COMPUTATIONAL INTELLIGENCE TECHNIQUES FOR DIABETES PREDICTION

BY

A. K. M. SAZZADUR RAHMAN

ID: 143-25-433

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

Supervised By

Dr. Syed Akhter Hossain
Professor and Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2018

APPROVAL

This Thesis titled **"Computational Intelligence Techniques for Diabetes Prediction"**, submitted by A. K. M. Sazzadur Rahman (ID: 143-25-433) to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **December 12, 2018.**

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor and Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor & Coordinator of MIS
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this thesis have been done by me under the supervision of **Dr. Syed Akhter Hossain, Professor and Head, Department of CSE** Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Syed Akhter Hossain
Professor and Head
Department of CSE
Daffodil International University

Submitted by:

Sazzad

A. K. M. Sazzadur Rahman
ID: 143-25-433
Program: M.Sc
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty Allah for His divine blessing makes me possible to complete this thesis successfully.

I am sincerely and heartily grateful to my supervisor, **Dr. Syed Akhter Hossain**, **Professor and Head,** Department of CSE Daffodil International University, Dhaka, for the support and guidance he showed me throughout the thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

I would like to express my heartiest gratitude to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledgement with due respect the constant support and patients of my parents.

ABSTRACT

This study is on “**Computational Intelligence Techniques for Diabetes Prediction**”. The main objective is to examine the performance of various Machine Learning algorithms in order to reduce the high cost of chronic disease diagnosis by prediction. A huge number of individuals in Bangladesh alone experience the ill effects of undiscovered or late-analyzed unending sicknesses, for example, Endless Kidney Ailment (CKD), Coronary illness, diabetes, Bosom Malignant growth and some more. Most of the time such types of disease diagnosis is very costly and complicated. Considering diabetes, early prediction of diabetes is an important issue in Health Care Services (HCS). So, there is a need of an application that can effectively diagnosis thousands of patient using medical specification. I examine different machine learning algorithms for predicting diabetes in real time by sketch and gather from concepts and tools in the field of machine learning. This work uses 4 classification techniques for diabetes prediction. Such as, Artificial Neural Network, Random Forest , Naive Bayes and Support Vector Machine. The performance of different classification technique was evaluated on different measurements technique. In addition, my present investigation for the most part centers around the utilization of restorative code information for infection forecast, and investigate distinctive courses for speaking to such information in my expectation calculations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLES

CHAPTER 1
INTRODUCTION

1.1 Motivation

Diabetes is a prominent disease that affects huge amount of human around the world. It is an incessant illness that happens either when the pancreas does not gather enough insulin or when the body can't successfully utilization the insulin it creates. It is growing tremendously, because of unhealthy life style, take richer and junk food and lack of physical activity. Moreover, diabetes is a key reason for visual impairment, kidney disappointment, heart assaults, stroke and lower appendage removal.

In order to discover my hypothesis and to understand how I am going to build a predicting model, I would first take a look at what causes diabetes. Lake of awareness about health and taking unhealthy food are the two major causes of this disease. It would be safe to assume that patients with diabetes are more likely to have higher blood pressure and higher sugar level. However, diabetes might have other effects on blood. As mentioned before,

It is very important for doctors to be able to diagnose diabetes in its early stages and to prescribe proper drugs that