray gene expression data. In case of cancer patients, a comparative study of different ML-based systems were presented with two major objectives as (i) identification of high-risk differential genes using 4 statistical tests and (ii) propose an ML-based classifier for predicting cancer patients. Four statistical tests as WCSRS, t-test, KW, and F-test are adapted for the identification of differential expressed genes using p-values and then 10 classifiers like LDA, QDA, NB, GPC, SVM, ANN, LR, DT, AB, and RF for were also implemented to classify cancer patients. The overall average ACC of 90.50% was supported by a ML-based system combining with 4 statistical tests and 10 classifiers. It was confirmed that the combination of WCSRS test and RF-based classifier yields the highest ACC (99.81%). Finally, we may say that RF-based classifier performed better for accurate detection and prediction of both diabetes and cancer disease in multi-center clinical trials. It will be very helpful for physicians and doctors to make a decision to early diagnosis of diabetes and cancer disease.

List of Publication from this Thesis

From Peer Reviewed Journal

- **Maniruzzaman, M**., **Rahman, M.J.,** Ahammed, B., Abedin, M.M., (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems (**Springer***) (**SJR Q1**). 8:7. https://doi.org/10.1007/s13755-019-0095-z.

- **Maniruzzaman, M., Rahman, M. J.,** Ahammed, B., Abedin, M. M., Suri, H. S., Biswas, M., & Suri, J. S. (2019). Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Computer Methods and Programs in Biomedicine (**Elsevier**) (**SJR Q1, IF: 3.632**)*. 176, 173-193. https:doi.org/10.1016/j.cmpb.2019.04.008.

- **Maniruzzaman, M**., **Rahman, M. J.,** Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate Diabetes Risk Stratification using Machine Learning: Role of Missing Value and Outliers. *Journal of Medical Systems (**Springer**) (**SJR Q2, IF: 3.058**)*. 42(5), 92. https://doi.org/10.1007/s10916-018-0940-7.

Book Chapter

- **Maniruzzaman, M., Rahman, M. J.,** Ahammed, B., Abedin, M. M., Suri, H. S., Biswas, M., & Suri, J. S. (2020). Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. In Cognitive Informatics, Computer Modelling, and Cognitive Science (pp. 273-317). Academic Press. https://doi.org/10.1016/B978-0-12-819443-0.00014-3.

From Conference Proceedings

- **Maniruzzaman, M., Rahman, M. J**., Ahammed, B., Abedin, M. M. (2019). Logistic Regression based Feature Selection and Classification of Diabetes Disease using Machine Learning Paradigm. Proceedings of International Conference Data Science and SDGs: Challenges, Opportunities and Realities, 18-19 December, 2019, organized by Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh.

List of Tables

List of Figures

List of Abbreviations

DM	Diabetes Mellitus
PID	Pima Indian Diabetes
UCI	University of California Irvine
FS	Feature Selection
RF	Random Forest
LR	Logistic Regression
MI	Mutual Information
PCA	Principal Component Analysis
ANOVA	Analysis of Variance
FDR	Fisher's Discriminant Analysis
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
NB	Naïve Bayes
GPC	Gaussian Process Classification
SVM	Support Vector Machine
ANN	Artificial Neural Network
DT	Decision Tree
AB	Adaptive Boosting
BN	Bayes Net
PLS	Partial Least Square
SD	Standard Deviation
FFNN	Feed Forward Neural Networks
BPA	Back-Propagation
NA	Not Applicable
MLP	Multilayer Perceptron
AIS	Artificial Immune Systems
HM	Hierarchical Multi-Level Classifiers
BagMoov	Bagging with Multi-Objective Optimized Voting
JK	Jackknife
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ACC	Accuracy
SE	Sensitivity
SP	Specificity
PPV	Positive Predictive Value
NPV	Negative Predictive Value
ROC	Receiver Operating Curves
AUC	Area Under the Curve
RI	Reliability Index
FM	F-measure
MAE	Mean Absolute Error
RMSE	Root Mean Square Error

IQR	Inter-Quartile Range
MLE	Maximum Likelihood Estimator
PIM	Permutation Importance Index
GIM	Gini Importance Measure
OOB	Out of Bag
ORT	Outlier Removal Techniques
KNN	K-Nearest Neighborhood
CFS	Correlation Based Feature Selection
CI	Confidence Interval
ML	Machine Learning
EM	Expectation Maximization
FKM	Fuzzy K-means
OR	Odds Ratio
NHANES	National Health and Nutrition Examination Survey
PT	Protocol Types
mRMR	Minimum Redundancy Maximum Relevance
GB	Gradient Boosting
p-value	Probability Value
DS	Data Size
DL	Deep Learning
SVD	Singular Value Decomposition
KW	Kruskal-Wallis
WCSRS	Wilcoxon Sign Rank Sum
CT	Classifier Types
Poly-2	Polynomial Kernel with Order Two
RBF	Radial Basis Function
GSA	Gravitational Search Algorithm
BCGA	Binary Coded Genetic Algorithm
RCGA	Real Coded Genetic Algorithm
PSO	Particle Swarm Optimization
PPSO	Pure Particle Swarm Optimization
HPSOT	Hybrid Particle Swarm Optimization
PNN	Probabilistic Neural Network
PTS	Pure Tabu Search

List of Symbols

$\mathbf{X}$	Data Matrix
$\boldsymbol{\mu}$	Overall Mean Vectors
$\mathbf{I}$	Identity Vectors of 1
$\mathbf{S}$	Sample Variance-covariance Matrix
λ_j	j^{th} Eigen Values
e_j	j^{th} Eigen Vectors
μ_j	j^{th} Sample Mean Vectors
$\mathbf{S}_j^2$	j^{th} Sample Variance
p-value	Probability Values
$\mathbf{S}_w$	Within Scatter Matrix
$\mathbf{S}_B$	Between Scatter Matrix
P(X,Y)	Joint Probability Distribution of X and Y
P(X)	Marginal Probability Distribution of X
P(Y)	Marginal Probability Distribution of Y
P(y|x)	Conditional Probability Distribution of y given x
P(x|y)	Conditional Probability Distribution of x given y
ξ_n	Reliability Index
σ_n	Standard Deviation
$\bar{\mathbf{X}}_\mathbf{i}$	Mean Vectors of i^{th} Groups
$\bar{\mathbf{X}}_\mathbf{i}^\mathbf{T}$	Transpose of the Mean Vector of i^{th} Group
$\mathbf{K}(\mathbf{x}, \mathbf{x}^\mathbf{T})$	Kernel Matrix
$\bar{\boldsymbol{\mu}}_\mathbf{m}(\mathbf{c})$	Overall Mean Accuracy of the Classifier
O1	Data Contains Outliers
O2	Impute Outliers by Median
K2	Two-fold Cross-Validation
K4	Four-fold Cross-Validation
K5	Five-fold Cross-Validation
K10	Ten-fold Cross-Validation

Chapter 1 General Introduction

1.1 Introduction

Healthcare is a broad concept that refers to improve the health of patients and other neurodevelopmental disorders. It is provided by doctors and allied areas of health. Data on healthcare is any form of data relating to health conditions, causes of mortality, and wellbeing (Tzourakis & Melissa, 1996). There are different types of healthcare data including medical (cancer, diabetes, strokes, etc.), omics (genomics, transcriptomics, and proteomics), and sensor data (wearable and wireless devices), which are easily handled using different classical and machine learning (ML) models (Koh and Tan, 2011). With the advent of ML technique, it has been tested all over the world and accumulated study data suggested that ML technique for classification of healthcare data is superior performance to classical method. In recent times, ML-based systems have gained popularity like never before. This field is expected to grow exponentially in the coming years. ML is a branch of study in which a model can be learned automatically from the experiences based on data without exclusively being modeled like statisti cal models. Over a period and with more data, model predictions will become better. Although recently developed ML-based methods have attracted increasing attentions and substantial amount of research works have been carried out, the research for improving the accuracy of classification models has never been stopped (Zhang, 2003). To facilitate an adequate level of accuracy, the developer has to be responsive to the characteristics of different methods, and determine if a particular method is appropriate for the defined situation before embarking its usage in real application. As a result, the choice of a ML-based model is one of the crucial factors that will affect the classification accuracy. Researchers have made a great deal of effort over many years to develop effective models to improve classification accuracy. As a result, various important ML-based classification models have been evolved in literatures.

Recently, the application of ML-based system on many domains, including healthcare datasets has received great attention day by day. ML-based systems have also dominated in the field of medical healthcare (Srivastava et al., 2018; Shakeel et al., 2018; Bauder et al., 2018; Shah et al., 2019), mental-health (Grunerbl et al., 2014; Osmani et al., 2013), human behaviours

(Ertin et al., 2011; Muaremi et al., 2013), Time series (Zhang, 2003; Siami et al., 2018) and medical imaging (Deniz et al., 2018; Ashour et al., 2018; Banchhor et al., 2018). ML-based systems is mainly used for early diagnosis and prediction or detection of different cardiovascular disease as diabetes, cancer, heart disease, liver, and so on to improve the classification accuracy (Maniruzzaman, et al., 2017; Srivastava et al., 2018; Shakeel et al., 2018; Bauder et al., 2018; Shah et al., 2019). Moreover, ML-based system is also used for prediction and progonosis of cancer (Cruz and Wishart, 2006; Kourou et al., 2015).

Despite of the rapid development of theories for computational intelligence, application of ML-based classifiers to diabetes and cancer diagnosis remains a challenging issue. Still these ML-based classiiers did not give a satisfactory accuracy and therfore cannot correctly classify healthcare data like diabetes and cancer patients. Because most of the diabetes and cancer dataset are complex in nature and contains missing values, unusual observations, multi-collinearity problems and so on (Maniruzzaman et al., 2017; Lee and Yoon, 2017). In most of the existing research, the researcher did not use feature selection (FS) techniques to identify the risk factors of cancer and diabetes disease. They applied limited classifiers to classify and predict the diabetes and cancer status but they did not tune the hyper parameter of the classifiers, as a result, their accuracy and AUC were low (Maniruzzaman et al., 2017; Lee and Yoon, 2017). Thus, an attempt has been made in this study to increase the accuracy of the classifiers in diabetes and cancer data by considering the above factors in ML-based algorithm. The main objective of this study is to comparison the performances of ML-based methods in healthcare data and suggests a best model with better performance compared to the models published in the existing research.

In this study, we have chosen two types of healthcare data namely; diabetes and cancer due to the prevalence of diabetes and cancers are increasing worldwide. Cancer and diabetes are the 2^{nd} and 12^{th} leading causes of death globally (Lopez et al., 2006). Globally, 108 million people were affected by diabetes in 1980, exceeding 285 million in 2009, 366 million in 2011, 382 million in 2013, 422 million in 2014, 425 million in 2017, and 463 million in 2019. This amount will be expected to exceed 578 and 700 million in 2030 and in 2045 (Saeedi et al., 2019). Additionally, about 1.6 million individuals died directly from diabetes (Bharath et al., 2017). Different forms of cancer disorder, such as colon, lung, breast, prostate, and so on are found in human bodies. Approximately, 12.4 million new cases were diagnosed by cancer worldwide in

2008 (Giovannucci et al., 2010), 18.1 million in 2018 (Bray et al., 2018), and this figure extended to 19.30 million in 2020 (Ferlay et al., 2020). Moreover, 9.6 million people died of cancer in 2018 and this figure was reached about 9.9 million in 2020. Among them, 18% of deaths were occurred due to lung cancer, 9.4% death for colorectal, 8.3% for liver, 7.7% for stomach, 4.7% for pancreas, and 3.8% for prostate. Therefore, it is need to diagnosis and prognosis of cancer and diabetes disease using detection and prediction models.

This study aims at building several predictive models using the risk factors of diabetes and cancer through ML-based approaches. Risk factors are selected by different feature selection (FS) methods. FS is one of the most important preprocessing steps in the field of ML. FS can avoid the irrelevant features (risk factors) but automatically select a subset of relevant features for building effective predictive learning model. In this study, we have used different FS methods such as t-test, F-tests, Kruskal-Wallis (KW) test, logistic regression (LR), principal component analysis (PCA), Fisher discriminant analysis (FDR) and so on. Furthermore, different classifiers like linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machine (SVM), random forest (RF), logistic regression (LR) have been conducted on significant risk factors for classification of diabetes and cancer patients. This study proposes a ML-based systems combining a FS method and a classifier would yield the highest accuracy compared to existing research.

1.2 Research Gap

In the existing research, several studies in the literature had been, conducted in the diabetes and cancer datasets. In most of the papers, they used or did not utilize any FS-based methods to identify the risk factors of cancer/diabetes disease. They applied limited classifiers to classify and predict the diabetes/cancer status but they did not tune the hyper parameter of the classifiers, as a result, their accuracy and AUC are low.

1.3 Objective of the Study

The main objective of this study is to compare the performance of ML-based techniques in healthcare data and suggests the best model with the better performance compared to the models published in existing research. The specific objectives of the study are

- To identify the high-risk factors of healthcare data as diabetes and cancer.
- To classify and predict the healthcare data as diabetes and cancer using different ML-based classifiers.
- To compare the performance of the ML-based techniques on the healthcare data as diabetes and cancer.

1.4 Layout of the Study

We have organized this thesis paper in **seven Chapters**. The **first Chapter** is general introduction, including introduction of machine learning in healthcare data, research gap, objective of this study and layout of the study. **Chapter 2** represents the literature review. **Chapter 3** represents methodology, including introduction, feature selection, and machine learning techniques, data partitioning, model based performance evaluations are discussed. **Chapter 4** discussed about the classification and prediction of diabetes disease using machine learning paradigm. Accurate risk prediction of diabetes using machine learning: role of missing value and outliers are discussed in **Chapter 5**. The statistical characterization and classification of colon cancer using machine learning paradigms is discussed in **Chapter 6**. Finally, the conclusion and areas of further research are discussed in **Chapter 7.**

Chapter 2 Literature Review

2.1 Introduction

The application of machine learning (ML)-based techniques in healthcare datasets has been increased day by day. In this study, we have used three healthcare datasets, among them two datasets on diabetes and another one on colon cancer. Various ML-based systems like linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naïve Bayes (NB), decision tree (DT), support vector machine (SVM), artificial neural network (ANN), random forest (RF), artificial immune system (AIS), logistic regression (LR), etc. are used to classify and predict the patients in healthcare datasets. The literature on these issues is reviewed in the following two sections.

2.2 Diabetes Datasets

A large number of works in the literature related to the diagnosis and prediction of diabetes which were presented in Table 2.1. Karthikeyani et al. (2012) adapted SVM with RBF kernel on PI dataset. The PID dataset, comprised of 8 factors and 768 respondents having 268 diabetic patients. The dataset contained some meaningless value replaced these values with mean. They adopted SVM to predict diabetes patients and achieved 74.80% ACC. The same authors applied partial least square (PLS) as a feature selection (FS) method and extracted 3 factors out of 8 (Karthikeyani et al., 2013). They adapted LDA to classify and predict diabetes and obtained 74.40% ACC. Kumari and Chitra (2013) also applied SVM classifier along with RBF kernel to predict diabetes disease. They had been selected 460 observations for analysis after excluding the meaningless observations from the dataset. They took 200 observations as a training set and rests were used as test/validation set. They indicated that 75.50% ACC was obtained by SVM. A study on diabetes was conducted by Parashar et al. (2014). They selected 2 factors out of 8 using LDA and adapted two classifiers as SVM and FFNN and 75.65% ACC was achieved by SVM. Bozkurt et al. (2014) also used two ML-based systems as AIS and ANN to classify diabetic patients. It was noted that ANN gained a higher ACC of 76.00% compared to AIS. Iyer et al. (2015) applied two ML-based models as NB and DT to classify of diabetes which contained missing values, were imputed by mean. Using correlation-based FS, 2 features out of 8 were

extracted and they demonstrated that 74.79% ACC was achieved by DT. Two ML-based models as MLP and BN were applied by Kumar Dewangan and Agrawal (2015) for diabetes prediction, where, the best ACC of 81.19% was obtained by MLP. Bashir et al. (2016) presented HM-Bag-Moov-based method to classify diabetes and compared it performances along with different ML-based classifiers (i.e.,NB, SVM, LR, QDA, KNN, RF, and ANN). They demonstrated the highest ACC (77.21%) was obtained by HM-Bag-Moov-based method. A study J48-based algorithm was proposed for prediction of diabetes by Sivanesan et al. (2017) and achieved 76.58% ACC. Another study, 4 ML-based classifiers as NB, LR, J48, and RF were implemented by Nabi et al. (2017) and 80.43% ACC was obtained by LR. Maniruzzaman et al. (2017) implemented LDA, QDA, NB, and GP-based classifier to classify diabetic patients. It was noted that the largest ACC of 81.97~82.00% was given by GP-based classifier. Zou et al. (2018) conducted a study on the classification of diabetes disease. The diabetes dataset contained 14 attributes and 220680 patients (69% diabetic patients). Two FS methods (PCA vs. mRMR) were implemented for dimension reduction. They also implemented 3 prediction models (DT, NN, and RF) for prediction of diabetes and validated that the highest ACC of 80.84% was achieved by RF.Ahuja et al. (2019) propose a ML-based technique followed by the combination of LDA-based FS and a classifier out of 5 (SVM, MLP, LR, RF, and DT) provides better results. Missing values were substituted with median and they used LDA-based FS to pick the most important features. It was demonstrated that 78.70% ACC was obtained by LDA-MLP. Sisodia et al. (2018) applied SVM, NB, DT for diagnosis of diabetes and NB obtained 76.30% ACC. Yu et al. (2010) introduced SVM-based classifier on 6214 respondents (23.51% diabetic patients) while the dataset, extracted from 1999-2004 NHANES. They optimized the kernel among 4 kernels as linear, polynomial, sigmoid, and RBF using ACC. They the maximum ACC of 83.50% was given by SVM. Semerdjian et al. (2017) extracted 5515 individuals from the 1999-2004 NHANES diabetes dataset. The most significant diabetic risk features were extracted using RF. Furthermore, 5 prediction models based classifiers as LR, KNN, RF, GB, and RF were adopted and GB-based model gave AUC of 0.84. Mohapatra et al. (2019) showed that MLP-based classifier can correctly classify (77.50% ACC) diabetic patients. Pei et al. (2019) indicated that 94.20% ACC was obtained by DT.

Table 2.1. Previous studies of diabetes dataset in literature.

Authors	Year	Dataset	DS	FS method	Classifiers	ACC (%)
Karthikeyani et al.	2012	PID	768	NA	**SVM**	74.80
Karthikeyani et al.	2013	PID	768	PLS	**LDA**	74.40
Kumari and Chitra	2013	PID	460	NA	**SVM**	75.50
Parashar et al.	2014	PID	768	LDA	**SVM**, FFNN	75.65
Bozkurt et al.	2014	PID	768	NA	AIS, ANN	76.00
Iyer et al.	2015	PID	768	CFS	**DT**, NB	74.79
Kumar et al.	2015	PID	768	None	**MLP**, BN	81.19
Bashir et al.	2016	PID	768	NA	NB, SVM, LR, QDA, KNN, RF, ANN, **HM-Bag Moov**	77.21
Sivanesan et al.	2017	PID	768	NA	J48	76.58
Meraj Nabi et al.	2017	PID	768	NA	NB, **LR**, J48, RF	80.43
Maniruzzaman et al.	2017	PID	768	NA	LDA, QDA, NB, **GPC**	81.97
Zou et al.	2018	Diabetes	220680	PCA, mRMR	DT, NN, RF	80.84
Ahuja et al.	2019	PID	768	LDA	SVM, **MLP** LR, DT, RF	78.70
Sisodia et al.	2018	PID	768	NA	SVM, **NB**, DT	76.30
Yu et al.	2010	NHANES	6314	NA	**SVM**	83.50
Semerdjian et al.	2017	NHANES	5515	RF	LR, KNN, RF, **GB**, SVM	AUC:0.84
Mohapatra et al.	2019	PID	768	NA	**MLP**	77.50
Pei et al.	2019	Diabetes	10436	CS	**DT**	94.20

DS: Data size; ACC: Accuracy; AUC: Area under the curve; LDA: Linear discriminant analysis; NA: Not Available; CFS: Correlation feature selection; PCA: Principal component analysis; mRMR: Minimum redundancy maximum relevance; RF: Random forest; CS: Chi-square; SVM: Support vector machine; FFNN: Feed-forward neural network; AIS: Artificial immune system; DT: Decision tree; NN: Neural network; MLP: multilayer perceptron; LR: Logistic regression; NB: Naïve Bayes; BN: Bayes net; QDA: Quadratic discriminant analysis; KNN: K-nearest neighborhood; HM-Bag Moov: Hierarchical Multi-level classifiers bagging with Multi-objective optimized Voting; GPC: Gaussian process classification; GB: Gradient boosting; PID: Pima Indian diabetes; NHANES: National Health and Nutrition Examination Survey. Bold value indicates proposed method results

2.3 Colon Cancer Dataset

A lot of work had been done in the previous studies of cancer data in the literature that are showed in Table 2.2. Four sets of algorithms (GSA, BCGA, RCGA, and PSO) were applied for the prediction of cancer patients by Kumar et al. (2012), where, the cancer dataset extracted from PubMed and comprised of 2000 biomarkers and 62 respondents having 40 cancer patients. They confirmed that the maximum ACC of 58.70% was achieved by GSA. Shen et al. (2008) implemented 3 classifiers like HPSOTS, PTS, and PPSO for the classification of cancer patients. The largest ACC of 89.55% was obtained by HPSOTS-based model. Alladi et al. (2008) extracted top10 biomarkers out of 2,000 using t-test (p- value<0.05). About 80% dataset was taken as a training set and remaining of the dataset as a test set. 3 sets of classifiers (LR, NN, and SVM) were also adopted and 85.80% ACC was obtained by SVM. Vanitha et al. (2015) extracted 3 significant biomarkers/genes using MI and also implemented 3 sets of classifiers (k-NN, NN, and SVM) for cancer risk prediction and the highest ACC of 74.19% was achieved by SVM. Sun et al. (2006) implemented DWT to extract biomarker; they also implemented PNN for cancer risk prediction and obtained 92.00% ACC. Chen et al. (2007) proposed a MK-SVM

method as a FS and classifier. They adopted MK-SVM to cancer risk prediction, where, 11 differential expressed genes were extracted by MK-SVM and obtained 93.50% ACC. Three types of pathway biomarker datasets KEGG, Biocarta, and Reactome was used by Liu and Gao (2018). The risk biomarker had been identified based on entropy (p-values<0.05) and they extracted 190 biomarkers out of 205 for KCGG 197 out of 2017 for Biocarta, and 644 out of 1068 for Reactome datasets. They adopted SVM for biomarker risk prediction of diabetes and obtained the largest AUC of 96.00.

Table 2.2. Previous studies of cancer dataset in the literature.

Authors	Year	DS	FS method	Classifier	Protocols	ACC (%)
Kumar et al.	2010	62	NA	**GSA**,BCGA, RCGA, PSO	JK	58.70
Shen et al.	2008	62	NA	**HPSOTS**, PTS, PPSO	NA	89.55
Alladi et al.	2015	62	t-test	LR, NN, **SVM**	K10	85.80
Vanitha et al.	2015	62	MI	KNN, NN, **SVM**	JK	74.19
Sun et al.	2006	62	DWT	**PNN**	JK	92.00
Chen et al.	2007	62	MK-SVM	**MK-SVM**	JK	93.50
Liu and Gao	2018	--	Entropy	**SVM**	JK	96.00

DS: Data size; FS: Feature selection; MI: Mutual information; DWT: Discrete wavelet transformation; MK-SVM: Multiple kernel support vector machine; GSA: Gravitational search algorithm; BCGA: Binary coded genetic algorithm; RCGA: Real coded genetic algorithm; PSO: particle swarm optimization (PSO); HPSOTS: Hybrid particle swarm optimization; PTS: Pure tabu search; PPSO: Pure particle swarm optimization; PNN: Probabilistic neural network. Bold value indicates the proposed method results.

Chapter 3 Methodology

3.1 Introduction

In this chapter, we discussed different feature selection (FS) techniques, machine learning (ML)-based techniques, data partitioning, and model performance evaluations. There are different FS methods as PCA, FDA, ANOVA, MI, LR, RF and different statistical tests (t-test, WCSRS test, F-test) are used to identify the relevant factors as well as different ML-based techniques as: LDA, QDA, NB, AB, DT, RF, ANN, SVM, GPC. These are briefly discussed as follows:

3.2 Feature Selection Techniques

3.2.1 Principal Component Analysis

PCA is one of the popular dimension reduction technique in ML/statistics (Shrivastava et al., 2016). The calculation steps of pooling-based PCA algorithm are given below:

1. Compute the mean vectors for every features as follows:

$$\boldsymbol{\mu}_{(K\times 1)}=\frac{1}{N}\mathbf{X}^{\mathbf{T}}\mathbf{I} \tag{3.1}$$

   Here, **X** is a data matrix with dimensions N×K; where, N is total no. of observations, K is the total no. of factors, and **I** is identity vector of 1's of size N×1.

2. Subtract the mean vectors from data matrix to make normalize the data as

$$\mathbf{B}_{(N\times K)}=\mathbf{X}-\boldsymbol{\mu} \tag{3.2}$$

3. Compute the variance-covariance matrix using the formula

$$\mathbf{Q}_{(K\times K)}=\frac{1}{N}\mathbf{B}^{\mathbf{T}}\mathbf{B} \tag{3.3}$$

4. Compute the eigenvalues $(\lambda_1,\lambda_2,\ldots,\lambda_P)$ and eigenvectors $(e_1, e_2,\ldots,e_P)$ from the variance variance-covariance matrix (**Q**).
5. Rank the eigenvalues from the largest to smallest and also rank the corresponding eigenvectors in the same order.
6. Take the number of PC's (r) that satisfies the following criterion:

$$\frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{k}\lambda i}>G \tag{3.4}$$

   Where,G is the cutoff point varying from 0.90 to 0.99.

7. Calculate the contribution of each factors based on the following formulae

$$c_n = \sum_{z=1}^{r} |e_{hn}| \; ; \qquad n=1,2,\ldots,P \tag{3.5}$$

Where,e_{hn} indicates the n^{th} entry of e_n which is the h^{th} eigenvectors, n=1, 2,…,P. Select the factors after sorting c_n from the largest to smallest that will be given the significant factors (r) (without modifying original factors).

3.2.2 Analysis of Variance

The aim of ANOVA test is perform test whether or not all the different classes of **Y** have the same mean as **X**. (Elssied et al., 2014). ANOVA test is conducted based on the following notations:

N_j=Number of classes with **Y=j**.

$\boldsymbol{\mu}_j$= The sample mean of the predictors **X** for the target variables **Y**=j.

$\mathbf{S}_j^2$=The sample variance of the predictors **X** for the target variables **Y**=j:

$$\mathbf{S}_j^2 = \frac{\sum_{j=}^{N_j} \left(\mathbf{X}_{ij} - \boldsymbol{\mu}_j\right)^2}{N_j - 1} \tag{3.6}$$

$\boldsymbol{\mu}$= The overall mean of the predictors **X**: $\boldsymbol{\mu} = \frac{\sum_{j=1}^{N} N_j X_j}{N}$. The test statistic is

$$F = \frac{\frac{\sum_{j=1}^{J} N_j \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}\right)^2}{(J-1)}}{\frac{\sum_{j=1}^{J} (N_j - 1)\mathbf{S}_j^{\,2}}{(N-1)}} \tag{3.7}$$

Which follows F-distribution with $(J-1)$and $(N-1)$ degrees of freedom respectively.

The p-value is calculated based on the F-statistic as follows: Prob. {F (J-1, N-1)>F}

3.2.3 Fisher's Discriminant Ratio

Fisher discriminant ratio (FDR) is used to extract the most informative factors in that way which have high within-class distance and small between-class distance (Shrivastava et al., 2017). The general calculation steps of FDR are given as:

1. Calculate the sample mean vectors $\boldsymbol{\mu}_j$ for different classes:

$$\boldsymbol{\mu}_j = \frac{1}{N_j} \sum_{X \in D_j}^{N} \mathbf{X}_k \quad ; j=1, 2. \tag{3.8}$$

2. Compute the within class scatter matrix $\mathbf{S}_w$ by the following formula:

$$\mathbf{S}_w=\sum_{j=1}^{K}\mathbf{S}_j\,,\quad \text{where, } \mathbf{S}_j=\sum_{X\in D_j}\left(\mathbf{X}-\boldsymbol{\mu}_j\right)\left(\mathbf{X}-\boldsymbol{\mu}_j\right)^T \tag{3.9}$$

3. Also compute the between-class scatter matrix $\mathbf{S}_B$ as follows:

$$\mathbf{S}_B=\sum_{j=1}^{K}N_j(\boldsymbol{\mu}_j-\boldsymbol{\mu})(\boldsymbol{\mu}_j-\boldsymbol{\mu})^T \tag{3.10}$$

Where, $\boldsymbol{\mu}$ is the overall mean vector, $\boldsymbol{\mu}_j$ is the j^{th} sample mean vectors and N_j is total no. of classes.

4. Finally, the FDR is calculated as follows:

$$\text{FDR}=\mathbf{S}_W^{-1}\mathbf{S}_B \tag{3.11}$$

5. Compute the eigenvalues $(\lambda_1,\lambda_2,\ldots,\lambda_P)$ and the corresponding eigenvectors $(e_1,e_2,\ldots,e_P)$ for the scatter matrices of FDR.
6. Rank the eigenvectors from the largest to lowest and choose the K eigenvectors with the the largest eigenvalues to form a P× K dimensional weighted data matrix W.
7. Use this P× K eigenvector matrix to transform the samples into the new subspace as:

$$Y=XW \tag{3.12}$$

Where, **X** is a N× P -dimensional matrix representing the N samples, and Y is the N× K dimensional samples in the new spaces.

3.2.4 Mutual Information

Mutual information (MI) is also used to detect a subset of the most informative factors/features (Peng et al., 2005). It can easily the handle the over-fitting problems and detect the dominant factors based on MI. For two discrete variables x and y, MI (x, y) is defined as

$$\text{MI }(x,y)=\sum_{i,j}p\left(x_i,y_j\right)\log\frac{p\left(x_i,y_j\right)}{p(x_i)p\left(y_j\right)} \tag{3.13}$$

Where, p (x, y) is the joint pdf of x and y; p(x)and p(y)are the marginal pdf of x and y.

3.2.5 Logistic Regression

Logistic regression (LR) is a supervised learning that is adapted to forecast the probability of dichotomous response variable (1/0) based on discrete/continuous predictors. If Y is a dichotomous response variable (1/0) and X is the set of all predictors, then their linear combinations can be written as:

$$\text{logit}\,(P_j) = \log_e\left(\frac{P_j}{1-P_j}\right) = \sum_{i=0}^{K} \beta_i X_i \qquad (3.14)$$

Where, P_j is defined as the probability for Y=1 (yes) and $1-P_j$ (no)when Y=0 . $\beta_I^{'}s$ (i=0,1,..,K) are the unknown parameters is the total no. of predictors and X_0=1. We estimate the parameters based on MLE and take exponent of parameters to get odds ratio (OR). One can easily test the parameters /OR's based on normal test and detects the corresponding factors whose p-values are less than 0.05. After estimating the regression parameters, one can easily calculate probability of outcome variable (1/0) based on predictors and select the cutoff of point values that yields the highest classification accuracy. If the calculated probability value is superior to the cutoff of point, it goes to one class (yes) and vice-versa (Tabaei and Herman, 2002).

3.2.6 Random Forest

Random forest (RF) implement as a FS while the rules of classification are formulated. Two importance measurements methods are used for variable selection as (i) Gini importance index (GIM), and (ii) permutation importance index (PIM) (Hasan et al., 2016). PIM index is used to order the features while RF selects the best combination of features for classification (Hasan et al., 2013). These reduced features are used for classification.

3.2.7 Wilcoxon Sign Rank Sum Test

Wilcoxon sign rank sum (WCSRS) test is used to decide whether two population mean ranks are differed or not. Let x_{1i} and x_{2i} (i=1, 2, 3,…, n) be a set of two measurements. In 1st step, compute the absolute difference between two measures ($|x_{2i}-x_{1i}|$) and calculate their sign as $\text{sgn}|x_{2i}-x_{1i}|$. If $|x_{2i}-x_{1i}|=0$, then we are excluded this pairs from the final analysis and the sample size is reduced (n_r).

The, the pair is ordered in ascending order of absolute difference. The test statistic is

$$W=\sum_{i=1}^{N_r}\left[sgn\left|x_{2i}-x_{1i}\right|.\ R_i\right] \tag{3.15}$$

3.2.8 t-test

The t-test is used to show the difference between two group's means (disease vs. control) while the data follows normal populations with unknown variance. The test statistics is

$$t=\frac{\left|\boldsymbol{\mu_{1i}}-\boldsymbol{\mu_{2i}}\right|}{\sqrt{\frac{s_{1i}^2}{n_1}}+\sqrt{\frac{s_{2i}^2}{n_2}}},\ i=1,\ 2,\ 3,\ldots,\ k. \tag{3.16}$$

Where, $\boldsymbol{\mu_{1i}}$ and $\boldsymbol{\mu_{2i}}$ are the mean of the two groups: disease vs. control; s_{1i}^2 and s_{2i}^2 are the variance of disease and control class; n_1:total no. of disease and n_2: total no. of control groups. The above Eq. (3.16) follows t- distribution with (n_1+n_2-2) df.

3.2.9 Kruskal-Wallis Test

Kruskal-wallis (KW) test is used when data violates the normality assumptions (Sawilowsky, 1990; Nahm, 2016). If n_1 and n_2 are the no. of the two groups as disease vs. control; R_1: Sum of the ranks of disease and R_2 are the sum of the ranks of control. Then the test statistic is

$$H=\frac{12}{n(n+1)}\left[\frac{R_1^2}{n_1}+\frac{R_2^2}{n_2}\right]-3(n+1) \tag{3.17}$$

This follows chi-square distribution with 1 degree of freedom.

3.3 Machine Learning Techniques

3.3.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a generalization of Fisher LDA that is used in ML or St atistics (Fisher, 1936). It is used to classify the objects into two/more than two classes which have common variance-covariance matrix ($\sum$) (Jain et al., 2000). The aimed of LDA is to classify in such a way that maximizes between classes and minimizes within classes. The mathematical formula of LDA is written as:

$$\mathbf{X}^T\sum^{-1}(\bar{\mathbf{X}_2}-\bar{\mathbf{X}_1})-\frac{1}{2}(\bar{\mathbf{X}_2}+\bar{\mathbf{X}_1})^T\sum^{-1}(\bar{\mathbf{X}_2}-\bar{\mathbf{X}_1})>d \tag{3.18}$$

Where, **X** is data matrix, $\bar{\mathbf{X}}_2$ and $\bar{\mathbf{X}}_1$ is the mean vectors for two groups (disease vs. control), and d is the cutoff of point of decision boundary. The value of d may be zero, >0 or <0. If the value of d=0, classes are similar and If d>0, then it is classified objects into disease and vice-versa.

3.3.2 Quadratic Discriminant Analysis

QDA is an extension of LDA that is used to classify the objects into two/more classes by quadratic that does not assume equal variance-covariance matrices amongst the groups (Jain et al., 2000) and mathematical formula of the QDA can be expressed as:

$$\mathbf{X}^{\mathbf{T}}\left(\sum_{\mathbf{1}}^{-1}-\sum_{\mathbf{2}}^{-1}\right)\mathbf{X}+2\left(\bar{\mathbf{X}}_{\mathbf{2}}^{\mathbf{T}}\sum_{\mathbf{2}}^{-1}-\bar{\mathbf{X}}_{\mathbf{2}}^{\mathbf{T}}\sum_{\mathbf{1}}^{-1}\right)\mathbf{X}-\left[\bar{\mathbf{X}}_{\mathbf{2}}^{\mathbf{T}}\sum_{\mathbf{2}}^{-1}\bar{\mathbf{X}}_{\mathbf{2}}-\bar{\mathbf{X}}_{\mathbf{1}}^{\mathbf{T}}\sum_{\mathbf{1}}^{-1}\bar{\mathbf{X}}_{\mathbf{1}}+\log\left(\frac{\left|\sum_{\mathbf{2}}\right|}{\left|\sum_{\mathbf{1}}\right|}\right)\right]>d \qquad (3.19)$$

Where, $\sum_1$ and $\sum_2$ are the sample covariance matrices for two groups (disease vs. control), $\bar{\mathbf{X}}_{\mathbf{1}}^{\mathbf{T}}$ and $\bar{\mathbf{X}}_{\mathbf{2}}^{\mathbf{T}}$ is the transpose of the mean vector for disease and control group.

3.3.3 Naïve Bayes

Naïve Bayes (NB)-based classifier is a family of simple probabilistic classifiers based on Bayes' theorem (Webb et al., 2005). The Bayes theorem states as follows:

$$P(y|x_1,...,x_n) = \frac{P(y)\ \pi_{i=1}^{n}P(x_i|y)}{\pi_{i=1}^{n}P(x_i)} \qquad (3.20)$$

Where, $P(y|x_i)$ and $P(x_i|y)$ is the conditional probability of y given x_i and x_i given y; $P(y)$and $P(x)$ is the marginal probability of y and x; π is the product symbol. The probability of given set of inputs $(x_i)$for class variable $(y)$and the mathematical model can be expressed as:

$$y=\text{argmax}_y\ P(y)\ \pi_{i=1}^{n}P(x_i|y) \qquad (3.21)$$

3.3.4 Adaboost

Adaboost (AB) is a ML-based technique that was known as out-of-the-box classifier, introduced by Freund and Schapire 1996 (Hu et al., 2008) and received the golden award in 2003. It is utilized to combine with various types of algorithms to improve the efficiency of the classifier. It is quite sensitive for handling noisy data, outliers, and over-fitting problems.

3.3.5 Decision Tree

Decision tree (DT) is a ML-based model, was used in both in regression and classification (Quinlan, 1986). Three steps of DT as (i)make a tree with its nodes as input features;(ii) Extract features for prediction which gives the highest information gain;(iii) repeat the above steps to form sub trees based on features which was not used in the above nodes.

3.3.6 Random Forest

Random forest (RF) is a supervised learning that was proposed by Breiman in 2001 (Breiman, 2001). The steps of RF as (i)divide the given dataset into two parts as training set and test or validation set. Create another dataset based on training set using bootstrapping method; (ii) Construct a DT for every sample and calculate the prediction result for each DT; (iii) Also compute the vote for each prediction result; (iv) Choose the highest votes that belongs to the label of classification and compute classification accuracy.

3.3.7 Artificial Neural Network

One of the popular methods in ML is artificial neural network (ANN) that is brain-inspired device designed (Shah et al., 2019). ANN comprises of input and hidden layers. The hidden layer also comprise of units that convert the input into anything that can be used by outlier layer. We have used Back-propagation algorithm for training ANN. Further, we have used various number of hidden layers; ranging from 1 to 50 for getting better performance (Shah et al., 2019).

3.3.8 Support Vector Machine

Support vector machine (SVM) is a ML-based method that was introduced by Cortes and Vapnik in 1995 (Cortes and Vapnik, 1995). Suppose that the dataset **S** consists of a sets of input observations $x_1,x_2,\ldots,x_p \in \mathbf{X}$ and class labels $y_1,y_2,\ldots,y_n \in \mathbf{y}$ associated with the observation. A separating hyper plane learn a function $f:\mathbf{X}\rightarrow\mathbf{y}$ from **S** used to predict the class label for any new observation $x\in\mathbf{X}$ by f(x)and classified as $\mathbf{y}\in\{-1,+1\}$. Then a separating hyper plane has the property that $\mathbf{W}^T\mathbf{X}+b>0$ if $y_i=1$ and $\mathbf{W}^T\mathbf{X}+b<0$ if $y_i=-1$ (Figure 3.1). SVM uses kernel trick while the data does not linearly separable for the purpose. The kernel function may be likely linear, polynomial, sigmoid, radial, and so on.

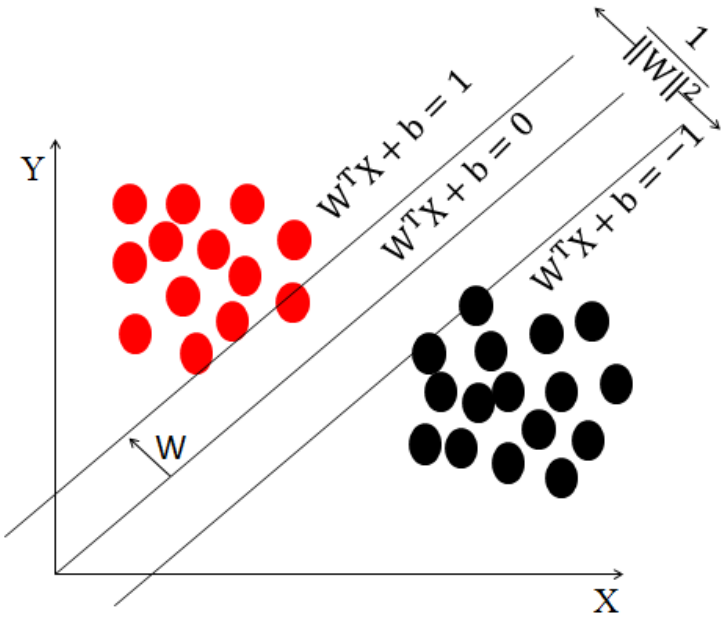


Figure 3.1. Hyper plane separating two classes.

3.3.9 Gaussian Process Classification

Gaussian process (GP) is a ML-based system that is mainly used for prediction both of classification and regression. It is a set of random variables that has a Gaussian distribution. It is also specified by mean and covariance function (Brahim-Belhouari, 2004; Rasmussen, 2004). It is mathematically defined as follows:

$$f \sim GP\ (\mathbf{m}\ (\mathbf{x}), \mathbf{K}(\mathbf{x}, \mathbf{x}^T)) \tag{3.22}$$

Where, $\mathbf{m}\ (\mathbf{x})$:the mean vector of $\mathbf{x}$ and $\mathbf{K}(\mathbf{x}, \mathbf{x}^T)$: Kernel function defined as

$$\mathbf{K}(\mathbf{x}, \mathbf{x}^T) = E\left\{\left(\mathbf{x}-\mathbf{m}(\mathbf{x})\right)\left(\mathbf{x}^T-\mathbf{m}(\mathbf{x}^T)\right)\right\}. \tag{3.23}$$

3.4 Data Partitioning

Partitioning of data is well-known as cross-validation (CV) protocol. The provided dataset divided into 2 sets as (i) training set and (ii) test set. There are lots of protocols K2, K5, K10, etc. that is utilized to minimize the variability of data. The 10-fold CV protocol is divided the given dataset into 10 equivalent parts of the provided dataset, while the 9 parts are treated as training set and remaining as a test set. The term K10 designates the total no. of CV-based protocol during ML-based system. Similarly, K4, and K5, CV-based protocols based on percentage of the training set as 75%, and 80% while remaining parts are treated as a test set.

3.5 Model Performance Evaluations

The performances of model based classifiers are evaluated based on accuracy (ACC), sensitivity (SE), specificity (SP),positive predictive value (PPV),negative predictive value (NPV). These evaluations criterion are computed using true positive (T_P), true negative (T_N), false positive (F_P) , and false negative (F_N) which are defined as follows:

Accuracy

It is the ratio of the sum of the true positive and true negative to total no. of population that can be presented as:

$$\text{ACC}\ (\%) = \left(\frac{T_P+T_N}{T_P+F_N+F_P+T_N}\right)\times 100 \tag{3.24}$$

Sensitivity

It is the ratio of the true positive condition to the predicted condition is positive that can be presented as:

$$\text{SE}\ (\%) = \left(\frac{T_P}{T_P+F_N}\right)\times 100 \tag{3.25}$$

Specificity

It is the ratio of the negative condition to the predicted condition is negative that can be presented as:

$$\text{SP}\ (\%) = \left(\frac{F_P}{F_P+T_N}\right)\times 100 \tag{3.26}$$

Positive predictive value

It is the ratio of the predicted positive condition to the true condition is positive that can be presented as:

$$\text{PPV}\ (\%) = \left(\frac{T_P}{T_P+F_P}\right)\times 100 \tag{3.27}$$

Negative predictive value

It is the ratio of the predicted negative condition to the true condition is negative that can be presented as:

$$\text{NPV}\ (\%) = \left(\frac{T_N}{F_N+T_N}\right)\times 100 \tag{3.28}$$

F-measure

It is the harmonic mean of recall and precision that can be presented as:

$$\text{FM}\ (\%) = \left(\frac{2T_P}{2T_P+F_P+F_N}\right)\times 100 \tag{3.29}$$

Chapter 4 Prediction of Diabetes Disease using Machine Learning Paradigms

4.1 Introduction

Diabetes is the 12th leading cause of death globally (Lopez et al., 2006). It is a collection of metabolic disorders that are identified by elevated blood sugar levels (Lonappan et al., 2007; American Diabetes Association, 2010; Sarwar, et al., 2010). It may lead to various complicated long-term disease as stroke, kidney failure, heat attack, blood vessels, and nerves (Nathan, 1993; Krasteva et al., 2011). Globally, 108 million people were affected by diabetes in 1980, exceeding 285 million in 2009, 366 million in 2011,382 million in 2013, 422 million in 2014 (Risk, 2016), 425 million in 2017, 463 million in 2019. This amount will be expected to exceed 578, 642, and 700 million in 2030, 2040, and in 2045 (Zimmet et al., 2016; Saeedi et al., 2019). Additionally, about 1.6 million individuals died directly from diabetes (Bharath et al., 2017). It is noted that the incidence of diabetes and subsequent death have been increased globally day by day. Three types of diabetes can be distinguished: type I (T1D), (ii) type II (T2D) and (iii) gestational diabetes (Danaei et al., 2011). T1D is usually found in young adults below 30 years of age which are connected with polyuria, thirst, chronic malnutrition, losing weight, change in vision and fatigue (Iancu et al., 2008). T2D happens in adults over the age of 45. T2D is also related to obesity, high blood pressure, dyslipidemia, arteriosclerosis, and so on (Robertson et al., 2012). Generally, pregnant women are affected by gestational diabetes.

The handling of diabetic data is a complicated task since most of the healthcare data have missing values, correlation-based structure, nonlinear, non-normal, the course of dimensionality, and complex in nature (Maniruzzaman et al., 2017). ML-based systems have dominated in the field of medical healthcare (Srivastava et al., 2018; Shakeel et al., 2018; Bauder et al., 2018; Shah et al., 2019) and medical imaging (Deniz et al., 2018; Ashour et al., 2018;Banchhor et al., 2018). Furthermore, ML-based models may be used as both FS techniques and classifiers. It helps the doctor community for correctly diabetes risk prediction. Various ML-based systems as LDA, QDA, NB, SVM, ANN, FFNN, AB, DT, J48, RF, GPC, LR, and so on have been commonly used for the identification and prediction of diabetes (Zhao et al., 2014; Bashir et al.,

2016; Maniruzzaman et al., 2017; Sisodia, 2018; Ahuja et al., 2019). The overview of the proposed ML-based framework has been presented in Figure 4.1. We believe that the combination of LR-based FS method with 4 classifiers will accurately classify of diabetes. Four applicable and relevant ML-based models as NB, DT, AB, and RF have adopted for prediction of diabetes disease. The goal of this study (**Chapter 4**) was to use LR-based FS model for the identification of risk factors of diabetes as well as propose a ML-based classifier for diabetes prediction.

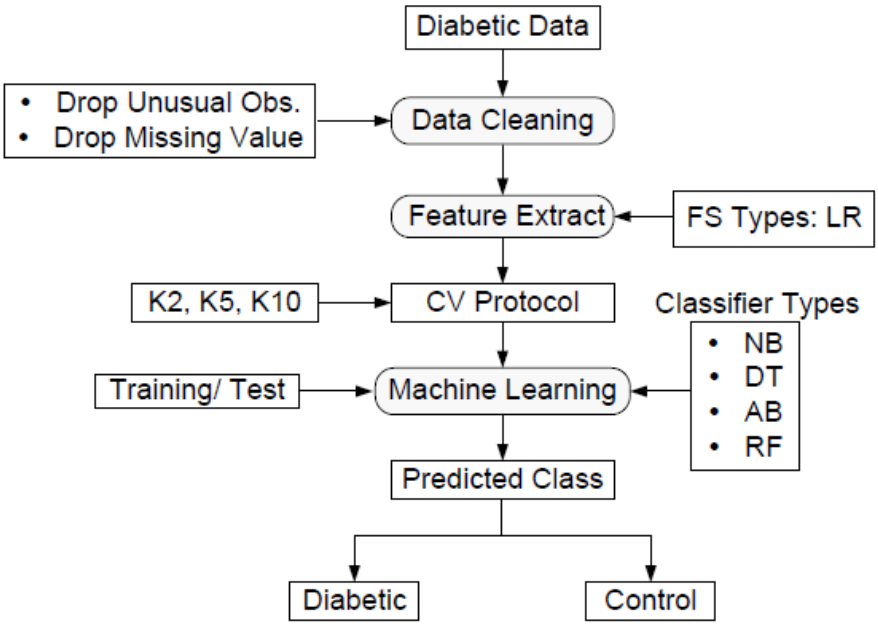


Figure 4.1. Overview of the proposed ML-based framework.

4.2 Materials and Methods

4.2.1 Dataset

The diabetes dataset, comprised of 9858 respondents having 760 diabetic respondents, derived from the 2009- 2012 National Health and Nutrition Examination Survey (NHANES), USA. The Respondents were identified as diabetic if they were met with at least one of the following criteria : plasma fasting glucose≥126mg/dl, serum glucose≥200 mg/dl; glycohemoglobin ≥6.5%. After excluding the missing values, the dataset consisted of 6561 respondents with 10% diabetic respondents.

4.2.2 Statistical Analysis

The baseline characteristics of the study population were presented as mean ± SD for continuous & number (percentages) for categorical variables, respectively. Differences in variables between diabetic patients and control were analyzed by independent paired t-test for continuous and Chi-Square test for categorical variables. The demographic and clinical characteristics of the diabetic patients have been described in Table 4.1. All of the tests were two-tailed and considered as the significant factors whose p-values were less than 0.05. Data were analyzed using Stata version 14.10 and R-i386 3.6.1. The training/test set paradigm of the entire ML-based system has been shown in Figure 4.2.

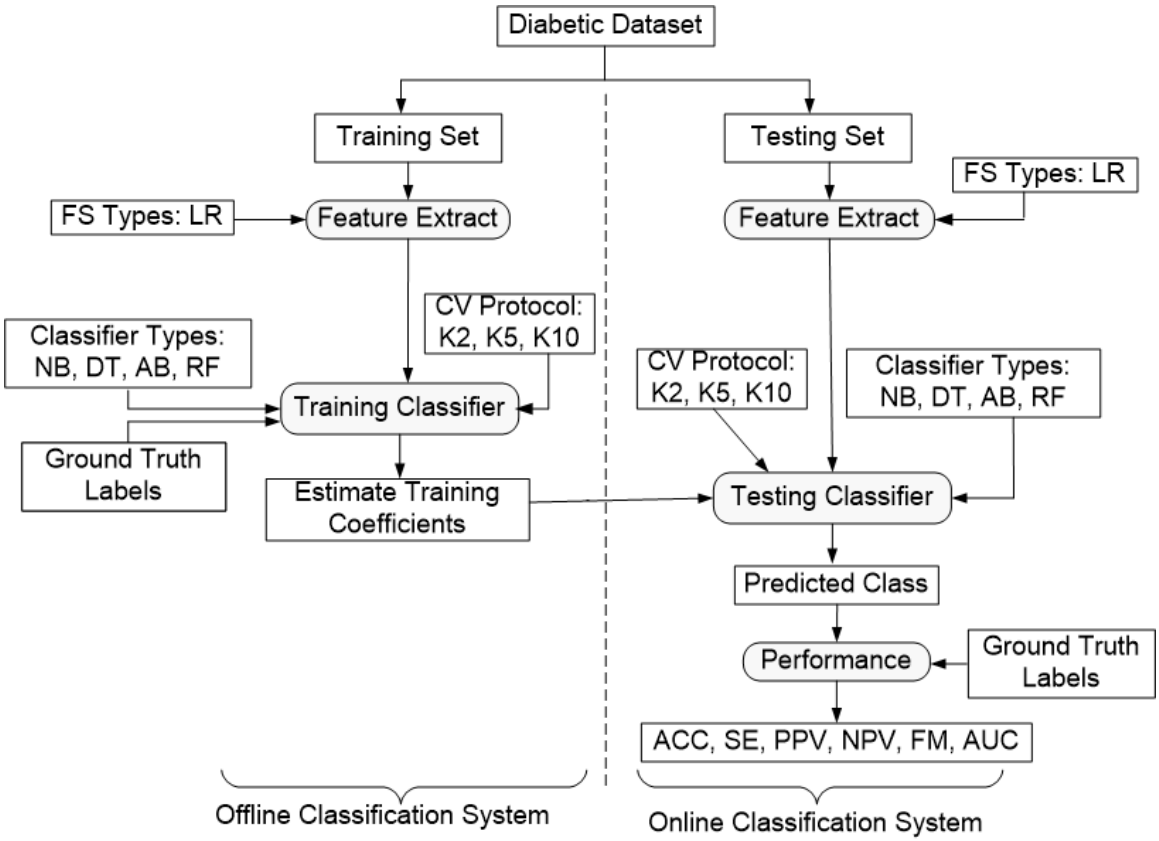


Figure 4.2. The training/testing set paradigm of the ML-based system.

4.3 Results and Discussion

4.3.1 Baseline and Demographics of the Respondents

The baseline and clinical information of patients were presented in Table 4.1. There is a total of 10% respondents out of 6561 are diabetic patients. The average ages of the respondents were 47.18 ±16.79 years. There were about 361 (55%) male diabetic patients with average age 59.85±13.22 years. It was observed that all factors were highly statistically ($p<0.001$) associated with diabetes.

Table 4.1. Baseline and clinical characteristics of respondents.

Factors	Overall (6561)	Diabetic (657)	Control (5904)	p-value[1]
Age (years)	47.18 ± 16.79	59.85± 13.22	45.77± 16.56	<0.001
Gender, male n (%)	3257 (49.64)	361 (54.95)	2896 (49.05)	<0.001
Race, white n (%)	4474 (68.19)	400 (60.88)	4074 (69.00)	<0.001
Education, college n (%)	3994 (60.87)	337 (51.29)	3657 (61. 94)	<0.001
Marital Status, married n (%)	4132 (62.98)	409 (62.25)	3723 (63.06)	<0.001
Occupation, working n (%)	4084 (62.25)	272 (41.40)	3812 (64.57)	<0.001
Weight (kg)	82.48 ± 21.24	92.40 ± 25.17	81.38 ± 20.47	<0.001
Height (m)	1.69 ± 0.10	1.68 ± 0.11	1.69 ± 0.10	<0.001
BMI (kg/m^2)	28.78 ± 6.65	32.70 ± 8.10	28.34 ± 6.32	<0.001
Systolic BP (mm Hg)	120.86 ± 16.96	128.30 ± 19.42	120.04 ± 16.46	<0.001
Diastolic BP (mm Hg)	70.24 ± 12.32	68.37 ± 14.05	70.46 ± 12.10	<0.001
Direct cholesterol (mg/dL)	1.37 ± 0.42	1.24 ± 0.35	1.38 ± 0.42	<0.001
Total cholesterol (mg/dL)	5.07 ± 1.05	4.80 ± 1.15	5.10 ± 1.04	<0.001
Physical activity, yes n (%)	3497 (53.30)	249 (37.90)	3248 (55.01)	<0.001

[1]p-value is obtained from an independent t-test for continuous and a Chi- square test for a categorical variable.

4.3.2 Identification of Risk Factors of Diabetes

Table 4.2 presented that identification the risk factors of diabetes based on LR. It was found that age, education, BMI, SBP, DBP, direct cholesterol, and total cholesterol were considered as signi ficant factors for diabetes (p-value<0.001).

Table 4.2. Identification the risk factors of diabetes using logistic regression.

Factors	OR	p-value	95% CI for OR	
			Lower	Upper
Age (years)	1.055	<0.001	1.047	1.064
Gender				
Female (*Ref*)	1.000			
Male	1.270	0.086	0.967	1.670
Race				
Black (*Ref*)	1.000			
Hispanic	0.746	0.206	0.470	1.167
Mexican	0.858	0.465	0.567	1.290
Other	1.109	0.623	0.732	1.670
White	0.469	0.263	0.358	0.619
Education				
8th Grade (*Ref*)	1.000			
9 - 11th Grade	0.577	0.005	0.394	0.844
College Grad	0.641	0.019	0.441	0.932
High School	0.746	0.010	0.526	1.060
Some College	0.786	0.030	0.559	1.112
Marital Status				
Divorced (*Ref*)	1.000			
Live Partner	0.562	0.240	0.335	0.915
Married	0.780	0.083	0.591	1.038
Never Married	0.600	0.112	0.404	0.887
Separated	0.985	0.957	0.552	1.709
Widowed	0.541	0.162	0.368	0.796
Occupation				
Looking working (*Ref*)	1.000			
Not working	1.061	0.838	0.619	1.932
Working	0.776	0.371	0.457	1.402
Weight (kg)	0.974	0.135	0.940	1.008
Height (m)	1.029	0.155	0.989	1.072
BMI (kg/m^2)	1.170	0.002	1.061	1.292
Systolic BP (mm Hg)	1.007	0.007	1.002	1.013
Diastolic BP (mm Hg)	0.994	0.080	0.986	1.001
Direct cholesterol (mg/dL)	0.510	<0.001	0.381	0.678
Total cholesterol (mg/dL)	0.809	<0.001	0.738	0.885
Physical activity				
No (*Ref*)	1.000			
Yes	0.906	0.319	0.747	1.099

Ref: Reference.

4.3.3 Effect of Keeping Data Size Fixed on Performance of ML-based System

The effect of keeping fix data size (n=6561) on accuracy of 4 ML-based classifiers over 3 CV protocols were presented in Figure 4.3. It was also presented that the ACC of 4 ML-based classifiers have been increased with increasing the number of CV protocol (K2 to K5, to K10). It was also noted that the highest ACC of **94.25%** was provided by RF-based classifier for K10 protocol (see Table 4.3). Furthermore, the highest SE, PPV, NPV, and FM were also supported by RF- based classifier for K10 protocol (see Table 4.4).

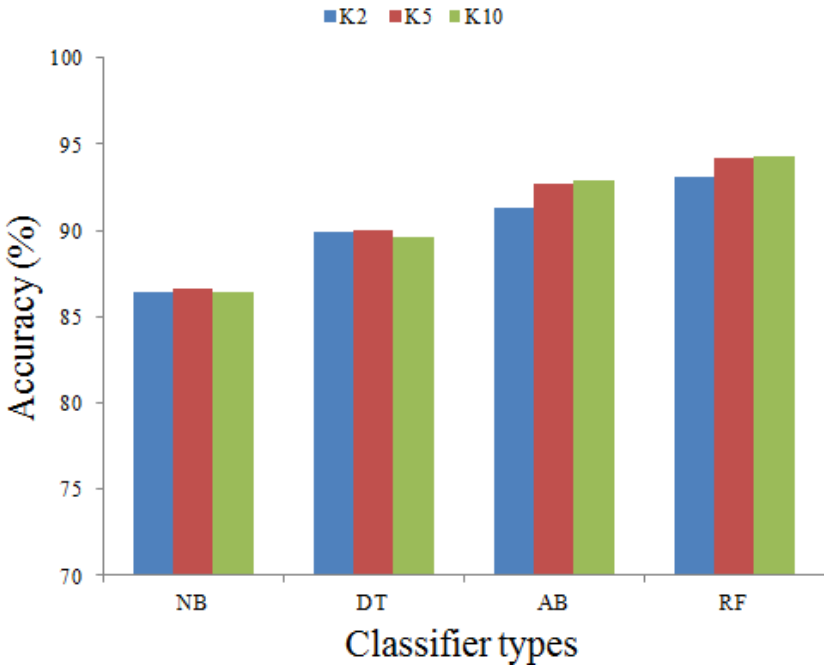


Figure 4.3. Effect of keeping data size fixed on accuracy (%) of 4 classifiers for 3 protocols.

Table 4.3. Comparison of accuracy (%) of 4 classifiers for 3 protocols.

Classifier types	Protocol types		
	K2	K5	K10
NB	86.42	86.61	86.70
DT	89.90	89.97	89.65
AB	91.32	92.72	92.93
RF	**93.12**	**94.15**	**94.25**

Bold values indicate the result of proposed method.

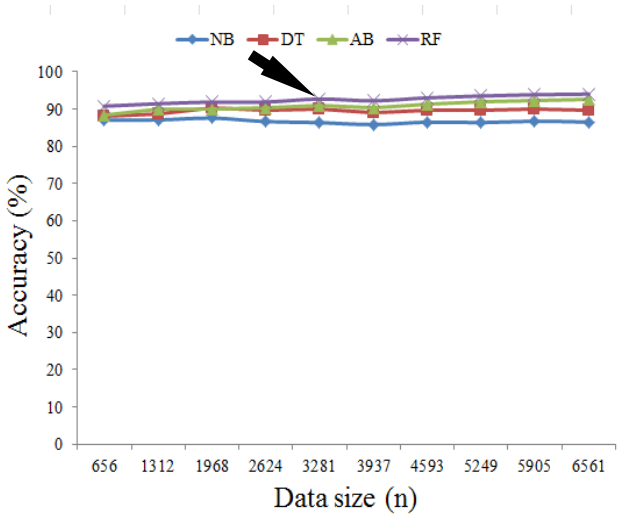
Table 4.4. Four performance evaluation parameters of 4 classifiers for 3 protocols.

Protocol types	Classifier Types	Performance evaluation parameters (%)			
		SE	PPV	NPV	FM
K2	NB	92.11	92.75	33.16	92.43
	DT	99.12	90.56	37.28	94.64
	AB	96.04	94.42	57.91	95.22
	RF	**99.56**	**93.25**	**89.98**	**96.30**
K5	NB	92.05	93.02	33.71	92.53
	DT	99.18	90.59	38.00	94.68
	AB	96.60	95.39	64.93	95.99
	RF	**99.54**	**94.29**	**91.53**	**96.84**
K10	NB	92.13	92.74	34.08	92.43
	DT	99.48	90.06	40.25	94.52
	AB	96.81	95.40	67.49	96.09
	RF	**99.57**	**94.34**	**92.59**	**96.88**

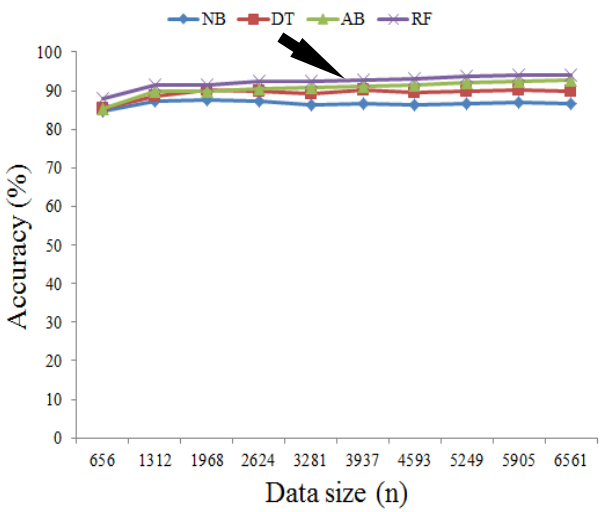
Bold values indicate the result of proposed method.

4.3.4 Effect of Varying Data Size on Performance of ML-based Systems

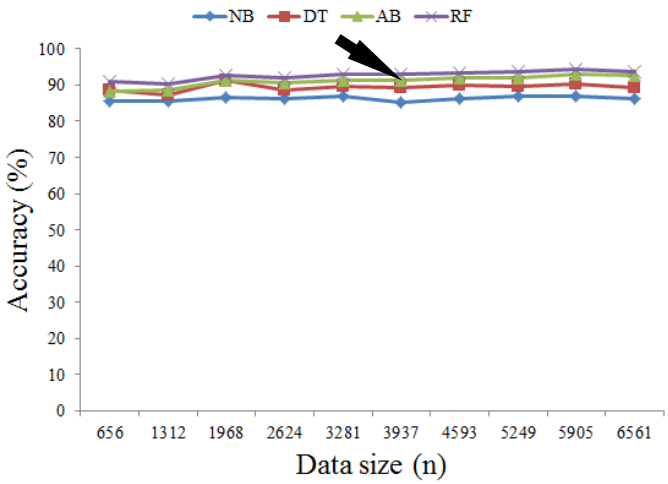
The effects of varying data size on the efficiency of ML-based models for 3 CV protocols were presented in Figure 4.4. We have separated the training data size into 10 parts as 656, 1312, 1968, 2624, 3281, 3937, 4593, 52459, 5905, and 6561. We have trained the 4 ML-based models on these training datasets and computed their ACC for 3 CV protocols. It was noted that at least 70% respondents (4593) were needed for the net generalizations. Figure 4.4 also showed the ACC of 4 ML-based classifiers were increased with increasing in data size and RF-based system was achieved better performance compared to others. The system ACC of 4 classifiers with varying data sizes for K2 protocol were presented in Appendix A1: Table A1.1, K5 protocol in Appendix A1: Table A1.2 and K10 protocol in Appendix A1: Table A1.3. Then we have computed the system mean ACC by averaging ACC of 4 classifiers over varying data sizes for 3 CV protocols which were shown in Table 4.5. It was also indicated that RF-based classifier was performed better compared to others.



(a) K2 protocol.



(b) K5 protocol.



(c) K10 protocol.

Figure 4.4. Effect of accuracy over varying data size (n) for 3 CV protocols. (a) K2 protocol; (b) K5 protocol; and (c) K10 protocol.

Table 4.5. Systems mean accuracy (%) of 4 classifiers for 3 partition protocols.

Protocol types	Classifier types			
	NB	DT	AB	**RF**
K2	86.67	89.52	90.79	**92.54**
K5	86.61	89.28	90.58	**92.33**
K10	86.24	89.38	91.08	**92.75**

Bold values indicate the result of proposed method.

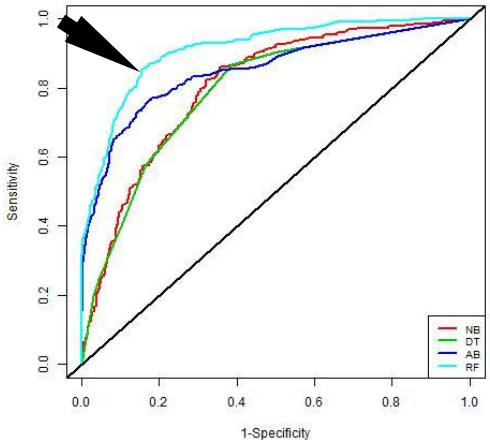
4.3.5 Receiver Operating Characteristics Analysis

The ROC is a graphical procedure that is plotted based on sensitivity vs. '1-specificity'. The ROC curves of 4 classifiers for 3 partition protocols were presented in Figure 4.5. It is observed that the higher AUC of 0.95 was obtained by RF-based classifier for K10 protocol compared to others and the corresponding AUC values were presented in Table 4.6.

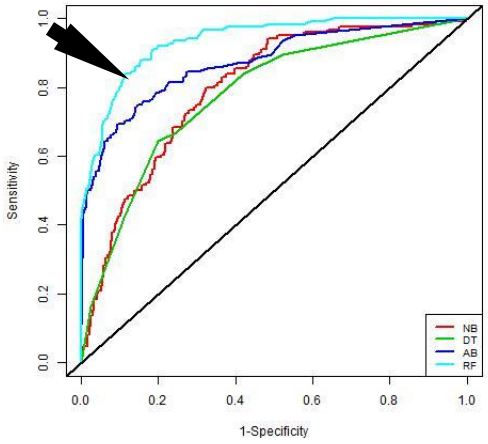
Table 4.6. Comparison of AUC of 4 classifiers for 3 CV protocols.

Classifier types	Protocol types		
	K2	K5	K10
NB	0.80	0.81	0.82
DT	0.78	0.78	0.78
AB	0.86	0.89	0.90
RF	**0.91**	**0.94**	**0.95**

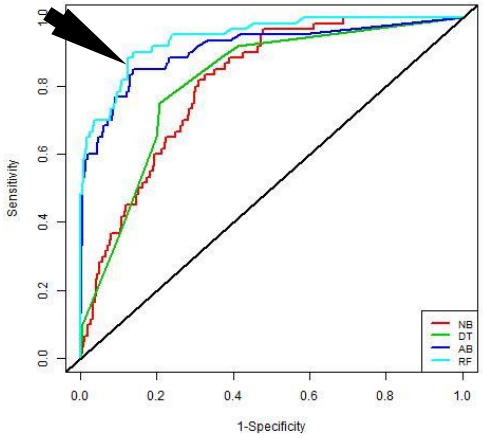
Bold values indicate the result of proposed method.



(a) K2 protocol.



(b) K5 protocol.



(c) K10 protocol.

Figure 4.5. ROC curves of 4 classifiers for 3 CV protocols: (a) K2 protocol; (b) K5 protocol and (c) K10 protocol.

4.3.6 Validations of the Proposed ML-based System

To validate our proposed method of this study (**Chapter 4**), liver dataset was used, taken from UCI data repository (Ramana et al., 2012). The dataset consisted of 10 factors as well as 583 respondents having 416 liver patients. Our findings illustrate that RF-based system gave ACC of **70.59%** (see Table 4.7) compared to others which was validated to our proposed (RF) method.

Table 4.7. Validation of the proposed (RF) method for liver dataset.

Classifier types	Protocol types		
	K2	K5	K10
NB	55.85	55.38	55.21
DT	66.83	67.29	67.62
AB	69.24	69.6	70.45
RF	**70.42**	**70.44**	**70.59**

Bold values indicate the result of proposed method.

4.4 Summary of the Chapter

In this section, **Chapter 4** is summarized as:

1. Data:
   - Extraction: Extract the diabetes dataset into a dta (Stata) format from NHANES.
   - Data cleaning: Drop the missing values and unusual observations from the analysis.
   - Feature extraction: Extract 7 high-risk factors of diabetes using LR.
2. Modeling:
   - CV protocol: 3 CV-based (K2, K5, and K10) protocols were used.
   - Model selection: Apply 4 classifiers as NB, DT, AB, and RF for diabetes prediction.
3. Model performance evaluation:
   - Metrics: Use ACC, SE, PPV, NPV, FM, and AUC for classifiers evaluations.
   - Interpretation: Finding show that RF classifier performed better compared to others.
   - Validation of proposed method: Liver dataset was used for validation of RF-based classifier.

Chapter 5 Accurate Risk Prediction of Diabetes based on Machine Learning: Role of Missing value and Outliers

5.1 Introduction

Accurate classification is an essential and exhaustive issue for diagnosis and prognosis of diabetic (Barakat et al., 2010) since most healthcare data have missing values, correlation-based structure, nonlinear, non-normal, the course of dimensionality, and complex in nature (Maniruzzaman et al., 2017). The adequacy of ML-based systems is impaired when diabetic data have missing value or outliers. Several ML-based models, such as LDA, QDA, NB, SVM, ANN, FFNN, DT, J48, RF, GPC, LR, and KNN have been carried out for diagnosis and monitoring of diabetes disease (Bashir et al., 2016; Maniruzzaman et al., 2017; Lee and Yoon, 2017). These ML-based classifiers are unable to accurately classify diabetes while data includes missing values/outliers. As a result, ML-based classifiers do not achieve greater accuracy (Cokluk et al., 2011; Baneshi et al., 2012; Leys et al., 2013; Zainuri et al., 2015; Maniruzzaman et al., 2017). Outlier removal and the monitoring of missing values is a critical issue in statistics and have never been neglected.

Previous ML-based models (Bashir et al., 2016) were ineffective since their models are either (a) directly on the raw data without feature extraction or (b) on raw data without outlier removal or (c)without inserting replacing values for missing values or (d)simply filling missing values with the mean. In addition, the replacements of outliers, measured by mean are very sensitive (Manikandan, 2011). As a consequence, their ML-based models are unable to achieve greater accuracy. Several authors have been attempted to substitute outlier or missing values, but in the non-classification framework (Cokluk et al., 2011; Baneshi et al., 2012; Leys et al., 2013; Zainuri et al., 2015;). Our methods are motivated by the spirit of these statistical measures embedded in a classification framework. The accuracy of ML-based models may be improved if one can be replaced the missing values and outliers by group median and median. Further, this study proposes a ML-based system by selecting the combination of a FS method and a classifiers from 6 FS methods and 10 classifiers. The hypothesis has been laid out in Figure 5.1, while the diabetic input data has a two-stage for data preparation as replace (a) the missing value by group

median and (b)outliers by median. The comparator helps in comparing the ACC, when the data has (a)no missing values but has outliers (b) no missing values and no outlier. RF-based classifier has adapted (Hasan et al., 2016) for both feature extraction and diabetic prediction. In this study, we hypothesize that by (a) replacing missing values with group median and outlier by median, and (b) using feature extraction by RF with the RF-based system yields the highest accuracy.

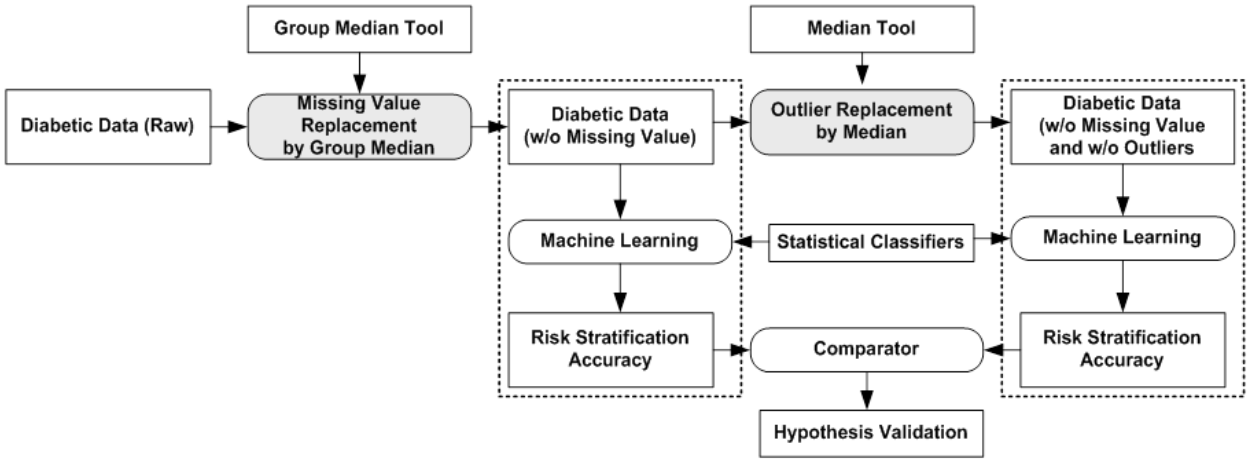


Figure 5.1. Data preparation of diabetic data by missing value replacement and outlier removal.

Thus, compared to the previous research, the following are the novelties of this present study:

(i) Establishing an ML-based method where the missing values can be replaced by group median and outliers by median if there are outliers, while outliers can be tested based on inter-quartile range (IQR).

(ii) Optimizing the ML-based framework by choosing the best combination of a FS and a classifier among the 6 FS approaches (RF, LR, MI, PCA, ANOVA, and FDR) and 10 classifiers (RF, LDA, QDA, NB, GPC, SVM, ANN, AB, LR, and DT).

(iii) Appling 5 sets of CV protocols (K2, K4, K5, K10, and JK) for the generalization of the ML-based system and computed their performance parameters, namely ACC, SE, SP, PPV, NPV, and AU. Reliability index (RI) and stability index are also used to check the validity of our study.

5.2 Materials and Methods

5.2.1 Dataset

The dataset, contained 768 female respondents with 21 years, having 268 (34.90%) diabetic and has been extracted from UCI repository. The database had 5 zeros in glucose, 35 zeros in blood pressure, 27 zeros in BMI, 227 zero in triceps and 374 zeros in insulin. We have divided the dataset into 2 groups like diabetic vs. control, and these meaningless values were replaced by group median. We have also checked the outliers/unusual observations by IQR and replaced these outliers by the median if there exists outliers/unusual observations in the dataset. Descriptions of the attributes and brief statistical summary were presented in Table 5.1.

Table 5.1. Demographics of the diabetic patient cohort.

Attributes	Descriptions	Attributes type's	Mean ± SD
Pregnant	Number of times pregnant	Continuous	3.84 ±3.36
Glucose	Plasma glucose (2-hours)	Continuous	121.67 ±30.46
Pressure	Diastolic blood pressure (mm Hg)	Continuous	72.38 ±12.10
Triceps	Triceps skin fold thickness (mm)	Continuous	29.08 ±8.89
Insulin	Two hours serum-insulin (μ U/ml)	Continuous	141.76 ±89.10
BMI	Body mass index (weight in kg/ (height in m)^2)	Continuous	32.43 ±6.88
Pedigree	Diabetes pedigree function	Continuous	0.47 ±0.33
Age	Age (years)	Continuous	33.24 ±11.76

5.2.2 Machine Learning System

Figure 5.2 presented the concept of the overall ML-based system (proposed). This followed the output of ML-based method while the input data was preprocessed by replacing the missing values and outlier with groups median and median. The first phase is divided the diabetic data in to two phases as training and test data. The following stage, extract features using 6 FS methods like RF, LR, MI, PCA, ANOVA, and FDR. The main role of this stage is to choose the dominant features by dropping the complexity of data. Additionally, 10 ML-based classifiers as LDA, QDA, NB, GPC, SVM, ANN, AB, LR, DT, and RF have added to the training database and estimated ML-based parameters. ML-based parameters and dominant features have extracted for test data is used to predict diabetic patients. Monitoring output of ML-based systems yields ACC, SE, SP, PPV, NPV, and AUC which were shown in Figure 5.3.

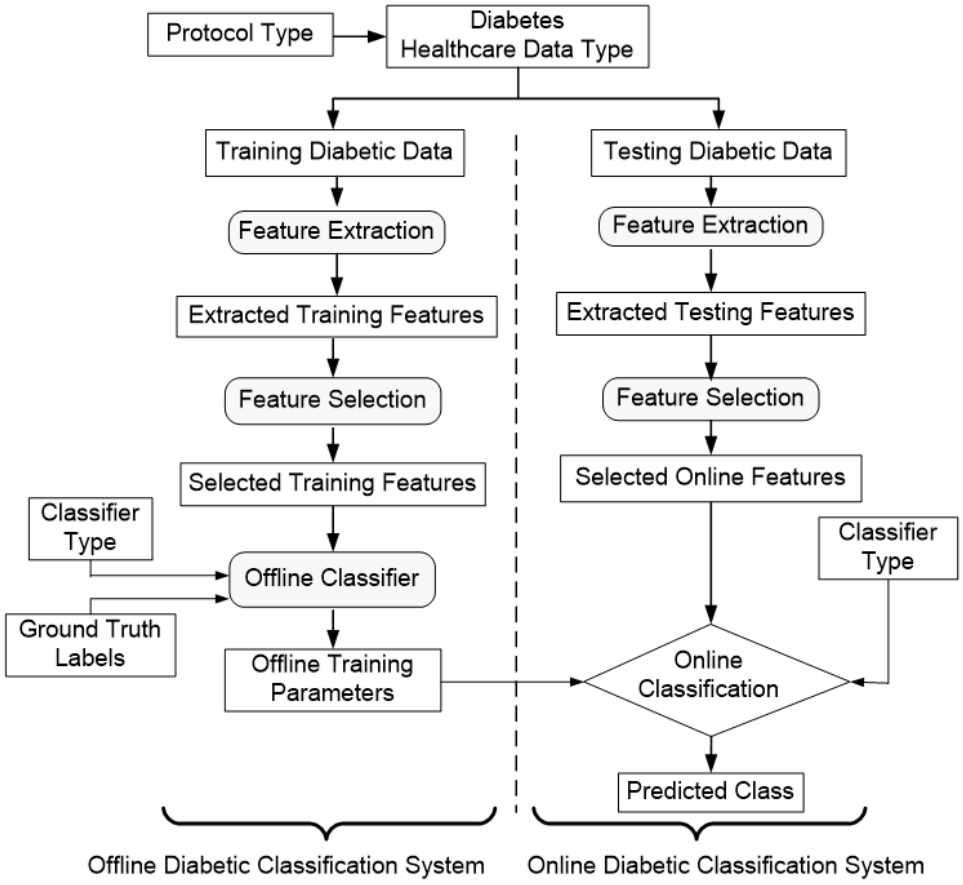


Figure 5.2. Architecture of the ML-based framework.

5.3 Results and Discussion

5.3.1 Select the Best FS Techniques over K-fold CV and ORT

We have used 6 FS methods and for our notational simplicity, we may be defined as F1 for RF, F2 for LR, F3 for MI, F4 for PCA, F5 for ANOVA and F6 for FDR on both presence of outliers (O1) and imputed outliers by median (O2) datasets. F5-based FS technique gave ACC of 81.94% for K2 protocol and presence of outliers. It was observed that ACC was increased by increasing the value of K for both presence (O1)and absence of outliers (O2). F2 gave the highest ACC of 84.66% of the same protocols for absence of outliers (O2). In the same way for K4, F2 and F4 gave the highest ACC of 82.73% and 86.16% for both presence and absence of outliers (O2). The highest ACC of 85.86% was obtained by RF for K10 and ACC of 88.45% for JK protocol while data have the absence of outliers (O2) which are given in detail in Table 5.2. So it was concluded that RF (F1)-based classifier was performed better for both presence (O1) and absence of outliers (O2) datasets.

Table 5.2. Comparison of accuracy (%) in presence and absence of ouliers for different FS's and protocols.

FST	Presence of outliers (O1)					Absence of outliers (O2)				
	K2	K4	K5	K10	JK	K2	K4	K5	K10	JK
F1	81.58	81.97	84.23	**84.66**	**86.05**	84.30	85.71	85.88	**85.86**	**88.45**
F2	81.84	81.45	83.23	83.56	86.77	84.66	86.16	84.40	84.40	88.40
F3	81.92	82.73	81.88	81.90	86.19	84.50	85.27	85.64	84.73	88.45
F4	81.48	81.98	83.09	82.23	85.66	83.71	84.60	83.73	84.60	87.91
F5	81.94	81.94	82.51	82.47	87.89	83.77	83.44	84.20	84.01	87.75
F6	71.48	73.51	74.90	74.82	78.13	75.53	75.82	76.77	77.35	79.32

Bold values indicate the result of proposed method.

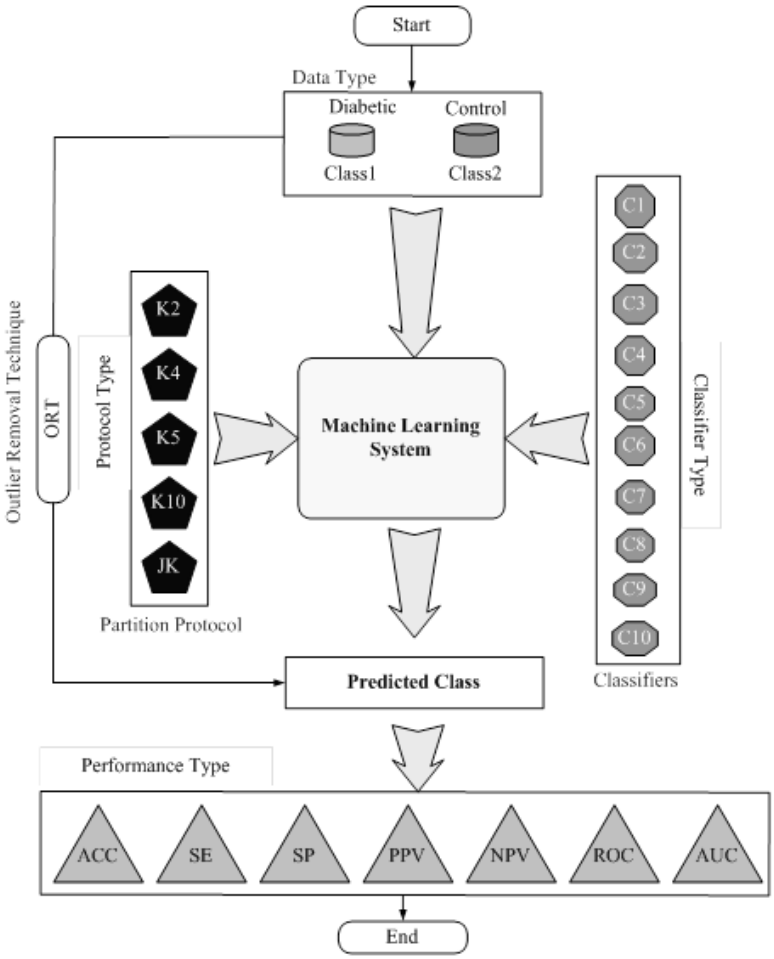


Figure 5.3. Performance of the ML-based system in the presence and absence of outliers.

5.3.2 Comparison of Performance of ML-based Classifiers

This section is implemented to investigate the 10 ML-based classifier's performances with changing CV protocols in presence and absence of outliers in the dataset. We may be defined 10

ML-based classifiers as C1 for LDA, C2 for QDA, C3 for NB, C4 for GPC, C5 for SVM, C6 for ANN, C7 for Adaboost, C8 for LR, C9 for DT, and C10 for RF. Table 5.3 and Table 5.4 indicates that ACC was increased with increasing the value of K for both presence and absence of outliers in the dataset. For K2 protocols, the combination of F1-C10 based classifier gave ACC of 89.09% for the presence of outliers and 88.98% ACC for the absence of outliers in the dataset; (ii) ACC of C10-based classifier is also increased with increasing the value of K (2 to 4).

Table 5.3. Comparisons of accuracy (%) of 10 classifiers and 6 FS methods for presence of outliers.

PT*	FST	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
K2	F1	77.21	73.83	76.56	85.23	85.18	78.88	86.33	78.54	86.93	**89.09**
	F2	77.76	76.38	77.86	84.71	84.92	76.88	85.76	79.17	87.24	87.73
	F3	77.24	74.27	77.03	83.88	83.93	81.98	87.16	78.57	86.25	88.88
	F4	77.55	75.13	77.45	82.89	82.99	79.82	85.08	79.90	86.51	87.47
	F5	77.73	75.36	77.55	84.48	85.08	78.70	85.78	79.56	87.21	87.97
	F6	69.64	67.97	68.78	72.29	71.61	68.62	73.67	71.20	75.18	75.81
K4	F1	76.30	73.49	75.73	86.25	86.46	79.90	86.41	78.39	86.93	**89.79**
	F2	77.34	74.84	76.93	85.68	83.39	75.68	84.74	79.27	88.02	88.65
	F3	78.02	75.26	77.71	85.52	84.90	81.56	87.55	80.10	86.93	89.79
	F4	77.55	75.10	77.86	84.11	82.97	80.62	85.21	80.94	86.98	88.49
	F5	78.96	76.72	78.28	85.68	86.35	80.10	85.00	80.73	87.66	89.06
	F6	70.94	68.85	69.95	75.83	73.70	70.52	75.57	73.28	77.92	78.49
K5	F1	80.32	77.40	79.48	88.51	87.21	81.17	87.34	82.40	88.57	**90.78**
	F2	79.22	77.27	78.64	87.53	86.56	79.68	86.30	80.78	87.34	88.96
	F3	77.47	73.77	76.62	84.81	84.22	81.36	85.91	78.96	87.01	88.70
	F4	77.92	76.30	78.18	85.39	84.03	82.66	86.36	82.60	87.86	89.55
	F5	77.79	75.19	77.21	85.58	84.81	81.04	86.30	80.32	87.73	89.16
	F6	71.62	71.82	71.88	78.70	75.39	72.53	74.55	74.74	78.57	79.22
K10	F1	77.62	74.48	77.03	85.57	85.00	80.93	86.63	79.62	87.07	89.59
	F2	78.18	76.36	78.44	88.05	85.58	78.31	88.70	81.69	89.35	**90.91**
	F3	76.75	73.38	76.10	84.81	83.12	81.69	85.71	80.39	86.88	90.13
	F4	77.92	75.32	77.92	85.84	83.25	78.96	85.45	82.34	86.23	89.09
	F5	76.75	75.58	77.53	86.88	85.19	81.43	84.29	80.91	87.01	89.09
	F6	72.73	69.87	71.56	79.09	75.58	72.34	74.29	75.97	77.66	79.09
JK	F1	77.92	76.05	77.44	89.01	90.41	82.16	99.92	78.32	89.28	**99.99**
	F2	78.27	81.22	78.78	88.12	89.24	83.20	99.49	79.16	90.23	**99.99**
	F3	78.09	76.24	77.04	88.49	89.99	83.82	99.87	78.37	90.03	**99.99**
	F4	77.77	75.60	77.34	86.77	88.97	82.98	99.62	80.29	87.31	**99.99**
	F5	83.67	83.84	82.84	88.42	88.28	79.06	98.63	84.13	88.59	**99.99**
	F6	70.10	70.20	68.88	77.30	76.63	75.82	96.02	71.26	76.97	**99.99**

Bold values indicate the proposed method results.

Table 5.4. Comparisons of accuracy (%) of 10 classifiers and 6 FS methods for the absence of outliers.

PT*	FST	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
K2	F1	83.88	83.78	84.37	87.34	86.15	79.14	86.67	85.29	86.54	**88.98**
	F2	84.40	84.11	84.71	86.82	85.05	77.66	84.87	85.23	86.02	87.76
	F3	83.39	83.41	83.70	86.48	85.21	79.87	85.42	84.82	85.05	87.66
	F4	82.50	82.84	82.50	85.16	84.32	80.03	84.19	83.70	85.13	86.72
	F5	83.10	83.83	83.10	85.81	83.93	77.63	84.11	84.27	84.82	86.75
	F6	76.56	76.28	76.90	77.37	75.99	72.16	72.16	77.94	73.88	76.04
K4	F1	85.31	85.21	85.05	88.28	86.61	79.58	87.50	87.45	87.03	**89.58**
	F2	83.80	84.17	84.06	87.24	85.99	79.90	86.35	85.83	86.41	88.96
	F3	85.10	84.43	84.90	88.28	86.61	79.79	86.09	86.61	85.83	89.48
	F4	83.49	83.75	83.23	86.41	84.90	79.90	84.84	85.42	85.89	88.13
	F5	82.45	82.50	82.40	86.15	83.91	77.71	83.33	84.01	85.26	86.67
	F6	76.61	75.99	76.82	78.33	75.47	72.03	71.15	79.06	76.46	76.30
K5	F1	84.94	84.48	84.29	88.70	86.49	79.29	86.49	86.82	87.53	**89.81**
	F2	82.47	82.27	82.27	86.75	84.94	77.99	86.88	85.26	86.17	88.96
	F3	84.22	84.61	83.77	88.31	86.82	79.94	87.08	85.91	86.10	89.61
	F4	82.08	82.08	81.95	86.17	83.70	80.58	83.05	84.35	85.71	87.66
	F5	83.12	83.38	83.38	87.08	84.35	79.16	82.21	84.94	85.91	88.44
	F6	77.08	76.36	77.14	78.57	77.34	72.66	73.57	79.68	76.62	78.64
K10	F1	83.38	84.16	82.73	89.35	86.49	81.17	84.42	85.97	87.27	**92.26**
	F2	85.45	85.71	85.97	89.61	86.62	78.83	87.27	88.05	87.79	90.26
	F3	82.86	82.60	83.38	88.05	85.19	76.75	86.62	86.62	86.49	88.70
	F4	82.60	83.77	82.73	87.14	83.64	79.74	82.86	87.40	87.79	88.31
	F5	82.86	83.12	82.73	87.79	85.06	76.49	82.73	86.36	85.97	87.01
	F6	77.66	77.27	77.53	80.13	77.14	75.32	72.99	80.26	76.62	78.57
JK	F1	84.12	84.31	83.74	88.43	88.66	80.40	99.82	84.78	90.20	**99.99**
	F2	84.01	84.79	84.03	88.72	88.30	79.14	99.24	85.66	90.14	**99.99**
	F3	84.12	84.31	83.74	88.44	88.62	80.50	99.82	84.78	90.20	**99.99**
	F4	83.45	84.00	82.76	87.30	87.45	81.67	99.81	84.13	88.58	**99.99**
	F5	81.10	82.26	83.66	88.17	89.09	83.21	99.82	81.39	90.23	**99.99**
	F6	71.10	70.32	69.88	78.30	77.63	77.82	97.02	72.26	76.97	**99.97**

*Protocol Types. Bold values indicate the proposed method results.

Table 5.3 and Table 5.4 also showed that F1-C10 based combination also gave the highest ACC for both the presence (89.79%) and absence of outliers (89.58%) in the datasets. Similarly it can be showed that for K10 protocols; F1-C10 gave the highest ACC of 90.91% and 92.26% for the presence and absence of outliers. It was also noticed that the combination of all FS with RF-based classifier gave 99.99~100.00% ACC for both the presence/absence of outliers in datasets. Therefore, it may be concluded that F1-C10 based combination was performed better for both (presence/absence of outliers) datasets.

5.4 Hypothesis Validation and Performance Evaluation

5.4.1 Hypothesis Validation

The hypothesis of this study (**Chapter 5**) was RF-based in ML-based framework yields the highest ACC while the replacement of missing values by the group median and outliers by the median. The comparison of ACC of 10 classifiers and 6 FS methods in the presence of outliers (O1) and absence of outliers (O2) was presented in Table 5.5. It was also demonstrated that the hypothesis has been validated.

Table 5.5. Comparison of accuracy (%) of 10 classifiers and 6 FS methods in presence and absence of outliers over protocols.

CT*	Presence of outliers (O1)						Absence of outliers (O2)					
	F1	F2	F3	F4	F5	F6	F1	F2	F3	F4	F5	F6
C1	78.14	78.15	77.51	77.74	78.22	71.23	84.03	83.73	83.68	82.82	83.04	76.98
C2	75.45	77.21	74.58	75.49	76.75	69.63	83.99	83.73	83.82	83.29	83.33	76.48
C3	77.56	78.13	76.90	77.75	78.7	70.54	83.77	83.67	83.73	82.63	82.89	77.10
C4	87.67	86.82	85.50	85.00	85.92	76.48	88.40	87.46	87.70	86.44	87.05	78.60
C5	87.44	85.94	85.23	84.44	85.85	74.07	86.72	85.97	86.37	84.80	85.11	76.48
C6	81.15	78.75	82.08	81.01	80.62	71.00	79.87	78.47	79.39	80.38	78.01	73.04
C7	89.92	89.00	89.24	88.34	88.39	74.52	88.38	89.07	89.06	86.95	86.20	72.47
C8	80.13	80.01	79.28	81.21	80.35	73.80	85.71	85.77	85.59	85.00	84.74	79.24
C9	88.16	88.44	87.42	86.98	87.88	77.33	87.24	87.11	86.85	86.62	86.14	75.90
C10	**91.35**	91.25	91.50	90.92	90.84	78.15	**92.29**	91.05	90.98	90.16	89.81	77.39

* Classifier Types; Bold values indicate the proposed method results.

5.4.2 Performance Evaluation

5.4.2.1 Reliability

Reliability index (RI) and stability index is required the performance evaluation of ML-based system (Shrivastava et al., 2017) which was presented in Figure 5.4. RI was computed by the proportion of the SD (σ_n) and mean ACC ($\boldsymbol{\mu}_n$) over data size (n) and mathematically expressed as

$$\xi_n(\%)=\left(1-\frac{\sigma_n}{\boldsymbol{\mu}_n}\right)\times 100 \tag{5.1}$$

The system RI of $\bar{\xi}$ is also calculated using the following formula:

$$\bar{\xi}(\%)=\left(\frac{\sum_{r=1}^{n}\xi_r}{n}\right) \tag{5.2}$$

Figure 5.5 and Figure 5.6 show that RI for all F_i-C_j (i=1, 2…, 6 and j=1, 2… 10) based 60 combinations as data size (n). Further, the system RI was shown in Table 5.5 for the presence of

outliers and Table 5.6 for absence of outliers. It was confirmed that the best performance of F1-C10 based combination provided the best performance for both the presence and absence of outliers in the datasets. It was also found that the data system is stable within 2% tolerance limit.

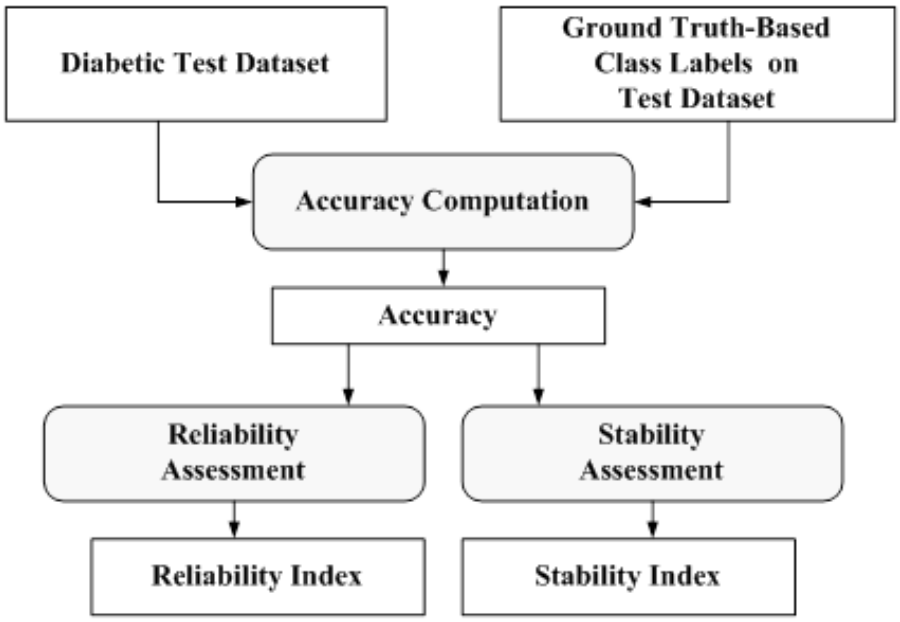


Figure 5.4. Performance evaluations of ML-based system in presence and absence of outliers.

Table 5.6. Comparison of RI (%) 10 classifiers and 6 FS methods for the presence of outliers.

FST	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
F1	96.97	96.17	96.82	97.84	98.25	96.50	97.73	96.94	97.78	**98.45**
F2	96.52	96.04	96.34	97.56	97.68	96.42	97.56	97.22	97.58	98.01
F3	96.08	95.22	95.88	96.98	97.14	96.18	97.37	96.73	97.53	98.01
F4	96.25	95.95	96.40	97.43	97.51	97.11	97.34	97.43	97.44	98.31
F5	96.79	96.49	97.13	97.44	97.03	96.77	97.68	97.17	97.13	98.24
F6	96.23	96.07	95.42	96.40	95.94	95.59	96.45	96.32	96.78	97.57

Bold values indicate the proposed method results.

Table 5.7. Comparison of RI (%) 10 classifiers and 6 FS methods for the absence of outliers.

FST	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
F1	97.73	97.21	97.14	97.77	97.74	96.81	97.06	97.54	97.16	**98.02**
F2	97.17	97.02	97.33	97.52	97.24	96.43	97.05	97.34	97.19	97.91
F3	97.00	96.43	96.80	97.59	97.47	96.41	96.69	97.31	97.22	97.72
F4	97.11	96.32	97.10	97.41	96.77	95.58	96.50	97.44	96.93	97.84
F5	97.23	97.21	97.01	97.67	97.34	96.67	96.53	97.35	97.39	97.73
F6	96.24	96.18	96.06	97.38	96.18	95.34	95.96	96.96	96.15	96.46

Bold values indicate the proposed method results.

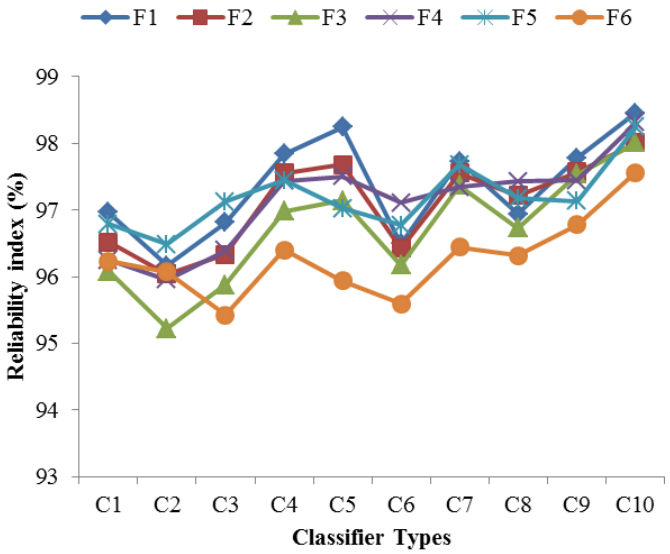


Figure 5.5. Comparison of RI (%) of 10 classifiers and 6 FST's for the presence of outlier.

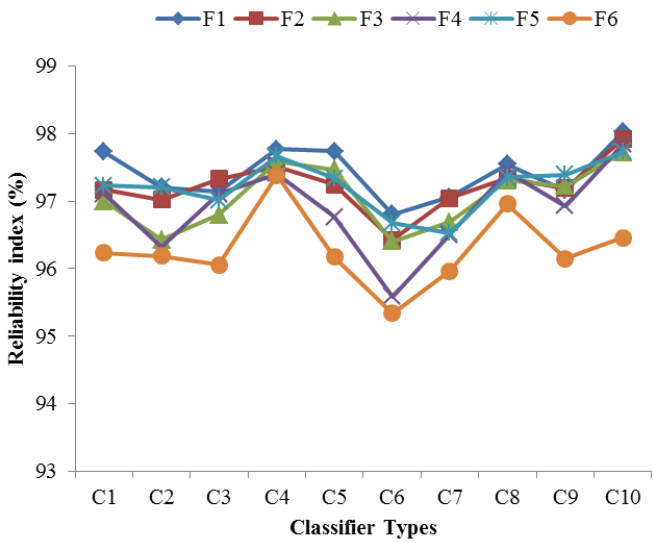


Figure 5.6. Comparison of RI (%) of 10 classifiers and 6 FST's for the absence of outliers.

5.5 Summary of the Chapter

In this section, we summarize **Chapter 5** at a glance as follows:

1. Data:
   - Extraction: Extract the Pima Indian diabetes dataset into a .csv format from UCI.
   - Data processing: Replace missing values and outliers by group median and median.
   - Feature extraction: Extract risk factors of diabetes using 6 FS method as: RF, LR, MI, PCA, ANOVA, and FDR.
2. Modeling:
   - CV protocol: 5 CV-based (K2, K4, K5, K10, and JK) protocols were used.
   - Classifiers: Apply 10 classifiers as LDA, QDA, NB, GPC, SVM, ANN, AB, LR, DT, RF.
3. Model based performance evaluation:
   - Metrics: Use ACC, SE, PPV, NPV, AUC, and RI as performance of classifiers
   - Interpretation: Interpret the performance evaluation metrics to compare the classifiers and our findings showed that the combination of RF-RF based FS and classifier gave **92.26%** ACC and **85.86%** RI for K10 protocol and **99.99%** ACC and **88.45%** RI for JK protocol.

Chapter 6 Statistical Characterization and Classification of Colon Cancer using Machine Learning Paradigms

6.1 Introduction

Globally, cancer is the 2nd major cause of death. Different forms of cancer disorder, such as colon, lung, breast, prostate, and so on are found in human bodies (Hollstein et al., 1991; Bray et al., 2018). Approximately, 12.4 million new cases were diagnosed by cancer worldwide in 2008 (Giovannucci et al., 2010), 18.1 million in 2018 (Bray et al., 2018), and this figure extended to 19.30 million in 2020 (Ferlay et al., 2020). Moreover, 9.6 million people died in 2018 and this figure was reached about 9.9 million in 2020. Among them, 18% of deaths were occurred due to lung cancer, 9.4% death for colorectal, 8.3% for liver, 7.7% for stomach, 4.7% for pancreas, and 3.8% for prostate. Therefore, it is clear that the no. of new cases and deaths from cancer has steadily increased over time. It is required to be diagnosed cancer patients and determined the high-risk genes of cancer.

Generally, gene expression datasets have a huge no. of genes as well as a limited sample size. As a consequence, there exists a high correlation among these genes. Previously, lots of supervised and unsupervised ML-based methods implemented for significant gene identification (Matthias et al., 2003; Monti et al., 2003). These ML-based methods face with over-fitting and multi-collinearity problems caused by limited sample size, noise and huge no. of genes (Hong and Cho, 2006; Hung and Wang, 2006). It is urgent to eliminate the noisy genes through a novel FS method and also predict the high-risk genes using ML-based methods based on different CV-protocols. Many authors implemented their studies unsupervised algorithms like likely hierarchical clustering (Yeoh et al., 2002), K-means clustering (Li et al., 2001), SOM (Hautaniemi et al., 2003), FNN (Tung et al., 2005), etc., for detecting responsible genes for cancer. In addition, various supervised techniques as ANN (Ando et al., 2003; Takahashi et al., 2004), SVM (Guyon et al., 2002; Mao et al., 2005) and so on were used both for gene extraction and prediction. Moreover, different parametric and non-parametric tests like t-test (Jeanmougin et al., 2010; Kuyuk et al., 2017), KW-test, (Chen et al., 2005; Shi et al., 2015) were also used only for significant gene extraction but not used for prediction of cancer patients. No attempt was

found to combine a statistical test along with a ML-based classifier for accurate gene identification and prediction. Therefore, we believe that the highest accuracy is provided by choosing a proper statistical test, CV protocol along with a classifier. The global system of high-risk gene extraction with ML-based method for classification is presented in Figure 6.1.To fulfill the above foundational assumptions, this study presents a two-stage systems where the first stage is identify of significant cancerous genes using 4 statistical tests as t, F, WCSRS, and KW based on p-values. The 2nd stage is to propose a classifier for performing better results for prediction of cancer among 10 ML-based classifiers like LDA, QDA, NB, GPC SVM, ANN, LR, DT, AB, and RF. The ML-based framework is assessed by 3 CV protocols as K2, K10, and JK. Reliability index is also used for the evaluation of ML-based framework. Therefore, this study offers the following contributions:

(i) Propose a ML-based framework by choosing a suitable statistical test and classifier from 4 statistical tests (WCSRS, t, KW, and F) and 10 classifiers (LDA, QDA, NB, GPC, SVM, ANN, LR, DT, AB, and RF).

(ii) Understands ML-based system using 3 CV protocols and 4 statistical tests combined with 10 classifiers. This further involved optimization of the best matching strategy between data normalization, detection and classification.

(iii) Observing the effect of fixing/varying data size on performance of ML-based systems and computing their performance based on ACC, SE, SP, PPV, NPV, and FM. RI and AUC was also used as a part of evaluation.

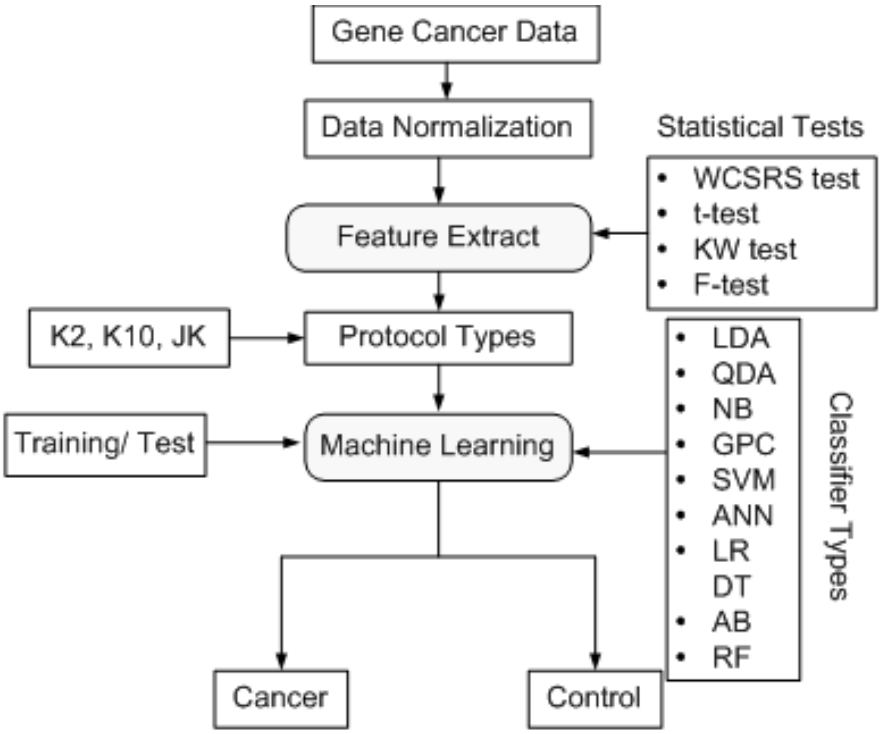


Figure 6.1. Global system of high-risk gene detection, CV protocols for ML-based system.

6.2 Materials and Methods

6.2.1 Dataset

The colon cancer dataset, extracted from Kent ridge biomedical data repository, USA that was publicly available (Alon et al., 1999). The dataset contained 2,000 genes, and 62 observations having 40 cancer patients. This dataset was in a matrix form while the genes were in the row and observations were in the column. The matrix of gene expression matrix was utilized for global system which was discussed in Figure 6.1.

6.2.2 Gene Expression Data Normalization

Data normalization is needed to avoid bias and redundancy of the gene expression data. The normalization formula is given, as below:

$$Z=\frac{\mathbf{X-\mu}}{\boldsymbol{\sigma}} \tag{6.1}$$

Where, **X** is the variable to be normalized, **μ** and **σ** is the mean and standard deviation.

6.2.3 Local System for the Machine Learning

The main goal of this study was to predict the high-risk genes based on ML system. Thus, we need to preprocess the input data for better characterization of gene expression. The overall system comprises of 4 statistical tests. The effect of the global system is in depicted Figure 6.1. The training/testing paradigm of the entire ML-based system is presented in Figure 6.2. The 1st step is to split the given dataset into two sets as: training and test. These two parts are split with a dotted line as training and test of gene expression data. The next stage is normalization of data, and then extract the top differential expressed (DE) genes using 4 statistical tests (WCSRS, t, KW, F) based on the p-value. The DE genes are trained based on ML framework. Estimating ML -based training parameters and then applied to test data that is transformed to predict cancer patients. Additionally, 10 ML classifiers as LDA, QDA, NB, GPC, SVM, ANN, LR, DT, AB and RF have been adapted to classify the cancer patients. Monitoring the output of these classifiers are evaluated using ACC, SE, SP, PPV, NPV, FM, and AUC which were displayed in Figure 6.3.

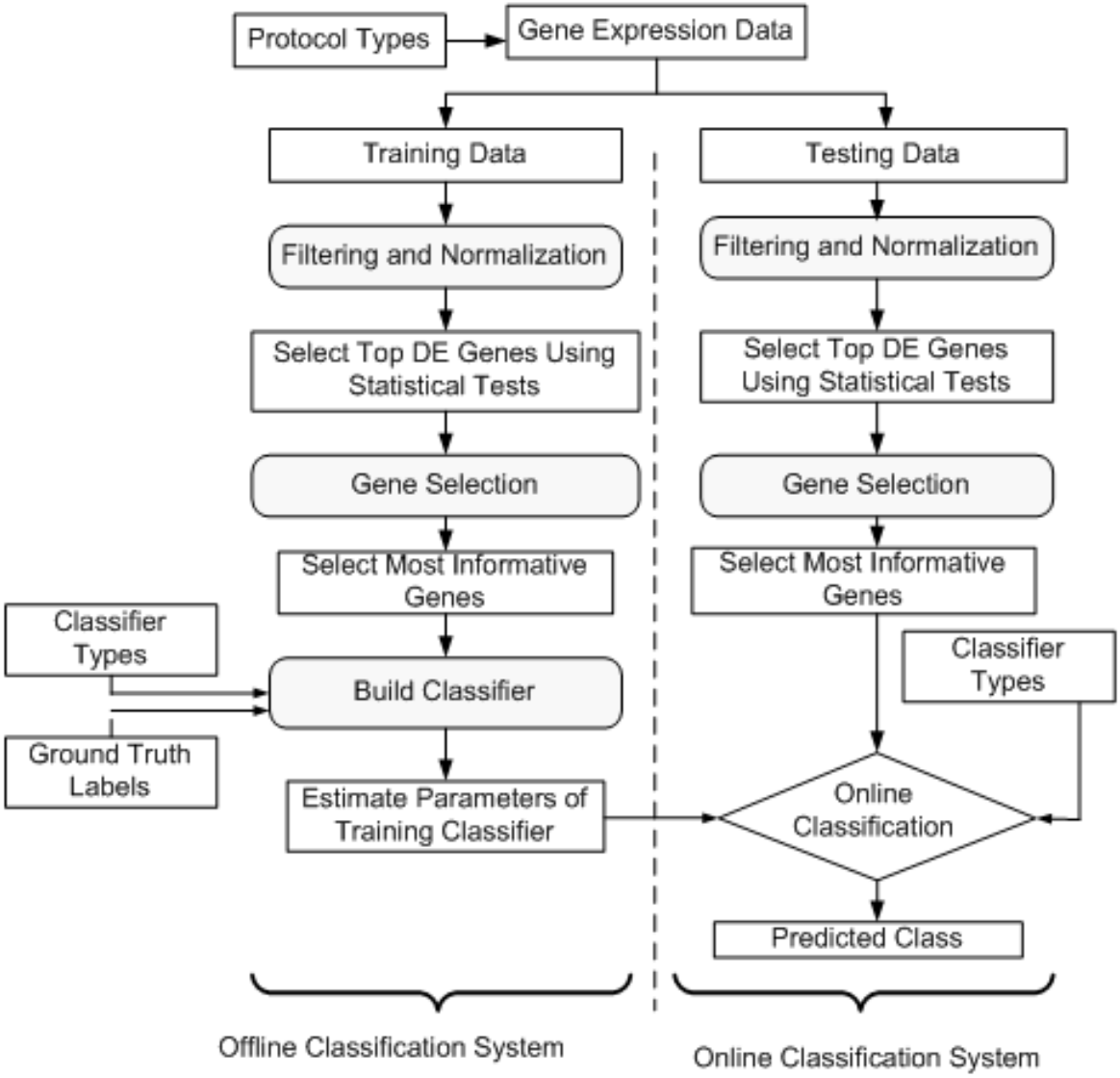


Figure 6.2. Local system for the machine learning.

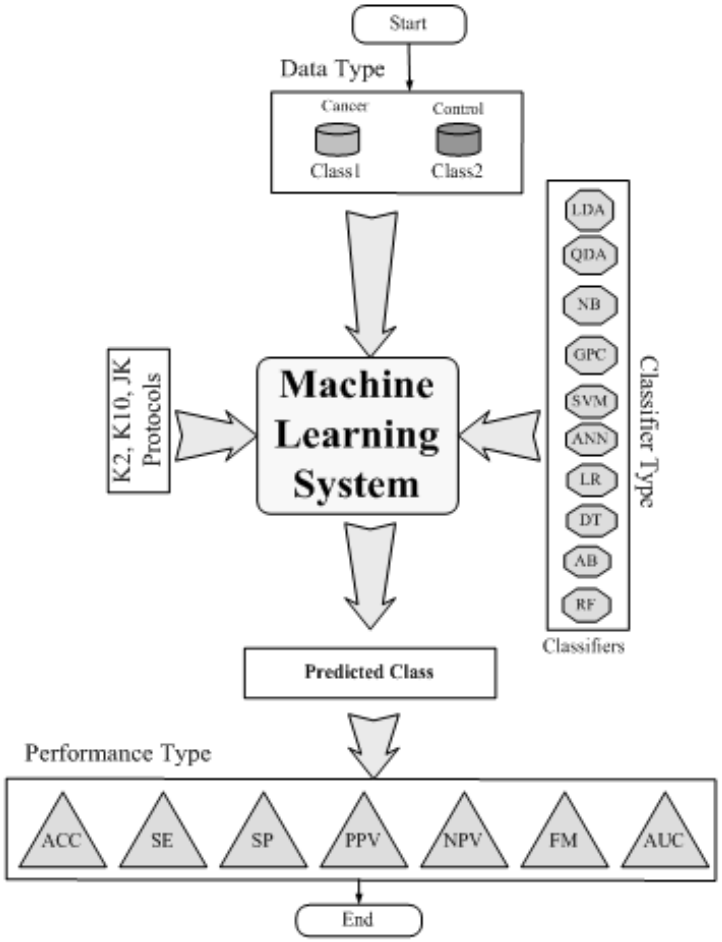


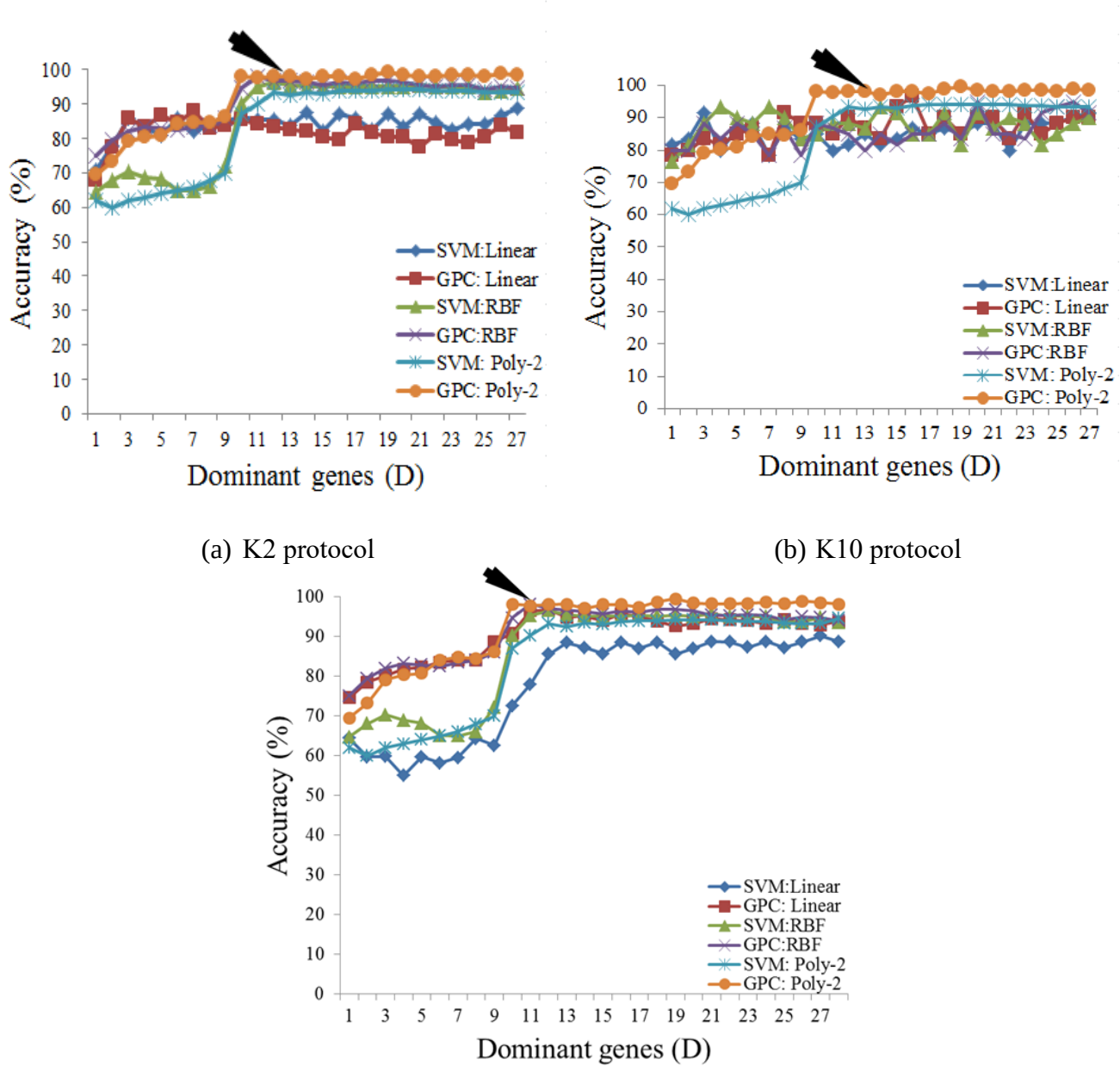
Figure 6.3. Performance of the ML-based framework.

6.3 Results and Discussion

6.3.1 Kernel Optimization

The main objective of this section was to optimize the kernel for SVM and GP-based classifiers during the training. We have used 3 types of kernels, namely linear, RBF, and Poly-2 and optimized the kernel whose gave the highest accuracy. The effect of dominant genes on accuracy was presented in Figure 6.4a for K2 protocol, Figure 6.4b for K10 protocol, and Figure 6.4c for JK protocol, respectively. The corresponding table was shown in Appendix A3: Table A3.1. The figures demonstrated the accuracy is improved with increasing in the dominant genes (D). These figures also demonstrated that the highest ACC is obtained with decreasing p-values by GP with Poly-2 based classifier. It was observed that the highest ACC was also obtained by SVM with RBF kernel.

However, classification accuracy of all classifiers has more variations for K10 protocol compared to K2 protocol due to small sample size. So we have chosen that RBF kernel for SVM and Poly-2 for GP-based classifier. The mean of the ACC over the dominant genes (D) for 3 kernels were depicted in Appendix A3: Table A3.1.



(a) K2 protocol (b) K10 protocol

(c) JK protocol

Figure 6.4. Kernels selection over 3 CV protocols: (a) K2, (b) K10, and (c) JK.

6.3.2 Effect of p-value on ML Performance

The most informative genes were extracted using 4 statistical cutoff of point of p-values as 0.05, 0.01, 0.001, and 0.0001, respectively. It was noticed that the no. of significant genes are reduced with reducing the cutoff point of p-values as displayed in Figure 6.5. It was also noticed that the ACC of 10 classifiers with 4 statistical tests was higher for p-value less than 0.0001 as displayed in Appendix A3: in Table A3.2, Table A3.3, Table A3.4 and Table 6.1.The ACC of 10 classifiers were increased for 4 statistical tests (Table 6.1). It was noted that the highest ACC was provided for 3 protocols with lowest number of genes by RF-based classifier (see Table 6.1) and the corresponding figures were designated in Appendix A3: Figure A3.1. It was also noted that the ACC of 10 classifiers were increased with increasing the training dataset with protocols from K2 to K10 to JK. In addition, the classification ACC were improved for each protocol as decrease the p-values. It is provided the evidence of classifiers trained well for the most significant genes with less noise.

Table 6.1. Mean accuracy (%) of 10 classifiers and p-values of WCSRS test over 3 protocols.

Protocol types	p-values	# of genes	Classifier types									
			LDA	QDA	NB	GPC	SVM	ANN	LR	DT	AB	RF*
K2	0.05	387	80.01	56.77	73.70	82.19	82.58	70.48	74.35	66.45	75.16	**84.56**
	0.01	194	79.68	62.74	75.97	82.74	82.97	74.35	73.39	72.58	79.36	**85.97**
	0.001	64	78.55	64.35	82.42	85.00	85.96	76.77	74.52	75.32	78.38	**85.32**
	0.0001	27	59.51	71.94	85.16	86.45	85.00	77.90	76.45	76.29	80.97	**85.80**
K10	0.05	387	80.83	75.83	76.67	88.33	87.50	71.05	78.33	74.16	82.91	**88.33**
	0.01	194	75.83	71.66	80.83	82.50	88.33	73.43	74.99	78.33	75.00	**92.50**
	0.001	64	72.83	73.33	80.00	85.00	90.00	70.63	72.08	77.50	87.92	**93.49**
	0.0001	27	75.83	75.83	81.67	87.50	87.50	70.00	71.25	78.34	81.84	**95.05**
JK	0.05	387	96.48	64.86	76.10	93.39	93.45	78.63	99.45	99.74	96.46	**99.79**
	0.01	194	96.43	66.65	82.60	92.54	93.29	79.87	99.74	99.69	96.35	**99.82**
	0.001	64	99.56	67.74	83.82	91.99	93.39	77.68	99.58	99.74	94.77	**99.77**
	0.0001	27	95.34	84.60	84.05	92.03	91.75	90.66	89.40	94.85	95.76	**99.81**

*Bold values indicate the result of proposed method.

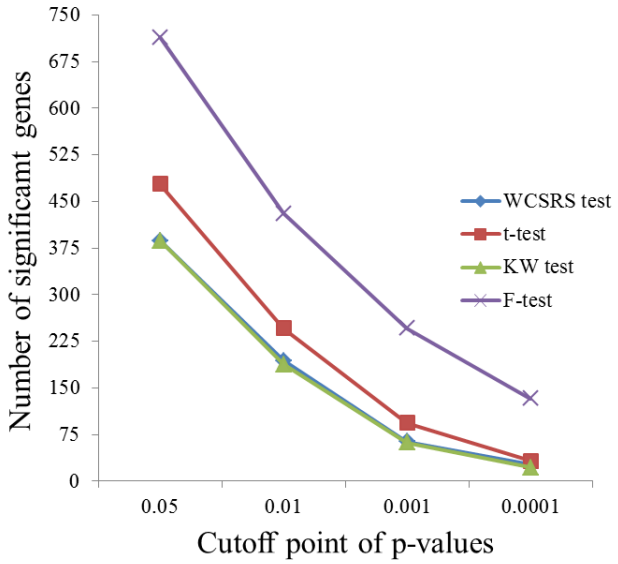


Figure 6.5. The Relation between p-value's cutoff point and the no. of genes.

6.3.3 Inter-comparison of the Classifiers

The accuracy of 10 classifiers with 4 statistical tests and 3 protocols, a total of 120 (10×4×3=120) combination with keeping data size fixed (n=62) is presented in Table 6.2. It was illustrated that the maximum ACC of **99.81%** was obtained by the WCSRS-RF-based classifier. It was also illustrated that the minimum ACC was obtained by NB followed by QDA. Five performance evaluation parameters, likely SE, SP, PPV, NPV, and AUC of WCSRS-RF-based classifier as described in Table 6.3. It was also confirmed that better performance was provided by the WCSRS-RF-based classifier and it was also validated based on AUC (see last column of Table 6.3).

6.3.4 Effect of Dominant Genes

The effect of dominant genes on ACC of WCSRS-RF-based model was displayed in Figure 6.6. It was demonstrated that the ACC of 10 classifiers is improved with increasing the number of genes. So, it was concluded that the best performance was given by RF-based classifiers.

Table 6.2. Accuracy (in %) of 4 tests along with 10 classifiers over 3 protocols.

Protocol types	Classifier types	Statistical tests			
		WCSRS test	t-test	KW test	F-test
K2	LDA	63.87	61.45	70.81	78.39
	QDA	85.83	72.90	84.03	70.00
	NB	85.97	80.00	83.39	75.81
	GPC	86.94	85.29	86.16	83.23
	SVM	85.16	84.35	84.68	77.42
	ANN	80.81	78.71	80.81	82.10
	LR	57.74	57.42	66.77	46.94
	DT	71.29	72.26	74.19	72.58
	AB	79.35	78.87	82.26	81.45
	RF*	**88.06**	**86.26**	**90.68**	**85.65**
K10	LDA	74.58	64.17	80.00	75.00
	QDA	87.50	79.17	85.42	77.50
	NB	85.42	78.75	82.50	77.50
	GPC	89.29	86.58	86.25	83.58
	SVM	84.17	85.42	87.92	78.33
	ANN	76.25	78.75	82.08	78.33
	LR	71.67	64.58	74.69	54.37
	DT	75.83	80.00	79.58	77.92
	AB	74.58	82.92	79.58	82.50
	RF*	**95.50**	**89.83**	**92.92**	**89.33**
JK	LDA	95.34	91.31	89.36	96.54
	QDA	84.60	67.82	80.72	74.14
	NB	84.05	80.67	84.08	83.92
	GPC	92.03	92.26	91.96	94.48
	SVM	90.66	91.75	89.00	90.48
	ANN	89.40	82.90	92.10	90.46
	LR	91.78	89.98	90.37	88.76
	DT	94.85	91.70	94.85	96.41
	AB	95.76	95.86	94.95	95.73
	RF*	**99.81**	**99.72**	**99.50**	**99.74**

*Bold and shaded values indicate the result of the proposed method.

Table 6.3. Performance evaluation parameters of 10 classifiers for WCSRS test over 3 protocols.

Protocol types	Classifiers types	Evaluation parameters (%)					
		SE	SP	PPV	NPV	FM	AUC
K2	LDA	65.56	61.29	76.95	49.12	69.75	67.42
	QDA	88.72	84.99	88.94	81.77	87.84	88.79
	NB	83.78	91.18	94.59	73.94	88.58	92.12
	GPC	90.83	80.15	89.23	82.23	89.91	87.34
	SVM	90.09	77.09	87.21	81.80	88.45	47.36
	ANN	84.45	75.40	84.94	74.40	84.30	88.54
	LR	60.84	52.42	69.98	43.35	64.06	58.03
	DT	76.81	63.79	79.80	66.19	76.40	75.80
	AB	85.47	71.46	84.03	75.80	83.84	89.13
	RF*	**92.78**	**80.08**	**89.88**	**85.75**	**91.03**	**93.00**
K10	LDA	82.74	60.58	81.23	61.00	81.08	82.02
	QDA	91.17	84.32	88.55	85.54	89.18	90.98
	NB	82.62	89.42	92.07	73.04	87.88	94.14
	GPC	91.92	82.32	90.06	87.82	91.68	88.73
	SVM	86.08	81.25	90.68	72.65	87.85	36.02
	ANN	80.99	68.93	81.26	66.86	80.46	85.53
	LR	77.33	64.83	78.70	62.65	76.80	75.34
	DT	82.82	64.60	81.43	68.81	80.61	75.53
	AB	82.98	62.60	74.44	76.40	76.58	82.66
	RF*	**93.08**	**87.29**	**92.94**	**84.00**	**92.22**	**97.78**
JK	LDA	95.44	95.16	97.31	92.05	96.36	99.51
	QDA	89.11	76.39	87.29	79.46	88.19	89.57
	NB	80.28	90.91	94.14	71.73	86.66	93.99
	GPC	92.98	90.30	94.59	87.65	93.77	94.33
	SVM	92.50	87.32	93.00	86.49	92.75	95.81
	ANN	93.22	82.46	90.75	87.39	91.90	95.18
	LR	89.96	95.09	97.11	83.95	93.39	96.20
	DT	97.22	90.54	94.94	94.78	96.05	98.88
	AB	94.60	97.87	98.80	90.95	96.64	99.47
	RF*	**99.84**	**99.75**	**99.87**	**99.72**	**99.85**	**99.95**

*Bold and shaded values indicate the result of the proposed method.

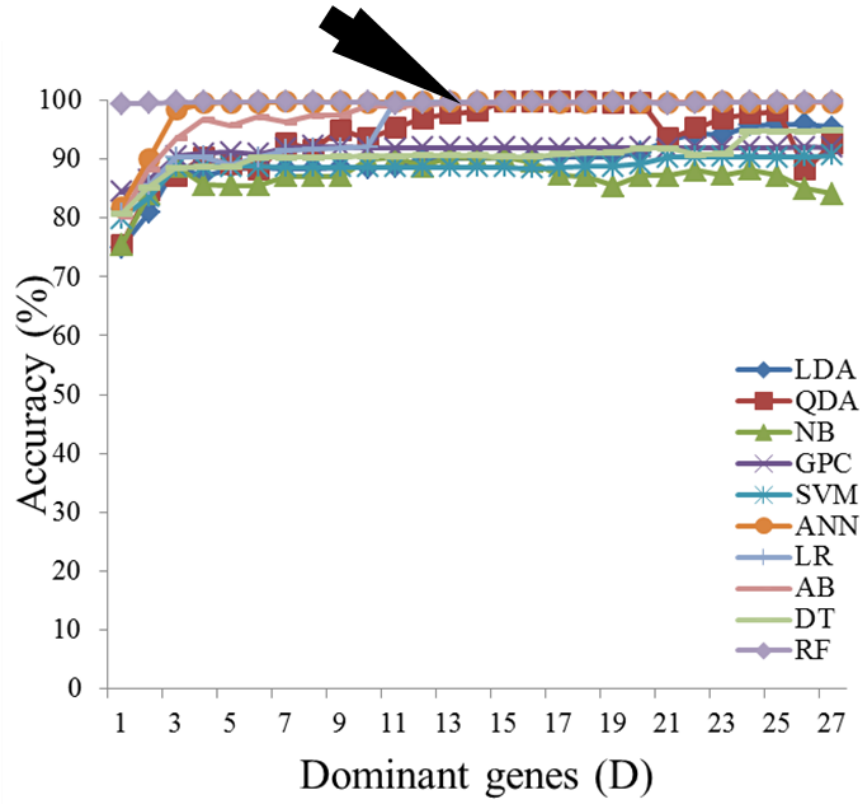


Figure 6.6. Accuracy vs. dominant genes of 10 classifiers for WCSRS test (p-value=0.0001).

6.3.5 Effect of Data size on Memorization vs. Generalization

This experiment presented the effect of varying data size (n)on accuracy of ML-based models. We have divided the training data size into 10 parts as10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. The corresponding 10 datasets were comprised of 6, 12, 19, 25, 31, 37, 43, 50, 56 and 62 patients. 10 ML-based classifier was implemented on training data size and calculates the classification accuracy. It was also noticed that at least 50% (32 patients) are needed for the net generalization. Figure 6.7 showed the change of accuracy with varying data size. It was noticed that accuracy of 10 classifiers were increased with increasing data size and the highest performance was provided by RF-based classifier (see Figure 6.7). We have computed system ACC by averaging ACC's over data size (n) for 3 CV protocols. Table 6.4 indicates that the highest ACC was supported by RF-based classifier for 3 protocols compared to others.

Table 6.4. Systems mean accuracy (%) of 10 classifiers over 3 CV protocols.

Protocol types	Classifier types									
	LDA	QDA	NB	GPC	SVM	ANN	LR	DT	AB	**RF***
K2	68.63	78.19	81.29	85.41	82.90	57.22	72.58	72.58	80.48	**87.66**
K10	73.44	82.40	86.43	78.85	66.33	78.33	79.90	91.90	93.14	**93.14**
JK	98.68	83.95	87.60	95.25	94.14	94.89	97.18	93.65	93.35	**99.77**

* Bold and shaded values indicate the result of proposed method.

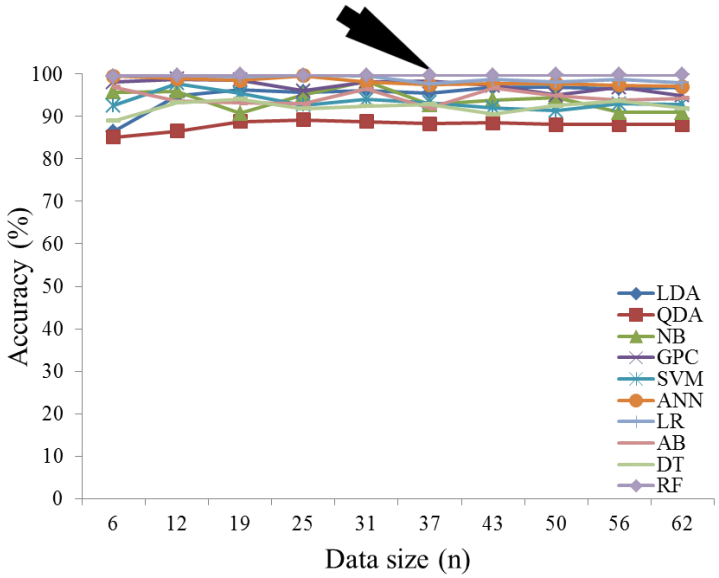


Figure 6.7. Effect of varying data size (n) on classification accuracy.

6.4 Performance Evaluation and Hypothesis Validations

6.4.1 Gene Separation Index

Gene separation index (nGSI) can be depicted the segregation power of the genes and it is mathematically represented as

$$c=|F_n-F_d| \quad (6.2)$$

Where,F_n and F_d is the mean value of the cancer patients and control. The relationship between nGSI, system mean accuracy along with p-values was presented in Table 6.5 and also effect of p-value on (nGSI) is presented in Figure 6.8.

Table 6.5. Relationship between nGSI and system mean accuracy.

p-value	Cancer	Control	nGSI	Accuracy
0.05	580.50 ± 386.42	456.66 ± 279.65	123.84	89.84
0.01	644.28 ± 439.74	512.32 ± 315.62	131.96	90.70
0.001	775.04 ± 497.78	614.40 ± 371.92	160.64	90.80
0.0001	1314.90 ± 756.62	1018.71 ± 537.11	296.19	91.83

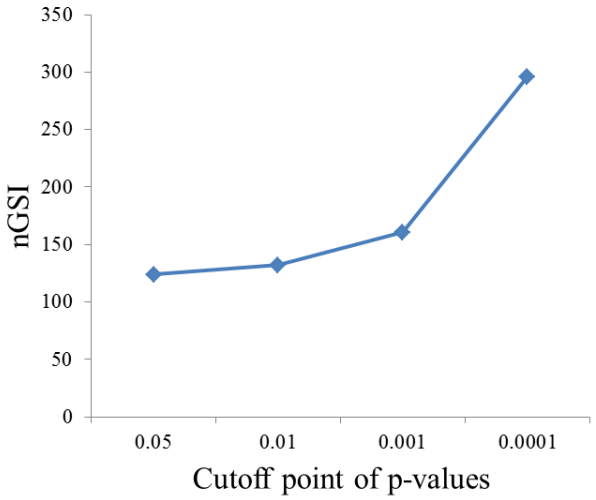


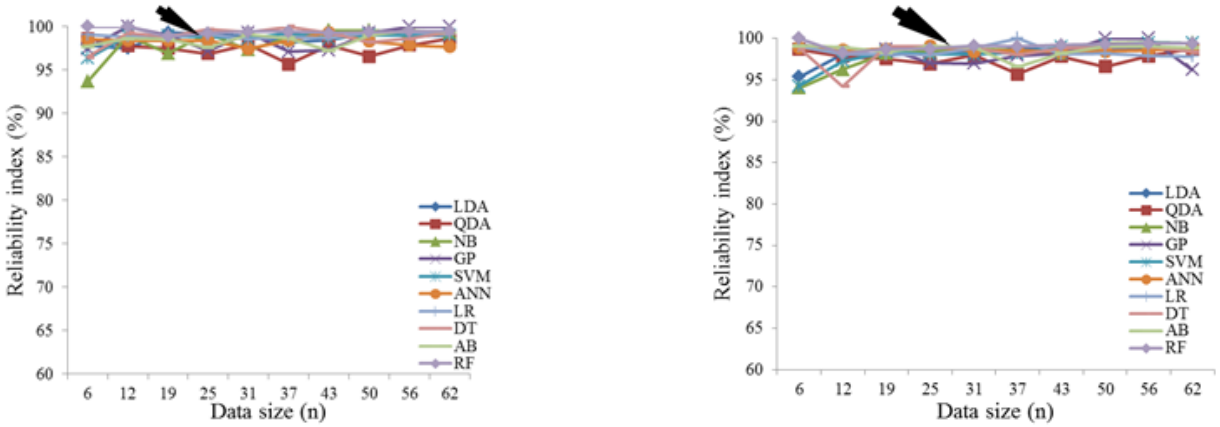
Figure 6.8 . Effect of p- values on (nGSI).

6.4.2 Reliability Index

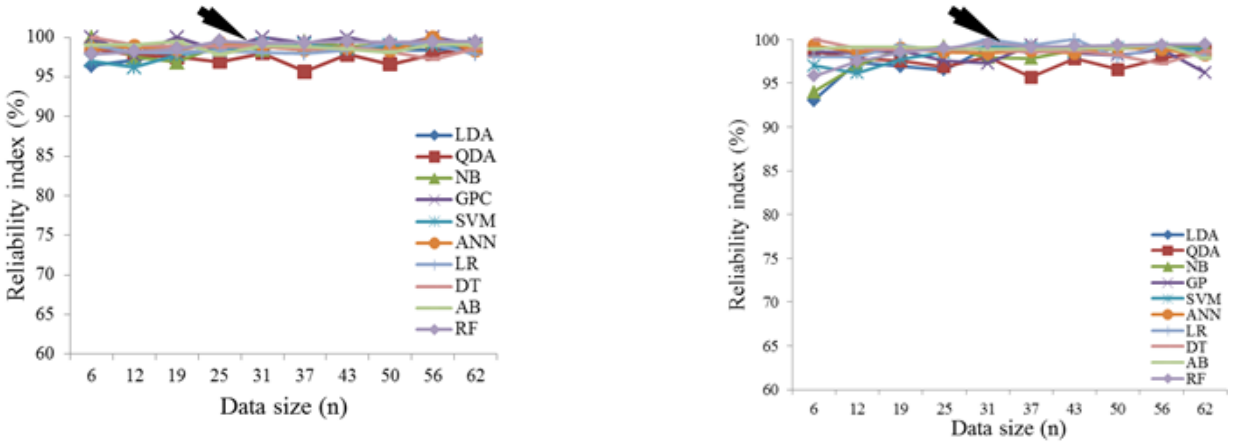
The system reliability index (RI) is used to evaluate the performance of the ML-based system which was computed as follows:

$$\zeta_{n_i} = \left(1 - \frac{\sigma_{n_i}}{\mu_{n_i}}\right) \times 100 \tag{6.3}$$

where, μ_{n_i} and σ_{n_i} are the mean and SD for all combinations of p-values with 4 statistical tests. The system RI is computed by averaging the RI over the data size which is shown in Figure 6.9. It shows the value of RI of 4 statistical tests, 10 classifiers over 3 protocols and also confirms that RF-based classifier is the best compared to others. The corresponding result is shown in Table 6.6.



(a) WCSRS test (b) t-test



(c) KW test (d) F-test

Figure 6.9. Reliability index vs. data size (n) for 10 classifiers for 4 statistical tests: (a) WCSRS test,(b) t- test,(c) KW test, and (d) F- test.

Table 6.6. System reliability index (%) of 4 statistical tests and 10 classifiers over 3 protocols.

PT	Test types	LDA	QDA	NB	GPC	SVM	ANN	LR	DT	AB	RF*
K2	WCSRS	98.91	97.97	97.33	99.29	98.64	97.48	98.55	99.33	99.12	**99.43**
	t	98.37	97.97	98.74	96.94	98.04	98.42	98.86	98.79	98.97	**99.11**
	KW	98.83	97.87	99.45	98.25	99.46	98.55	97.88	99.01	99.04	**99.37**
	F	98.79	97.97	98.47	97.25	99.11	98.86	98.03	98.86	98.97	**99.30**
K10	WCSRS	99.27	96.58	99.59	99.29	99.05	98.31	99.33	98.22	99.02	**99.40**
	t	98.37	97.97	98.74	96.94	98.04	98.42	98.86	98.79	98.97	**99.11**
	KW	98.83	97.87	99.45	97.80	99.46	98.55	97.88	99.01	99.04	**99.37**
	F	98.79	97.97	98.47	98.25	99.11	98.86	98.03	98.86	98.97	**99.30**
JK	WCSRS	98.67	97.55	98.26	98.66	98.67	98.29	98.99	98.91	98.55	**99.46**
	t	98.17	96.50	98.05	98.25	98.11	98.63	98.40	98.30	98.60	**99.09**
	KW	98.22	94.30	98.80	99.31	98.60	98.78	98.42	98.50	98.80	**99.05**
	F	97.83	95.20	98.09	98.17	98.30	98.72	98.89	98.68	98.91	**98.71**

*Bold and shaded values indicate the result of proposed method.

6.4.3 ROC Analysis

The value of AUC along with SE, SP, PPV, NPV, FM of 10 classifiers for WCSRS-RF tests and 3 protocols (K2, K10, and JK) were presented in Table 6.3. The similar results were presented for 3 tests (t, KW, F) in Appendix A3.2: (Table A3.5-Table A3.7) for K2, K10, and JK protocols. It was noted that the highest AUC was obtained by RF classifier for 4 statistical tests over K2, K10, JK protocols (see last column of Table A3.5 to Table A3.7, and Table 6.3). It was observed that the mean AUC of RF along with 4 statistical tests was provided almost close to unity which was also proved the hypothesis.

6.4.4 Validation of WCSRS-RF (Proposed) Method

Breast cancer dataset was used (Patrício et al., 2018) to validate the proposed WCSRS-RF method. 4 statistical test and 10 ML-based systems have been adapted for prediction of breast cancer. Our results indicate that the highest ACC of **95.25%** was provided by WCSRS-RF (see Table 6.7). So, our proposed WCSRS-RF methods was validated for both colon and breast cancer datasets.

Table 6.7. Validation of the WCSRS-RF (proposed) systems using K10 for breast cancer.

CT	WCSRS test	t-test	KW test	F-test
LDA	72.41	70.69	71.55	70.69
QDA	62.07	59.48	58.62	59.48
NB	62.93	62.07	61.21	62.07
GPC	81.76	77.63	80.13	76.81
SVM	80.17	71.55	77.59	71.55
ANN	73.66	74.91	79.96	73.66
LR	74.14	74.14	74.40	74.14
DT	82.76	80.17	81.90	80.17
AB	87.07	87.93	89.66	87.93
RF*	**95.25**	**94.31**	**94.40**	**93.79**

* Bold and shaded values indicate the result of proposed method.

6.5 Summary of the Chapter

In this section, **Chapter 6** is summarized as:

1. Data:
   - Data extraction: Extract colon cancer dataset format from PubMed.
   - Data normalization: It was needed to reduce the redundancy of the data.
   - Top gene extraction: Extract top high-risk genes of colon cancer using 4 tests like WCSRS, t, KW, and F-test.
2. Modeling:
   - CV protocol: 3 CV-based (K2, K10, and JK) protocols were used.
   - Kernel Optimization: Optimize kernel of SVM & GPC for prediction of colon cancer.
   - Model selection: Apply LDA, QDA, NB, GPC, SVM, ANN, LR, DT, AB, and RF for prediction of cancer.
3. Model performance evaluation:
   - Metrics: Use ACC, SE, PPV, NPV, FM, and AUC for classifiers evaluation.
   - Interpretation: Performance evaluation metrics to compare the classifiers and findings showed that RF-based classifier gave **99.81%** ACC for JK protocol.

Chapter 7 Conclusion and Areas of Further Research

7.1 Conclusion

In this research we have tried to show the comparison of the performance of ML-based systems in healthcare datasets. We have used three healthcare datasets, among them two datasets on diabetes (NHAMES and Pima Indian) and another one on colon cancer. For NAHNES diabetes dataset, our hypothesis is to propose an ML-based system combine with LR-based FS method and a classifier out 4 as NB, DT, AB, and RF along with 3 CV protocols (K2, K5, and K10) that will yield the highest accuracy. Results demonstrated that ML-based model with the combination of LR-based FS method and RF-based classifier obtained the highest ACC (**94.25%)** for K10 protocol.

For PID dataset, our hypothesis was that if missing values and outliers are replaced by group median and such a data when used in ML-based framework using RF-RF combination for a FS and a classifier should yield higher accuracy. We demonstrate our hypothesis by showing a improvement and reaching an accuracy of nearly 99.99~100% in JK-based CV protocol. Comprehensive data analysis is conducted consisting of 10 classifiers, 6 FS methods, and 5 set of CV protocols, 2 outlier's removal techniques.

For cancer dataset, this study presented an exhaustive evaluation of ML-based systems which has two major components: (a) identification of high-risk differential genes using tests and (b) development of a ML-based strategy for predicting the cancerous genes. 4 statistical tests as WCSRS test, t-test, KW test, and F-test are adapted for cancerous gene identification based on p-values. Further, 10 ML-based systems are designed using ten different classifiers as LDA, QDA, NB, GPC, SVM, ANN, LR, DT, AB, and RF. Our overalls mean accuracy of ML-based system using 4 tests and 10 classifiers was **90.50**%. The highest ACC of **99.81%** was obtained by adapting the combination of WCSRS test along with RF-based classifier. Finally, we may conclude that RF-based classifier performed better for both diabetes and cancer datasets.

7.2 Areas of Further Research

- In this study, I have used only median-based imputation method to impute missing values. One can easily extended different missing value imputation methods as expectation maximization (EM) algorithm, KNN, fuzzy K-means (FKM), etc.

- In this study, I have used only few FS methods to detect the high-risk factors/biomarker for healthcare disease. In future, I shall be used various feature extraction and FST's like singular value decomposition (SVD),correspondence analysis (CA), canonical correlation analysis (CCA), partial least square (PLS), SVM, GPC, pooling based FDR and so on.

- I have used only four statistical tests to select the top biomarkers of colon cancer disease. However, there were more statistical tests (parametric and non-parametric) available in literatures. In future, I shall be used more applicable and important statistical tests to identify the accurate biomarkers of any types of complex disease.

- Although the current work was focus on only ML, showing the role of detection , and prediction, one can extend this to adapt deep learning (DL)-based paradigm for detection and prediction of the segmentation of medical imaging, single RNA-sequencing data, and protein-protein interactions (PPI) data etc., and compare with our current study.

- Although this work was the application of ML-based techniques on diabetes and cancer disease and no mathematical/theoretical model was introduced or modified. In future, I will develop a new computational model to identify the biomarkers of complex disease. I shall also develop a novel ML-based system for addressing and predicting any complex types of disease.

Bibliography

Ahuja, R., Vivek, V., Chandna, M., Virmani, S., & Banga, A. (2019). Comparative study of various machine learning algorithms for prediction of Insomnia. *Inf Reso Manag Associ,* 234-257.

Alladi, S. M., Shinde Santosh, P., Ravi, V., & Murthy, U. S. (2008). Colon cancer prediction with genetic profiles using intelligent techniques. *Bioinformation,* 3 (3), 130-133.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci,* 96 (12), 6745-6750.

American Diabetes Association (2010). Diagnosis and classification of diabetes mellitus. *Diabetes Care,* 33 (Supplement 1), S62-S69.

Ando, T., Suguro, M., Kobayashi, T., Seto, M., & Honda, H. (2003). Selection of causal gene sets for lymphoma prognostication from expression profiling and construction of prognostic fuzzy neural network models. *J Biosci Bioeng,* 96 (2), 161-167.

Ashour, A. S., Hawas, A. R., & Guo, Y. (2018). Comparative study of multiclass classification methods on light microscopic images for hepatic schistosomiasis fibrosis diagnosis. *Health Inf Sci Syst,* 6 (1), 7.

Banchhor, S. K., Londhe, N. D., Araki, T., Saba, L., Radeva, P., Khanna, N., & Suri, J. S. (2018). Calcium detection, its quantification, and grayscale morphology-based risk stratification using machine learning in multimodality big data coronary and carotid scans: A review. *Comput Biol Med,* 101, 184-198.

Baneshi, Mohammad Reza, and A. R. Talei (2012). Does the missing data imputation method affect the composition and performance of prognostic models? *Iranian Red Crescent Med J,* 14 (1), 30- 31.

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *EEE Trans Inf Technol Biomed,* 14 (4), 1114-1120.

Bashir, S., Qamar, U., & Khan, F. H. (2016). IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. *J Biomed Inform,* 59, 185-200.

Bauder, R. A., & Khoshgoftaar, T. M. (2018). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health Inf Sci Syst,* 6 (1), 9.

Bharath, C., Saravanan, N., & Venkatalakshmi, S. (2017). Assessment of knowledge related to diabetes mellitus among patients attending a dental college in Salem city-A cross sectional study. *Braz Dent Sci*, 20 (3): 93-100.

Bozkurt, M. R., Yurtay, N., Yilmaz, Z., & Sertkaya, C. (2014). Comparison of different methods for determining diabetes. *Turkish J Electr Eng Compt Sci*, 22 (4), 1044-1055.

Brahim-Belhouari, S., & Bermak, A. (2004). Gaussian process for nonstationary time-series prediction. *Comput Stat Data Ana*, 47 (4), 705-712.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 68 (6), 394-424.

Breiman & Leo. (2001). Random forests. *Mach Learn*, 45, 5-32.

Chen, D., Liu, Z., Ma, X., & Hua, D. (2005). Selecting genes by test statistics. *Biomed Res Int*, 2005 (2), 132-138.

Chen, Z., Li, J., & Wei, L. (2007). A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artif Intell Med*, 41 (2), 161-175.

Cokluk, O., Kayri, M. (2011). The effects of methods of imputation for missing values on the validity and reliability of scales. *Educ Sci Theo Pract*, 11 (1), 303-309.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach Learn*, 20 (3), 273-297.

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*, 2, 1-19.

Danaei, G., Finucane, M. M., Lu, Y., Singh, G. M., Cowan, M. J., Paciorek, C. J., & Rao, M. (2011). National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet*, 378 (9785), 31-40.

Deniz, E., Şengür, A., Kadiroğlu, Z., Guo, Y., Bajaj, V., & Budak, Ü. (2018). Transfer learning based histopathologic image classification for breast cancer detection. *Health Inf Sci Syst*, 6 (1), 18.

Elssied, N. O. F., Ibrahim, O., & Osman, A. H. (2014). A Novel feature selection based on one-way ANOVA F-test for e-mail spam classification. *Res J Appl Sci Eng Tech*, 7 (3), 625-638.

Ertin, E., Stohs, N., Kumar, S., Raij, A., Al'Absi, M., & Shah, S. (2011). Auto Sense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. *Proc 9th ACM Conf Emb Net Sensor Sys,* 274-287.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., Bray, F. (2020). Global Cancer Observatory: Cancer Today. Lyon, France: *Int Agency Res Cancer*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann Eug*, 7 (2), 179-188.

Giovannucci, E., Harlan, D. M., Archer, M. C., Bergenstal, R. M., Gapstur, S. M., Habel, L. A., Pollak, M., Regensteiner, J. G., & Yee, D. (2010). Diabetes and cancer: a consensus report. *Diabetes Care*, 33 (7), 1674–1685.

Gruenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., & Lukowicz, P. (2014). Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. *Proc 5th Augm Hum Int Conf*, 1-8.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach Learn*, 46 (3), 389-422.

Hasan, A.M., M., Nasser, M., & Pal, B. (2013). On the KDD'99 dataset: support vector machine based intrusion detection system (ids) with different kernels. *Int J Electron Commun Comput Eng*,4 (4), 1164-1170.

Hasan, M. A. M., Nasser, M., Ahmad, S., & Molla, K. I. (2016). Feature selection for intrusion detection using random forest. *J Inform Secur*, 7 (3), 129-140.

Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A., Wolf, M., & Kallioniemi, O. P. (2003). Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps. *Mach Learn,* 52 (2), 45-66.

Hollstein, M., Sidransky, D., Vogelstein, B., & Harris, C. C. (1991). 53 mutations in human cancers. *Sci*, 253 (5015), 49-53.

Hong, J. H., & Cho, S. B. (2006). The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. *Artif Intell Med*, 36 (1), 43-58.

Hu, W., Hu, W., & Maybank, S. (2008). Adaboost-based algorithm for network intrusion detection. *IEEE Trans Syst Man Cybern B Cybern*, 38 (2), 577-583.

Huang, C. L., & Wang, C. J. (2006). A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst Appl*, 31 (2), 231-240.

Iancu, I., Mota, M., & Iancu, E. (2008). Method for the analysing of blood glucose dynamics in diabetes mellitus patients. *IEEE Int Conf Autom, Qual Testing Robot,* 3, 60-65.

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *Int J Data Min Know Manag Proce,* 5 (1), 1-14.

Jain, A. K., Duin, R. P., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Trans Pattern Anal Mach Intell*, 22 (1), 4-37.

Jeanmougin, M., De Reynies, A., Marisa, L., Paccard, C., Nuel, G., & Guedj, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies? *PloS One*, 5 (9), 12336-12345.

Karthikeyani, V., & Begum, I. P. (2013). Comparison a performance of data mining algorithms in prediction of diabetes disease. *Int J Comput Sci Eng*,5 (3), 205-210.

Karthikeyani, V., Begum, I. P., Tajudin, K., & Begam, I. S. (2012). Comparative of data mining classification algorithm in diabetes disease prediction. *Int J Comput Appl*, 60 (12), 26-31.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *J Healthc Inf Manag,* 19(2), 65

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*, 13, 8-17.

Krasteva, A., Panov, V., Krasteva, A., Kisselova, A., & Krastev, Z. (2011). Oral cavity and systemic diseases-diabetes mellitus. *Biotech Equip,* 25 (1), 2183-2186.

Kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *Int J Eng Appl Sci*, 2 (5), 145-148.

Kumar, P. G., Victoire, T. A. A., Renukadevi, P., & Devaraj, D. (2012). Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Expert Syst Appl*, 39 (2), 1811-1821.

Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *Int J Eng Res Appl,* 3 (2), 1797-1801.

Kuyuk, S., & Ercan, I. (2017). Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostat Epidemiol Int J,* 1 (1), 1-8.

Lee, C. H., & Yoon, H. J. (2017). Medical big data: promise and challenges. *Kidney Res Clin Pract*, 36(1), 3.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol*, 49 (4), 764-766.

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17 (12), 1131-1142.

Liu, Z. P., & Gao. (2018). Detecting pathway biomarkers of diabetic progression with differential entropy. *J Biomed Inform*, 82, 143-153.

Lonappan, A., Bindu, G., Thomas, V., Jacob, J., Rajasekaran, C., & Mathew, K. T. (2007). Diagnosis of diabetes mellitus using microwaves. *J Electr Wav Appl*, 21 (10), 1393-1401.

Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C. J. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, 367, 1747–1757

Manikandan, S. (2011). Measures of dispersion. *J Pharmacol Pharmacother*, 2 (4), 315-316.

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Comput Methods Programs Biomed*, 152, 23-34.

Mao, Y., Zhou, X., Pi, D., Sun, Y., & Wong, S. T. (2005). Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. *Biomed Res Int*, 2005 (2), 160-171.

Matthias EF, Anthony R, Nikola K. (2003). Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. *Artif Intell Med*, 28 (2), 165-189.

Mohapatra, S. K., Swain, J. K., & Mohanty, M. N. (2019). Detection of diabetes using multilayer perceptron. *Int Conf Intell Comput Appl*, 109-116.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn*, 52 (1), 91-118.

Muaremi, A., Arnrich, B., & Tröster, G. (2013). Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNano Sci*, *3*(2), 172-183.

Nabi, M., Wahid, A., & Kumar, P. (2017). Performance analysis of classification algorithms in predicting diabetes. *Int J Advan Res Comput Sci*, 8 (3), 456-461.

Nahm, F. S. (2016). Nonparametric statistical tests for the continuous data: the basic concept and the practical use. *Korean J Anesthesiol,* 69 (1), 8-14.

Nathan, D. M. (1993). Long-term complications of diabetes mellitus. *N Engl J Med*, 328 (23), 1676-1685.

Osmani, V., Maxhuni, A., Grünerbl, A., Lukowicz, P., Haring, C., & Mayora, O. (2013). Monitoring activity of patients with bipolar disorder using smart phones. *Proc Int Conf Advan in Mobile Comput Multime,* 85-92.

Parashar, A., Burse, K., & Rawat, K. (2014). A Comparative approach for Pima Indians diabetes diagnosis using lda-support vector machine and feed forward neural network. *Int J Advan Res Comput Sci Soft Eng*, 4 (4), 378-383.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*. 18 (1), 29-36.

Pei, D., Zhang, C., Quan, Y., & Guo, Q. (2019). Identification of potential type II diabetes in a chinese population with a sensitive decision tree approach. *J Diabetes Res,* 1-7.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell,* 27 (8), 1226-1238.

Quinlan, J. R. (1986). Induction of decision trees. *Mach Learn,* 1 (1), 81-106.

Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2012). A critical comparative study of liver patients from USA and India: an exploratory analysis. *Int J Comput Sci Issu,* 9 (3), 506.

Rasmussen, C.E. (2004). Gaussian processes in machine learning. *Advan Lect Mach Learn,* 3176, 63-71.

Risk, N. C. D. (2016). Factor Collaboration (NCD-RisC). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *Lancet,* 387 (10026), 1377-1396.

Robertson, G., Lehmann, E. D., Sandham, W., & Hamilton, D. (2012). Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. *J Electr Comput Eng*, 2011, 2-13.

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., & IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and

projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res Clin Pract*, 157, 107843.

Sarwar, N., Gao, P., & Seshasai, S. R. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease. *Lancet*, 375 (9733), 2215-2222.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Rev Educ Res*, 60 (1), 91-126.

Semerdjian, J., & Frank, S. (2017). An ensemble classifier for predicting the onset of type II diabetes. *arXiv preprint arXiv*. 1708.07480.

Shah, S., Luo, X., Kanakasabai, S., Tuason, R., & Klopper, G. (2019). Neural networks for mining the associations between diseases and symptoms in clinical notes. *Health Inf Sci Syst*, 7 (1), 1.

Shakeel, P. M., Baskar, S., Dhulipala, V. S., & Jaber, M. M. (2018). Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. *Health Inf Sci Syst*, 6 (1), 16.

Shen, Q., Shi, W. M., & Kong, W. (2008). Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. *Comput Biol Chem*, 32 (1), 53-60.

Shi, Y., Chinnaiyan, A. M., & Jiang, H. (2015). rSeqNP: a non-parametric approach for detecting differential expression and splicing from RNA-Seq data. *Bioinformatics*, 31 (13), 2222-2224.

Shrivastava, V. K., Londhe, N. D., Sonawane, R. S., & Suri, J. S. (2016). Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: a first comparative study of its kind. *Comput Methods Programs Biomed*, 126 (2), 98-109.

Shrivastava, V. K., Londhe, N. D., Sonawane, R. S., & Suri, J. S. (2017). A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification. *Comput Methods Programs Biomed*, 150 (2), 9-22.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. *17th IEEE Int Conf Mach Learn Appl*, 1394-1401.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia comput sci*, 132, 1578-1585.

Sivanesan, R., & Dhivya, K. D. R. (2017). A Review on diabetes mellitus diagnoses using classification on Pima Indian diabetes data set. *Int J Advan Res Comput Sci Manag Stud*, 5 (1), 12-17.

Srivastava, S., K., Singh, S. K., Suri J. S. (2018). Healthcare Text Classification System and its Performance Evaluation: A Source of Better Intelligence by Characterizing Healthcare Text. *J Med Syst*, 42 (5), 97.

Sun, G., Dong, X., & Xu, G. (2006). Tumor tissue identification based on gene expression data using DWT feature extraction and PNN classifier. *Neurocomputing*, 69 (6), 387-402.

Tabaei B, Herman W (2002). A Multivariate logistic regression equation to screen for diabetes. *Diabetes Care*, 25, 1999–2003.

Takahashi, H., Masuda, K., Ando, T., Kobayashi, T., & Honda, H. (2004). Prognostic predictor with multiple fuzzy neural models using expression profiles from DNA microarray for metastases of breast cancer. *J Biosci Bioeng*, 98 (3), 193-199.

Tung, W. L., & Quek, C. (2005). GenSo-FDSS: a neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data. *Artif Intell Med*, 33 (1), 61-88.

Tzourakis & Melissa (1996). The Healthcare Industry and Data Quality. *Int Conf Inf Qual*.

Vanitha, C. D. A., Devaraj, D., & Venkatesulu, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection, *Procedia Comput Sci*, 47, 13-21.

Webb, G. I., Boughton, J. R., & Wang, Z. (2005). Not so naïve Bayes: aggregating one dependence estimators. *Mach Learn*, 58 (1), 5-24.

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., & Cheng, C. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1 (2), 133-143.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak*, 10 (1), 16-23.

Zainuri, Nuryazmin Ahmat, Abdul Aziz Jemain, and Nora Muda (2015). A Comparison of various imputation methods for missing values in air quality data. Sains Malaysiana, 44 (3), 449-456

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.

Zhao, X., Zou, Q., Liu, B., and Liu, X. (2014). Exploratory predicting protein folding model with random forest and hybrid features. *Curr Proteomics.* 11, 289–299.

Zimmet, P., Alberti, K. G., Magliano, D. J., & Bennett, P. H. (2016). Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. *Nat Rev Endocrinol*, 12 (10), 616.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Front Genet*, 9 (515), 1-10.

Appendix

Appendix A1

Table A1.1 System accuracy (%) of 4 classifiers varying data sizes for K2 protocol.

Data size	Classifier types			
	NB	DT	AB	RF
656	84.75	85.34	85.25	87.82
1312	87.20	88.66	89.82	91.46
1968	87.48	90.30	90.00	91.57
2624	87.16	89.80	90.37	92.26
3281	86.35	89.14	90.72	92.45
3937	86.70	90.09	91.17	92.88
4593	**86.23**	**89.65**	**91.59**	**93.19**
5249	86.77	89.93	91.94	93.62
5905	86.85	90.07	92.37	93.96
6561	86.58	89.78	92.61	94.10

*Bold and shaded value indicates the result of proposed method.

Table A1.2. System accuracy (%) of 4 classifiers varying data sizes for K5 protocol.

Data size	Classifier types			
	NB	DT	AB	RF
656	84.75	85.34	85.25	87.82
1312	87.20	88.66	89.82	91.46
1968	87.48	90.30	90.00	91.57
2624	87.16	89.80	90.37	92.26
3281	86.35	89.14	90.72	92.45
3937	86.70	90.09	91.17	92.88
4593	**86.23**	**89.65**	**91.59**	**93.19**
5249	86.77	89.93	91.94	93.62
5905	86.85	90.07	92.37	93.96
6561	86.58	89.78	92.61	94.10

*Bold and shaded value indicates the result of proposed method.

Table A1.3. System accuracy (%) of 4 classifiers varying data sizes for K10 protocol.

Data size	Classifier types			
	NB	DT	AB	RF
656	85.61	88.64	88.18	90.91
1312	85.50	87.33	88.55	90.46
1968	86.55	91.22	91.42	92.69
2624	86.37	88.47	90.53	92.06
3281	86.80	89.70	91.22	93.11
3937	85.30	89.19	91.24	92.89
4593	**86.25**	**90.07**	**91.98**	**93.51**
5249	86.76	89.73	92.10	93.66
5905	87.07	90.20	92.94	94.31
6561	86.22	89.21	92.59	93.86

*Bold and shaded value indicates the result of proposed method.

Appendix A2

Table A2.1. Performance evaluation parameters (%) of 10 classifiers and 6 FS method between O1 and O2 for K2 protocol.

CT*	FST	O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C1	F1	88.74	56.37	80.62	74.74	84.50	88.22	72.58	82.49	77.17	89.33
	F2	88.80	56.97	77.46	75.00	85.21	87.69	78.34	88.03	81.60	89.16
	F3	87.54	58.65	76.56	71.17	85.06	89.28	72.40	88.14	74.81	89.86
	F4	86.59	59.65	82.35	70.54	84.99	86.43	75.08	85.93	70.25	88.95
	F5	86.56	61.16	83.14	69.11	85.40	88.28	73.60	86.27	75.97	89.00
	F6	82.64	46.15	68.83	78.33	74.55	87.82	56.33	81.02	73.64	83.62
C2	F1	87.62	48.97	75.74	68.97	82.77	86.17	76.15	86.76	75.42	88.42
	F2	87.57	55.28	78.79	70.14	83.30	85.62	81.41	89.62	75.06	88.73
	F3	87.25	50.97	76.42	68.86	82.56	87.42	75.95	87.25	76.37	89.33
	F4	85.42	54.78	78.89	65.57	82.20	85.43	77.98	88.16	73.57	87.98
	F5	86.00	55.39	78.40	67.78	82.70	87.45	77.13	87.63	76.94	88.45
	F6	80.46	45.26	72.99	56.20	73.40	86.23	58.37	78.91	70.19	82.08
C3	F1	86.80	58.04	79.00	71.02	84.77	85.96	78.24	87.74	75.58	89.41
	F2	88.59	57.53	79.85	72.61	85.39	86.01	82.38	90.15	75.82	89.47
	F3	85.92	60.91	80.02	70.40	84.49	86.94	77.72	88.02	76.10	90.05
	F4	85.15	62.19	81.69	67.96	84.14	84.96	77.90	88.05	72.94	88.88
	F5	85.44	62.69	81.19	69.63	84.90	87.45	77.13	87.63	76.94	88.45
	F6	80.37	47.64	73.84	57.14	73.25	87.13	58.49	79.13	71.64	82.99
C4	F1	88.35	79.50	88.78	78.98	89.84	91.01	77.71	88.29	83.16	91.25
	F2	87.79	78.65	88.77	77.53	89.63	92.86	75.31	87.72	85.52	91.90
	F3	85.73	80.23	89.10	75.65	89.11	90.36	79.19	89.21	81.70	91.75
	F4	85.75	77.12	88.21	73.39	88.10	88.92	77.77	88.59	78.71	90.35
	F5	87.39	78.99	88.78	77.07	90.09	91.77	74.90	87.16	83.22	91.01
	F6	85.41	47.76	75.35	64.57	75.74	86.01	61.74	80.38	71.46	82.53
C5	F1	89.60	77.41	87.80	80.40	89.45	90.52	75.98	87.25	81.66	90.53
	F2	89.18	77.35	88.19	79.17	89.09	89.01	77.74	88.23	78.96	90.82
	F3	89.42	74.08	86.22	79.44	89.28	89.91	76.27	87.77	80.15	90.53
	F4	85.93	77.20	88.23	73.53	88.22	88.13	77.09	88.10	77.20	89.21
	F5	88.52	78.66	88.70	78.64	90.15	89.56	73.61	86.29	79.23	88.85
	F6	83.29	50.56	75.61	62.12	74.00	88.15	54.14	77.66	71.73	79.46
C6	F1	84.07	69.80	83.50	70.97	80.39	87.69	62.56	80.97	73.79	80.41
	F2	84.24	63.32	81.29	68.01	80.26	83.78	66.29	82.36	68.53	77.50
	F3	86.49	73.97	85.77	75.28	85.81	87.50	65.55	82.74	73.78	79.65
	F4	80.24	78.99	88.33	66.90	85.01	84.57	71.37	85.01	70.64	84.51
	F5	83.62	69.53	83.80	69.53	82.76	84.32	65.46	81.87	69.35	78.16
	F6	87.33	34.25	71.10	60.24	67.03	89.52	40.67	73.36	70.11	70.88
C7	F1	89.36	80.88	89.48	80.68	93.30	89.30	77.51	87.82	80.01	91.55
	F2	90.06	77.59	88.42	80.57	91.67	87.47	80.04	89.20	77.16	92.34
	F3	89.55	82.97	90.51	81.33	93.88	89.27	78.11	88.55	79.57	92.25
	F4	86.24	82.75	90.84	75.22	92.67	87.33	78.20	88.49	76.38	91.38
	F5	88.53	80.71	89.63	78.93	92.90	86.78	79.14	88.49	76.57	91.03
	F6	77.05	67.71	81.50	61.82	80.67	76.07	65.19	79.87	60.14	78.05

(Continued Table A2.1)

CT*	FST	O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C8	F1	88.12	60.71	80.78	74.82	84.82	86.70	80.12	88.88	77.01	89.35
	F2	88.22	61.77	81.51	74.06	85.40	87.75	80.48	89.47	77.95	89.31
	F3	88.05	60.95	80.67	74.12	85.11	87.19	80.34	89.34	77.06	89.88
	F4	84.57	70.45	85.17	70.32	85.27	86.81	77.54	88.32	75.95	88.90
	F5	85.89	67.72	83.63	72.60	85.73	87.49	78.08	88.29	77.66	89.01
	F6	83.20	49.10	75.14	61.72	74.60	89.71	56.54	79.03	75.69	83.59
C9	F1	92.37	77.16	88.07	85.04	90.24	92.18	73.05	86.24	84.32	89.24
	F2	95.74	70.88	86.26	90.15	90.20	92.21	74.58	87.23	84.00	89.70
	F3	92.07	75.42	87.44	85.28	89.54	92.32	71.23	85.95	83.49	89.40
	F4	95.71	68.44	85.76	89.18	88.83	92.51	70.85	86.02	83.88	87.74
	F5	95.54	71.63	86.52	89.99	90.89	95.05	66.50	84.04	88.37	88.44
	F6	89.87	47.74	76.38	74.28	78.81	83.30	56.28	78.37	66.83	76.99
C10	F1	94.06	79.99	89.58	88.41	94.25	93.55	76.10	87.79	87.10	93.29
	F2	94.10	75.28	88.08	87.72	93.17	93.91	76.22	88.15	87.16	93.66
	F3	93.70	80.05	89.63	87.77	94.25	93.43	76.81	88.38	86.24	93.62
	F4	93.37	75.74	88.51	85.71	93.31	94.81	70.95	86.41	88.48	92.79
	F5	94.80	72.00	86.53	89.45	92.30	94.35	75.88	88.10	88.01	93.98
	F6	83.24	62.15	80.51	67.55	81.39	87.90	54.54	77.80	71.77	80.94

*Bold and shaded value indicates the result of proposed method.

Table A2.2. Performance evaluation parameters (%) of 10 classifiers and 6 FS method between O1 and O2 for K4 protocol.

CT*	FST	O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C1	F1	88.04	55.94	73.57	71.15	84.50	89.16	77.62	92.00	76.12	91.15
	F2	89.79	56.74	78.10	78.18	85.17	89.54	77.11	90.28	83.33	90.73
	F3	88.46	59.51	74.29	75.00	85.88	87.02	77.51	87.12	76.67	90.59
	F4	88.39	58.65	80.15	82.14	85.63	88.52	74.25	84.67	83.64	89.59
	F5	86.56	61.16	83.14	69.11	85.40	87.98	72.36	82.96	84.21	88.56
	F6	84.23	46.20	73.94	74.00	76.44	88.36	54.42	81.56	70.59	83.79
C2	F1	87.83	48.50	74.93	69.49	82.57	86.84	80.00	88.82	76.96	90.38
	F2	87.72	53.50	75.76	72.55	82.96	87.18	81.34	90.04	76.82	89.64
	F3	87.66	53.37	77.09	71.04	83.45	86.26	80.15	89.68	74.72	89.85
	F4	86.20	55.87	77.33	69.76	82.76	86.87	77.99	88.17	76.21	88.57
	F5	86.00	55.39	78.40	67.78	82.70	86.15	75.98	86.77	75.15	88.19
	F6	82.04	44.26	73.51	56.44	75.38	85.71	57.68	79.37	68.11	82.52
C3	F1	86.14	57.66	78.10	70.47	84.10	86.17	82.41	90.01	76.54	90.96
	F2	88.40	57.93	77.68	75.16	84.91	87.07	81.07	89.86	76.76	90.39
	F3	85.78	63.46	80.79	71.37	85.30	85.58	80.89	90.01	73.71	90.39
	F4	86.54	62.78	80.27	72.72	84.64	86.72	76.87	87.62	75.54	89.50
	F5	85.44	62.69	81.19	69.63	84.90	86.46	74.99	86.35	75.28	88.46
	F6	82.38	46.77	74.48	58.33	74.88	86.78	58.03	79.73	69.76	83.09
C4	F1	88.23	82.44	90.09	80.58	90.51	92.12	81.10	90.04	85.75	92.49
	F2	88.41	80.84	88.77	81.37	89.83	93.34	78.20	89.42	86.25	92.81
	F3	89.22	78.46	88.45	80.74	90.58	92.75	75.35	88.71	84.31	92.67
	F4	84.69	82.96	89.83	75.61	89.09	89.17	80.32	89.91	80.42	91.16
	F5	87.39	78.99	88.78	77.07	90.09	93.41	72.68	86.35	86.83	90.66
	F6	86.91	53.96	78.57	70.02	79.25	85.49	63.86	82.53	70.21	82.74
C5	F1	90.98	78.80	88.23	83.39	90.17	90.33	79.97	89.07	82.05	91.40
	F2	88.08	75.81	85.79	79.77	89.09	91.04	77.75	88.97	81.83	92.10
	F3	88.34	78.71	88.22	79.03	90.42	89.65	78.67	89.37	79.15	91.88
	F4	87.89	74.62	85.72	77.97	89.20	88.72	77.83	88.22	78.72	90.38
	F5	88.52	78.66	88.70	78.64	90.15	89.90	73.26	85.93	80.03	88.87
	F6	84.39	53.95	77.58	64.74	78.17	88.77	50.46	77.29	70.33	80.20
C6	F1	84.80	71.46	83.84	72.98	83.79	85.65	69.29	83.47	72.58	83.39
	F2	83.42	63.15	78.88	70.43	81.40	85.23	68.57	84.03	70.43	81.53
	F3	84.67	75.73	86.39	73.67	86.78	85.55	68.96	84.54	70.32	83.39
	F4	83.61	75.61	85.70	72.70	86.01	84.54	71.16	84.67	71.26	85.29
	F5	83.62	69.53	83.80	69.53	82.76	83.72	66.76	82.11	69.07	81.44
	F6	88.49	35.39	72.74	67.18	68.20	89.56	37.56	73.81	69.74	69.08

(Continued Table A2.2)

CT*	FST	O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C7	F1	88.67	82.54	89.90	80.60	93.49	88.36	81.94	89.88	79.44	92.21
	F2	88.62	78.26	87.10	80.80	92.14	91.25	79.84	89.93	82.86	93.73
	F3	89.05	84.93	91.31	81.27	94.02	88.00	82.97	91.21	77.47	92.60
	F4	88.05	80.60	88.75	79.47	92.92	88.14	78.46	88.51	78.00	91.02
	F5	88.53	80.71	89.63	78.93	92.90	85.70	79.32	88.31	75.72	91.22
	F6	81.05	65.25	81.63	64.48	82.45	75.34	63.36	79.73	57.55	77.87
C8	F1	86.40	63.72	81.51	73.98	84.74	88.93	81.91	90.27	80.54	91.21
	F2	86.42	67.66	81.92	76.26	85.25	89.32	83.18	91.60	80.62	90.78
	F3	86.86	67.66	83.36	75.37	86.09	89.25	78.33	89.50	78.95	90.56
	F4	85.71	72.19	84.88	75.02	85.98	90.07	76.22	87.95	80.83	89.56
	F5	85.89	67.72	83.63	72.60	85.73	88.63	75.66	86.94	78.56	88.51
	F6	82.97	53.90	77.65	63.38	76.59	88.62	60.26	81.28	74.29	83.83
C9	F1	92.73	76.39	87.66	86.86	90.98	92.95	72.57	86.61	86.56	89.83
	F2	96.74	73.72	85.92	93.33	91.93	92.76	75.73	88.26	85.28	91.47
	F3	90.88	79.60	89.26	83.44	91.87	94.30	69.77	86.63	87.43	90.37
	F4	95.10	72.60	85.96	90.30	90.57	96.18	66.22	84.45	91.05	88.75
	F5	95.54	71.63	86.52	89.99	90.89	94.81	67.15	84.45	89.42	89.63
	F6	87.33	58.46	81.02	74.31	81.26	86.14	57.70	80.06	70.14	79.64
C10	F1	93.59	83.04	90.74	88.44	94.30	95.12	79.00	89.31	90.47	94.48
	F2	94.20	79.56	88.51	89.42	94.04	95.67	77.40	89.37	90.96	94.38
	F3	94.80	8s0.68	89.96	90.02	94.87	93.94	77.82	90.11	87.82	94.27
	F4	93.57	79.28	89.09	88.71	93.70	94.42	76.16	88.28	88.14	93.27
	F5	94.35	75.88	88.10	88.01	93.98	96.60	68.55	84.82	92.08	92.26
	F6	83.79	66.65	83.65	69.92	84.23	89.43	50.47	77.94	72.50	80.85

*Bold and shaded value indicates the result of proposed method.

Table A2.3. Performance evaluation parameters (%) of 10 classifiers and 6 FS method between O1 and O2 for K5 protocol.

FST		O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C1	F1	90.79	61.35	78.38	74.42	87.29	88.25	78.69	89.36	73.33	90.57
	F2	89.57	59.20	80.00	85.29	85.85	88.25	72.56	79.80	85.45	89.38
	F3	87.23	59.74	76.53	62.50	83.73	89.01	75.79	88.42	77.97	89.55
	F4	88.61	58.04	82.24	72.34	87.35	87.23	72.68	92.23	64.71	89.59
	F5	87.81	59.83	77.19	75.00	84.88	88.29	72.69	84.11	82.98	89.31
	F6	86.53	45.34	84.43	53.12	78.16	88.14	56.88	78.57	71.43	85.03
C2	F1	89.68	55.32	78.66	74.24	85.19	86.00	81.64	90.19	74.99	90.09
	F2	87.89	56.76	79.55	71.12	83.33	85.55	76.65	86.52	75.03	87.75
	F3	86.10	51.26	76.43	66.81	81.12	87.38	79.68	88.57	77.64	88.82
	F4	87.29	55.65	78.77	70.06	84.42	85.46	75.83	87.18	72.79	88.50
	F5	85.42	56.88	78.23	68.13	81.30	86.68	76.79	88.39	74.10	88.82
	F6	85.62	47.10	74.44	64.5	78.35	86.6	57.78	78.78	70.22	83.23
C3	F1	88.96	62.22	81.26	75.17	87.33	85.74	81.68	90.16	74.63	90.56
	F2	88.29	59.96	80.84	73.26	85.49	86.02	75.79	86.11	75.68	89.12
	F3	85.33	60.82	79.98	69.31	83.12	86.27	79.30	88.16	76.15	89.64
	F4	86.83	62.02	81.14	71.33	86.63	86.05	74.58	86.60	73.46	89.45
	F5	86.20	61.08	80.02	70.79	83.76	87.25	75.60	87.93	74.61	89.06
	F6	84.35	49.57	75.04	63.68	77.5	86.97	59.34	79.42	71.51	83.94
C4	F1	89.82	85.63	92.32	82.48	92.63	92.63	80.99	90.64	85.66	92.30
	F2	91.07	80.64	90.07	82.74	91.65	91.78	77.74	87.91	84.56	91.80
	F3	87.69	79.02	88.80	78.34	89.07	93.41	78.81	89.10	87.08	91.71
	F4	90.01	76.36	88.14	80.96	90.51	90.50	77.35	88.76	80.77	91.12
	F5	93.03	71.44	85.93	86.31	89.98	92.86	74.91	88.43	84.46	91.38
	F6	85.2	64.57	82.42	73.21	81.62	88.31	60.31	80.44	74.69	83.69
C5	F1	91.31	80.19	89.32	82.95	92.21	89.83	80.13	89.91	80.27	91.68
	F2	90.11	79.54	89.61	80.86	91.38	90.38	75.69	86.60	81.59	90.81
	F3	87.31	78.70	88.23	77.06	89.04	91.00	79.28	88.72	82.96	90.39
	F4	89.28	74.65	86.84	78.67	90.45	89.32	73.51	86.52	78.05	90.51
	F5	88.12	78.87	88.34	78.42	89.87	89.49	74.07	87.53	78.10	89.15
	F6	87.93	53.16	76.99	70.72	79.72	89.44	55.41	78.39	74.09	82.02
C6	F1	85.57	73.46	85.47	73.26	85.31	85.42	67.49	83.77	70.38	83.68
	F2	85.15	69.08	84.17	71.07	83.85	82.80	69.69	82.71	69.74	82.00
	F3	85.24	74.43	85.93	73.18	86.19	86.37	68.71	83.20	73.51	82.89
	F4	85.46	77.76	87.80	74.29	87.85	86.11	70.86	84.87	72.38	86.22
	F5	85.81	72.84	85.07	74.04	85.25	84.73	68.00	84.32	68.65	83.59
	F6	85.00	48.86	75.27	64.36	72.51	90.68	38.77	73.47	73.86	70.99

(Continued Table A2.3)

CT*		O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C7	F1	88.83	85.00	91.47	80.43	94.82	88.68	82.39	90.87	78.85	92.37
	F2	92.13	75.47	87.73	83.31	92.67	91.06	79.45	88.71	83.33	92.86
	F3	87.03	84.09	90.90	77.79	93.21	89.01	83.64	90.66	80.76	93.24
	F4	89.37	81.28	89.88	80.52	93.75	86.47	77.05	87.73	74.47	91.23
	F5	88.61	82.17	89.99	79.84	93.80	84.69	77.27	88.34	71.69	91.28
	F6	81.02	62.68	79.77	64.34	83.12	76.86	67.55	81.09	61.98	80.22
C8	F1	89.08	69.94	85.02	78.48	87.66	91.59	77.01	88.96	83.09	90.54
	F2	90.28	61.95	82.14	78.29	86.18	85.88	83.69	90.35	77.64	89.45
	F3	83.41	70.07	84.14	70.70	84.00	88.33	81.18	89.56	79.48	89.57
	F4	86.60	74.64	86.74	75.05	87.48	87.03	79.68	89.42	76.38	89.54
	F5	89.14	64.14	82.12	77.40	84.93	89.01	76.34	88.68	77.94	89.26
	F6	84.02	55.95	78.15	66.37	78.14	90.96	59.64	80.22	79.1	85.11
C9	F1	94.00	78.19	88.97	88.47	93.49	93.10	76.51	88.78	86.03	90.89
	F2	94.82	72.68	87.09	88.71	92.22	94.06	72.01	85.76	87.95	90.34
	F3	95.05	72.51	86.40	89.23	91.29	93.69	72.18	85.96	87.36	89.20
	F4	95.13	74.39	87.62	90.28	92.72	95.77	65.68	84.60	90.65	89.91
	F5	94.35	75.63	87.79	89.27	91.62	94.91	67.36	86.04	89.37	90.75
	F6	86.08	61.67	82.59	77.17	82.59	88.72	54.84	78.4	74.12	81.22
C10	F1	95.55	81.69	90.66	90.93	95.75	94.26	80.88	90.73	88.71	94.18
	F2	94.46	78.40	89.48	88.69	94.58	95.25	77.37	88.44	91.06	93.66
	F3	95.15	76.64	88.39	90.24	93.64	94.61	80.29	89.86	89.60	93.81
	F4	95.18	78.63	89.52	90.52	94.70	93.87	74.11	88.21	88.19	93.00
	F5	94.98	78.67	89.06	89.91	94.23	96.22	72.59	87.76	90.66	93.29
	F6	85.22	65.61	83.1	75.79	84.55	91.62	55.1	79.05	79.55	82.96

*Bold and shaded value indicates the result of proposed method.

*

Table A2.4. Performance evaluation parameters (%) of 10 classifiers and 6 FS method between O1 and O2 for K10 protocol

CT*	FST	O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C1	F1	88.94	60.93	79.63	82.61	86.09	87.75	74.05	88.14	66.67	88.60
	F2	89.43	59.76	84.75	55.56	86.67	88.25	72.56	79.80	85.45	89.38
	F3	84.62	61.01	78.95	65.00	83.77	87.93	73.72	78.00	74.07	89.04
	F4	88.83	58.29	80.00	64.71	86.17	89.34	70.29	79.63	78.26	89.39
	F5	87.23	56.60	79.63	73.91	84.44	90.66	71.60	87.27	95.45	90.38
	F6	86.78	46.66	76.36	63.64	76.79	88.65	56.23	88.68	66.67	83.58
C2	F1	88.89	54.51	78.41	72.08	84.97	86.72	78.89	89.49	73.87	88.46
	F2	86.85	59.52	77.88	72.74	84.82	85.55	76.65	86.52	75.03	87.75
	F3	83.18	54.30	77.71	62.29	81.98	85.43	77.34	87.07	75.62	88.41
	F4	87.05	53.60	78.21	68.99	83.33	89.37	72.97	85.57	79.57	89.06
	F5	87.18	53.22	78.24	69.09	82.90	89.10	74.84	84.65	81.10	90.34
	F6	83.43	44.81	74.06	58.81	75.58	87.15	57.98	80.49	68.92	81.75
C3	F1	87.69	62.43	81.29	72.81	86.39	84.66	78.78	89.28	71.08	88.09
	F2	88.73	61.32	79.18	75.99	86.82	86.02	75.79	86.11	75.68	89.12
	F3	82.82	62.39	80.92	65.26	83.75	86.14	78.20	87.66	76.53	89.03
	F4	87.20	61.00	80.98	71.71	85.05	88.06	72.65	85.33	77.25	89.39
	F5	86.95	59.26	80.46	70.55	85.10	88.93	73.85	84.20	80.54	90.40
	F6	83.56	49.14	75.63	61.42	75.45	87.75	57.76	80.53	69.68	82.41
C4	F1	91.11	85.29	92.68	84.81	92.47	94.39	78.51	90.40	87.72	91.40
	F2	91.69	80.38	89.68	87.51	92.58	91.78	77.74	87.91	84.56	91.80
	F3	88.18	76.47	88.32	77.96	88.90	90.97	82.63	90.35	84.24	90.91
	F4	89.74	77.45	89.06	81.54	89.07	91.96	76.44	88.41	86.53	90.70
	F5	90.30	80.23	89.90	81.40	90.61	96.13	74.90	86.08	92.11	91.61
	F6	88.73	60.05	81.31	77.09	80.63	87.30	65.03	83.72	74.78	82.13
C5	F1	91.80	80.81	89.83	84.37	92.00	90.14	78.65	89.90	79.36	90.25
	F2	88.67	80.73	88.33	80.74	91.34	90.38	75.69	86.60	81.59	90.81
	F3	86.49	75.69	87.34	74.62	89.43	89.77	76.97	87.41	81.60	89.86
	F4	87.33	75.58	87.10	75.96	88.63	89.24	72.13	85.17	79.85	89.22
	F5	89.27	77.21	88.20	79.58	89.60	92.46	73.72	84.76	85.49	89.15
	F6	84.83	58.75	79.48	67.24	79.21	89.25	53.49	79.26	70.56	81.10
C6	F1	85.92	79.80	88.59	74.92	87.89	85.91	71.27	86.15	70.94	84.08
	F2	85.22	66.64	80.92	72.78	83.91	82.80	69.69	82.71	69.74	82.00
	F3	85.61	73.40	86.16	72.31	85.65	83.80	64.59	80.71	69.42	80.44
	F4	83.11	71.35	84.72	69.13	85.74	85.32	69.36	83.03	72.78	86.24
	F5	85.86	72.75	85.83	73.56	85.21	82.64	66.88	79.92	70.63	83.20
	F6	88.16	40.83	74.31	74.71	70.64	91.86	42.38	76.61	78.85	72.13

(Continued Table A2.4)

CT*	FST	O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C7	F1	91.47	86.32	92.50	84.34	95.45	86.63	79.90	90.05	74.23	91.14
	F2	92.46	81.99	89.77	86.65	93.76	91.06	79.45	88.71	83.33	92.86
	F3	85.68	86.40	91.94	75.30	93.96	89.05	81.86	89.80	81.71	91.77
	F4	89.68	77.47	88.51	79.61	92.70	87.78	73.36	85.53	76.63	89.79
	F5	84.97	83.04	90.54	74.51	93.03	87.70	75.98	85.25	79.06	91.81
	F6	78.73	66.44	81.69	62.13	81.67	76.20	66.50	82.05	58.57	77.77
C8	F1	88.75	71.42	86.24	78.70	86.57	91.15	74.79	88.64	80.32	88.53
	F2	87.28	69.37	83.99	81.15	86.89	85.88	83.69	90.35	77.64	89.45
	F3	87.94	62.19	83.31	74.96	83.81	89.11	81.69	90.00	81.78	89.13
	F4	91.57	63.76	83.50	82.54	86.37	90.99	78.94	89.67	84.09	89.34
	F5	87.63	67.12	84.94	77.98	84.90	89.29	81.39	88.81	82.26	90.30
	F6	89.44	50.12	77.60	71.26	76.95	93.71	53.90	80.33	82.44	83.44
C9	F1	94.93	77.73	89.32	90.52	94.32	94.53	71.82	87.66	87.17	87.45
	F2	94.31	80.50	89.45	90.01	92.83	94.06	72.01	85.76	87.95	90.34
	F3	92.09	74.59	88.37	85.24	91.99	90.93	78.39	88.74	84.73	89.85
	F4	94.92	69.15	85.59	89.06	90.32	96.81	70.68	86.24	92.75	89.99
	F5	93.15	75.25	87.99	86.09	90.29	89.38	78.32	88.09	84.99	90.45
	F6	92.49	48.29	77.77	79.22	80.64	87.52	54.17	79.75	72.27	80.12
C10	F1	95.36	86.15	92.50	90.97	95.90	95.96	79.72	91.14	91.20	93.11
	F2	94.66	83.55	91.31	91.99	95.40	95.25	77.37	88.44	91.06	93.66
	F3	93.83	81.93	91.37	88.16	94.59	94.64	77.83	88.76	90.74	92.54
	F4	95.31	76.26	89.04	91.59	93.77	94.52	75.92	88.08	90.57	91.97
	F5	97.66	72.92	87.35	94.81	94.30	96.23	71.93	85.25	92.51	92.48
	F6	85.60	65.84	84.08	75.29	84.06	92.11	51.68	79.33	79.44	80.74

*Bold and shaded value indicates the result of proposed method.

.

Table A2.5. Performance evaluation parameters (%) of 10 classifiers and 6 FS method between O1 and O2 for JK protocol.

		O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C1	F1	88.08	58.96	80.00	72.48	85.11	88.41	76.13	87.35	77.86	90.12
	F2	90.48	55.49	79.20	76.02	84.55	88.41	75.81	87.18	77.78	89.94
	F3	88.20	59.22	80.18	72.94	85.07	88.41	76.13	87.35	77.86	90.12
	F4	88.12	58.47	79.89	72.69	85.76	88.58	73.88	86.35	77.65	89.71
	F5	91.79	61.16	81.53	80.00	85.88	88.60	74.49	86.38	77.95	89.56
	F6	82.83	42.61	72.88	60.37	68.83	85.83	43.61	73.88	62.37	75.83
C2	F1	88.20	53.37	77.92	70.80	84.14	87.19	78.94	88.54	76.76	89.83
	F2	89.08	66.56	83.25	76.57	87.45	86.82	81.02	89.51	76.71	89.47
	F3	88.60	53.17	77.93	71.43	84.15	87.19	78.94	88.54	76.76	89.83
	F4	86.41	55.44	78.35	68.61	83.55	87.24	77.96	88.07	76.61	89.70
	F5	90.98	66.01	83.31	79.68	89.06	86.98	77.99	88.06	76.25	89.34
	F6	82.41	44.03	71.78	60.22	74.70	84.41	44.03	73.78	60.22	75.70
C3	F1	86.60	60.33	80.29	70.71	85.11	85.80	79.88	88.84	75.10	89.96
	F2	89.58	58.64	80.16	75.10	86.27	86.61	79.22	88.61	76.02	89.77
	F3	85.73	60.83	80.33	69.56	84.68	85.80	79.88	88.84	75.10	89.96
	F4	86.18	60.86	80.42	70.24	85.01	86.20	76.34	87.17	74.78	89.57
	F5	90.95	70.07	85.01	80.58	88.93	86.01	76.94	87.44	74.67	89.50
	F6	81.94	44.53	71.96	56.85	73.37	82.94	45.52	73.96	58.85	74.37
C4	F1	92.12	83.19	91.09	85.00	93.04	94.09	77.88	88.81	87.61	93.59
	F2	96.42	72.64	86.81	91.60	65.10	94.58	77.79	88.82	88.51	93.59
	F3	92.01	81.93	90.48	84.61	92.51	94.10	77.89	88.82	87.63	93.60
	F4	91.17	78.56	88.83	82.72	92.25	90.43	81.45	90.10	82.04	93.01
	F5	96.73	72.22	86.66	92.22	65.10	93.74	78.49	89.06	87.09	93.11
	F6	85.34	62.31	80.46	70.35	82.35	86.34	63.31	81.46	71.35	83.35
C5	F1	93.10	85.40	92.25	86.90	92.72	91.83	82.74	90.85	84.45	93.89
	F2	96.77	75.19	87.92	92.58	93.60	92.50	80.45	89.82	85.19	93.80
	F3	92.29	85.70	92.33	85.64	92.72	91.79	82.72	90.83	84.38	93.86
	F4	91.24	84.75	91.78	83.83	93.14	90.65	81.49	90.14	82.36	92.89
	F5	96.40	75.45	87.99	91.83	93.58	92.86	79.74	89.53	85.70	92.84
	F6	88.16	55.12	78.13	72.51	81.64	89.16	56.12	79.13	73.51	82.64
C6	F1	87.01	73.12	85.79	75.11	85.53	86.23	69.54	84.08	73.02	84.87
	F2	100.00	51.87	79.49	100.00	87.76	83.92	70.22	84.02	70.07	84.53
	F3	88.22	75.63	87.10	77.49	88.36	86.36	69.55	84.10	73.23	85.08
	F4	86.19	76.99	87.48	74.94	88.47	86.88	71.95	85.25	74.63	88.14
	F5	100.00	51.89	79.50	99.99	87.49	83.85	70.12	83.96	69.95	84.80
	F6	85.64	39.08	84.83	55.54	70.28	89.64	39.08	82.83	53.54	69.28

(Continued Table A2.5)

		O1					O2				
		SE	SP	PPV	NPV	AUC	SE	SP	PPV	NPV	AUC
C7	F1	99.88	99.98	99.99	99.78	99.99	99.74	99.95	99.97	99.53	99.99
	F2	99.51	99.46	99.71	99.09	99.99	99.33	99.07	99.50	98.76	99.98
	F3	99.82	99.98	99.99	99.66	100.00	99.74	99.95	99.97	99.52	99.99
	F4	99.43	99.97	99.98	98.96	99.99	99.75	99.93	99.96	99.53	99.99
	F5	99.73	99.97	99.98	99.51	99.99	98.17	99.47	99.71	96.70	99.96
	F6	95.96	94.13	97.45	93.49	98.75	96.96	97.13	98.45	94.49	99.75
C8	F1	88.51	59.30	80.23	73.45	85.44	87.74	79.26	88.76	77.61	90.10
	F2	85.31	67.67	83.12	71.18	84.54	86.20	84.65	91.29	76.68	89.90
	F3	88.62	59.24	80.22	73.62	85.44	87.74	79.26	88.76	77.61	90.10
	F4	83.79	73.76	85.63	70.92	86.07	84.69	83.09	90.34	74.42	89.72
	F5	92.11	61.39	81.65	80.67	85.89	85.54	81.49	89.62	75.16	89.55
	F6	81.70	51.77	75.57	60.06	73.85	82.70	52.77	76.57	62.06	75.85
C9	F1	95.81	77.09	88.64	90.80	96.54	92.79	85.37	92.22	86.42	95.13
	F2	98.20	75.38	88.15	95.73	96.25	92.66	85.42	92.24	86.25	95.16
	F3	97.00	77.03	88.74	93.22	96.84	92.79	85.37	92.22	86.42	95.13
	F4	96.00	71.09	86.11	90.50	94.16	93.49	79.44	89.47	86.77	94.43
	F5	98.20	75.38	88.15	95.73	96.25	93.48	79.45	89.47	86.76	93.99
	F6	85.89	60.32	80.16	69.66	84.07	85.89	60.32	80.16	69.66	84.07
C10	**F1**	**99.99**	**99.98**	**99.99**	**99.99**	**100.00**	**99.99**	**99.98**	**99.99**	**99.99**	**100.00**
	F2	**100.00**	**99.98**	**99.99**	**99.99**	**100.00**	**99.99**	**99.98**	**99.99**	**99.99**	**100.00**
	F3	**99.99**	**99.98**	**99.99**	**99.99**	**100.00**	**99.99**	**99.98**	**99.99**	**99.99**	**100.00**
	F4	**99.99**	**99.98**	**99.99**	**99.99**	**100.00**	**99.99**	**99.98**	**99.99**	**99.99**	**100.00**
	F5	**100.00**	**99.98**	**99.99**	**99.99**	**100.00**	**99.99**	**99.98**	**99.99**	**99.98**	**100.00**
	F6	**99.98**	**99.96**	**99.98**	**99.96**	**99.99**	**99.98**	**99.96**	**99.98**	**99.96**	**99.99**

Bold and shaded value indicates the result of proposed method.

Appendix A3

This **Appendix A3.1** demonstrates the optimization of kernels on the two machine learning classifiers namely: GPC and SVM including three types of kernel namely: Linear, Poly-2, and RBF. The **Appendix A3.1** also demonstrates the different classifiers performance while changing statistical tests. Ten classifiers are compared in each of the tables shown below. Each table corresponds to different set of statistical tests.

Table A3.1. Comparison of mean accuracy (%) for 3 kernels using 3 CV protocols.

PT	SVM: Linear	GPC: Linear	SVM: RBF*	GPC: RBF	SVM: Ploy-2	GPC: Poly-2*
K2	80.99 ±6.04	81.79 ± 3.84	**85.75 ± 13.21**	84.31 ± 5.25	82.50 ± 6.25	**91.11 ± 6.50**
K10	84.16 ± 3.68	86.85 ± 4.49	**86.53 ± 4.14**	85.43 ±4.61	83.54 ± 13.95	**92.27 ±9.20**
JK	86.06 ± 12.98	90.25 ± 6.29	**87.03 ±13.04**	91.41 ±6.82	83.94 ±13.85	**92.48 ±9.09**

Mean accuracy is expressed as accuracy ± SD. *Bold and shaded value indicates the selected kernels.

Table A3.2. Change in mean accuracy (%) of 10 classifiers and different p-values for **t-test**.

PT	p-values	# of genes	LDA	QDA	NB	GPC	SVM	ANN	LR	DT	AB	RF*
K2	0.05	478	81.45	59.03	71.94	79.35	79.84	77.58	72.74	69.67	76.29	**83.71**
	0.01	246	82.90	58.71	70.97	80.71	81.77	77.09	75.80	71.29	78.38	**83.90**
	0.001	94	79.35	58.06	77.58	81.45	82.42	78.87	70.32	75.00	78.06	**84.19**
	0.0001	33	58.71	66.29	76.13	82.90	80.00	76.45	75.48	73.39	81.61	**87.42**
K10	0.05	478	73.33	63.34	70.00	85.00	86.67	79.99	79.16	79.17	82.08	**87.52**
	0.01	246	75.83	63.34	77.50	88.33	85.83	71.84	70.83	77.50	80.00	**88.33**
	0.001	94	78.33	59.17	72.50	87.50	85.83	74.21	74.16	75.83	80.00	**91.67**
	0.0001	33	77.50	74.16	86.67	81.66	68.42	70.42	70.50	81.67	85.00	**92.32**
JK	0.05	478	96.54	65.95	71.20	93.47	93.42	73.10	90.22	90.66	96.43	**99.74**
	0.01	246	96.35	63.86	75.84	93.24	93.03	71.62	90.14	90.66	96.38	**99.79**
	0.001	94	97.87	61.24	77.58	92.25	93.42	69.69	90.27	90.79	94.12	**99.77**
	0.0001	33	91.31	67.82	80.67	92.26	91.75	82.90	89.98	91.70	95.86	**99.72**

*Bold and shaded value indicates the result of proposed method.

Table A3.3. Change in mean accuracy (%) of 10 classifiers and different p-values for **KW test**.

PT	p-values	# of genes	LDA	QDA	NB	GPC	SVM	ANN	LR	DT	AB	RF*
K2	0.05	387	80.97	58.06	75.16	80.80	82.42	70.32	73.07	75.00	74.84	**84.84**
	0.01	188	83.23	65.00	76.77	79.35	85.16	77.25	75.00	72.58	77.58	**85.16**
	0.001	62	80.65	65.48	83.87	85.97	82.90	77.42	79.83	75.65	79.52	**85.16**
	0.0001	22	70.96	67.74	83.07	86.93	86.94	72.10	77.42	76.78	80.32	**87.42**
K10	0.05	387	81.67	55.00	65.83	82.50	80.00	78.16	74.17	79.17	77.50	**85.83**
	0.01	188	75.00	60.83	79.17	83.33	82.50	77.10	74.58	70.83	80.00	**86.67**
	0.001	62	70.00	60.83	83.33	82.50	84.33	75.53	73.34	80.00	85.83	**87.50**
	0.0001	22	80.83	70.00	85.83	85.00	85.83	79.74	75.83	84.17	83.33	**89.17**
JK	0.05	387	99.56	67.74	83.82	91.99	93.39	77.68	99.58	99.74	94.77	**99.77**
	0.01	188	96.38	66.65	83.66	92.54	93.29	69.87	99.09	99.69	96.35	**99.79**
	0.001	62	98.83	67.74	83.97	91.97	92.74	78.15	98.99	99.71	94.77	**99.82**
	0.0001	22	89.36	80.72	84.08	91.96	89.00	92.10	90.37	94.85	94.95	**99.50**

*Bold and shaded value indicates the result of proposed method.

Table A3.4. Change in mean accuracy (%) of 10 classifiers and different p-values for **F-test**.

PT	p-values	# of genes	LDA	QDA	NB	GPC	SVM	ANN	LR	DT	AB	RF*
K2	0.05	714	72.26	53.39	65.64	75.32	69.35	76.61	76.77	69.52	73.22	**75.81**
	0.01	431	74.03	56.45	70.16	71.61	70.48	77.58	73.39	72.74	77.74	**76.78**
	0.001	246	76.68	56.77	75.00	76.29	71.61	70.81	73.87	72.58	77.26	**79.20**
	0.0001	133	79.03	54.19	77.26	78.07	72.25	71.77	77.58	73.39	78.71	**81.13**
K10	0.05	714	79.17	55.84	63.33	74.17	70.17	72.81	78.92	80.83	79.58	**84.17**
	0.01	431	73.34	55.00	68.34	79.17	60.63	74.58	70.00	80.41	77.25	**88.33**
	0.001	246	80.83	63.33	68.34	86.33	73.34	71.56	71.25	75.00	78.34	**88.34**
	0.0001	133	79.16	68.33	79.58	87.50	75.42	71.40	74.17	84.17	79.17	**88.42**
JK	0.05	714	96.33	57.83	70.94	95.03	90.79	71.36	99.09	99.66	96.43	**99.71**
	0.01	431	96.38	64.31	71.49	93.81	93.00	72.14	99.12	99.71	96.38	**99.77**
	0.001	246	96.28	63.45	75.91	94.33	93.00	68.31	98.99	99.66	96.38	**99.77**
	0.0001	133	96.54	74.14	83.92	94.48	90.48	90.46	88.76	96.41	95.73	**99.74**

*Bold and shaded value indicates the result of proposed method.

Table A3.5. Performance evaluation in parameters (%) of all classifiers for **t-test**.

PT	CT	SE	SP	PPV	NPV	FM	AUC
K2	LDA	60.01	64.88	78.21	43.79	67.36	67.78
	QDA	84.84	54.08	77.23	67.58	79.99	83.17
	NB	76.90	86.63	91.97	68.68	82.84	91.62
	GPC	87.88	77.32	87.96	84.36	89.54	87.80
	SVM	88.01	81.05	89.38	80.37	87.81	40.80
	ANN	78.61	79.99	88.31	67.57	82.52	87.59
	LR	55.04	62.38	73.35	43.33	61.98	58.92
	DT	80.65	59.21	77.97	64.12	78.48	75.58
	AB	77.70	82.09	88.48	67.85	82.08	87.03
	RF*	**91.90**	**64.29**	**81.25**	**70.74**	**82.51**	**92.75**
K10	LDA	68.11	62.20	73.14	55.93	69.08	70.97
	QDA	94.41	55.71	78.90	70.25	85.02	89.78
	NB	69.80	93.82	95.45	62.54	80.15	90.72
	GPC	94.56	78.75	90.07	92.38	91.67	88.79
	SVM	86.61	85.17	91.95	76.48	88.54	43.36
	ANN	81.88	74.90	86.68	65.24	83.33	88.05
	LR	63.24	69.33	84.45	42.74	71.17	67.13
	DT	84.46	62.90	85.48	76.50	84.48	76.40
	AB	84.22	80.24	88.50	76.57	84.90	92.14
	RF*	**87.82**	**72.56**	**81.64**	**80.00**	**83.90**	**90.56**
JK	LDA	90.69	92.45	95.64	84.59	93.08	98.12
	QDA	62.82	76.91	83.19	53.22	71.58	76.47
	NB	70.12	99.85	99.89	64.77	82.40	94.23
	GPC	93.06	90.81	94.97	87.99	93.95	95.88
	SVM	94.60	86.58	92.77	89.88	93.67	94.68
	ANN	87.44	74.65	86.71	77.69	86.82	91.58
	LR	90.20	89.59	94.10	83.49	92.08	93.74
	DT	91.41	92.23	95.61	85.73	93.41	97.75
	AB	94.56	98.24	99.02	90.96	96.72	99.41
	RF*	**99.80**	**99.59**	**99.78**	**99.65**	**99.79**	**99.96**

*Bold and shaded value indicates the result of proposed method.

Table A3.6. Performance evaluation parameters (%) of all classifiers for **KW test**.

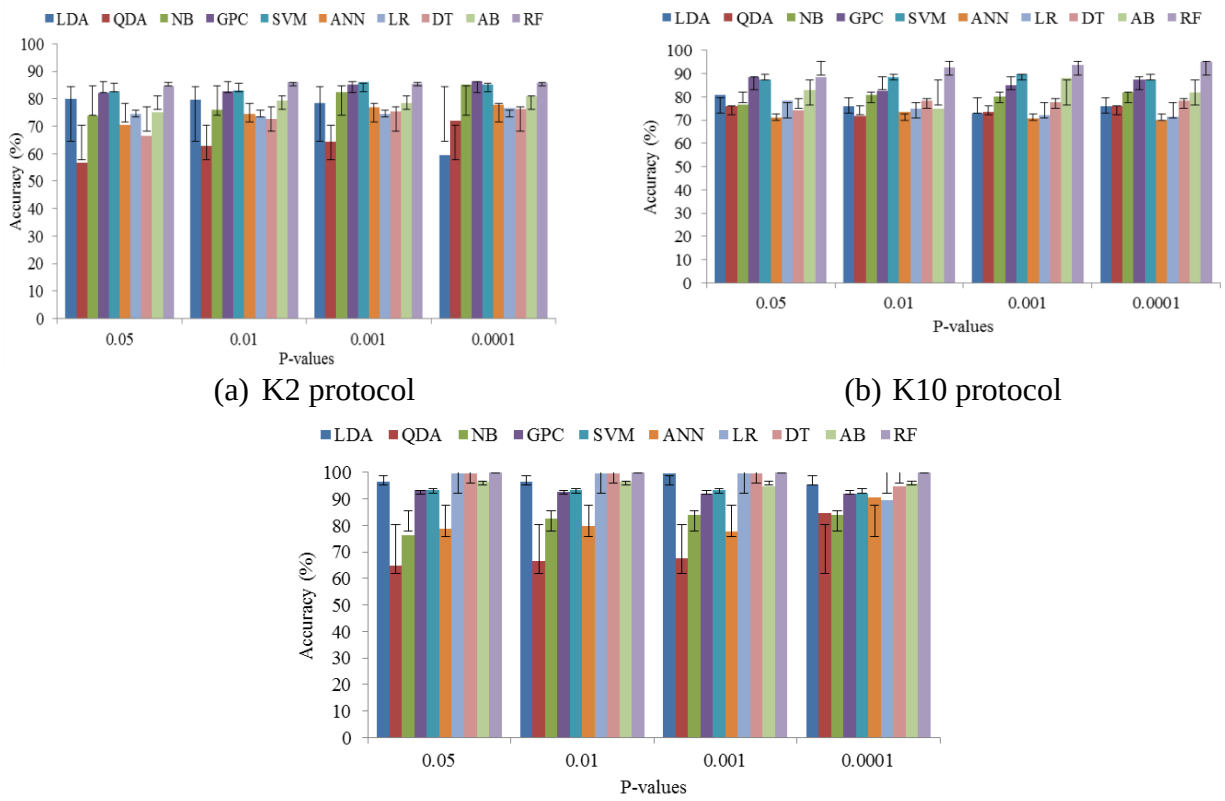
PT	CT	SE	SP	PPV	NPV	FM	AUC
K2	LDA	75.47	64.05	78.09	60.66	76.10	76.72
	QDA	92.33	71.17	84.80	83.62	87.92	90.11
	NB	83.43	83.22	90.13	73.57	86.48	91.03
	GPC	91.54	86.76	92.70	87.38	91.96	90.88
	SVM	91.36	73.10	86.00	83.46	88.31	41.91
	ANN	87.04	69.55	84.09	76.55	85.12	88.79
	LR	69.32	63.75	78.95	52.45	72.87	71.07
	DT	81.82	62.71	79.42	68.56	79.89	76.88
	AB	85.27	76.79	87.39	75.98	85.76	88.98
	RF*	**91.33**	**76.43**	**88.36**	**77.94**	**88.48**	**91.78**
K10	LDA	90.83	55.70	81.17	70.60	84.90	85.56
	QDA	91.85	78.02	85.26	83.75	87.46	89.01
	NB	80.04	86.15	93.22	68.74	85.21	92.69
	GPC	94.90	84.25	93.80	94.08	94.59	93.40
	SVM	93.52	76.42	89.59	84.50	90.98	19.60
	ANN	92.13	66.10	81.96	86.31	85.82	90.83
	LR	75.83	71.35	83.63	62.44	78.84	77.39
	DT	82.53	74.90	86.83	70.23	83.43	81.35
	AB	81.87	75.24	84.88	73.53	82.37	89.80
	RF*	**95.71**	**79.24**	**87.30**	**85.85**	**88.01**	**96.19**
JK	LDA	94.96	79.18	89.27	89.66	92.02	98.47
	QDA	92.50	59.31	80.52	81.30	86.10	87.82
	NB	80.36	90.84	94.10	71.80	86.69	94.53
	GPC	96.79	83.17	91.32	93.64	93.95	95.21
	SVM	92.46	82.70	90.69	85.78	91.56	95.96
	ANN	92.92	90.61	94.82	88.16	93.78	96.88
	LR	96.69	78.89	89.35	93.18	92.84	94.88
	DT	97.22	90.54	94.94	94.78	96.05	98.88
	AB	93.55	97.51	98.57	89.34	95.98	99.30
	RF*	**99.84**	**99.76**	**99.87**	**99.72**	**99.85**	**99.95**

*Bold and shaded value indicates the result of proposed method.

Table A3.7. Performance evaluation parameters (%) of 10 classifiers for **F-test**.

PT	CT	SE	SP	PPV	NPV	FM	AUC
K2	LDA	84.05	69.03	83.23	70.85	83.21	84.83
	QDA	78.29	58.16	77.19	59.28	76.79	76.39
	NB	72.73	81.35	87.71	64.32	78.73	84.03
	GPC	88.93	72.75	86.52	81.47	86.95	85.41
	SVM	85.25	43.81	76.64	83.48	84.72	51.19
	ANN	82.75	82.45	88.69	73.98	85.02	89.61
	LR	45.99	48.40	59.25	35.78	50.90	45.89
	DT	85.34	51.09	76.02	66.05	79.82	71.57
	AB	81.35	82.34	89.58	71.76	84.73	88.84
	RF*	**87.29**	**69.56**	**84.15**	**75.65**	**85.22**	**88.95**
K10	LDA	84.17	63.71	78.21	75.54	79.49	79.97
	QDA	93.90	54.71	75.96	88.33	83.07	81.63
	NB	72.63	83.33	88.81	65.44	79.03	86.04
	GPC	93.78	79.87	89.49	93.12	91.10	88.94
	SVM	92.66	54.36	76.94	84.17	83.69	25.69
	ANN	79.00	73.92	88.14	60.36	82.88	87.15
	LR	56.26	51.55	66.14	42.76	59.08	54.22
	DT	83.71	69.36	83.57	69.99	82.40	80.69
	AB	86.61	70.92	85.98	78.43	85.77	88.07
	RF*	**83.65**	**81.92**	**93.45**	**69.38**	**87.30**	**92.67**
JK	LDA	97.46	94.87	97.19	95.40	97.32	99.40
	QDA	68.15	85.04	89.28	59.51	77.27	84.98
	NB	75.16	99.85	99.90	68.86	85.78	93.33
	GPC	94.66	94.17	96.74	90.67	95.68	96.78
	SVM	95.00	82.26	90.71	90.04	92.80	96.85
	ANN	93.17	85.53	92.26	87.57	92.66	96.89
	LR	92.74	81.52	90.14	86.11	91.42	93.40
	DT	99.44	90.91	95.24	99.01	97.27	99.40
	AB	95.00	97.07	98.42	91.81	96.62	99.39
	RF*	**99.82**	**99.59**	**99.78**	**99.69**	**99.80**	**99.96**

*Bold and shaded value indicates the result of proposed method.



(a) K2 protocol

(b) K10 protocol

(c) JK Protocol

Figure A3.1. Mean accuracy (%) of 10 classifiers varying p-values for 3 CV protocols: (a) K2 protocol; (b) K10 protocol; and (c) JK protocol.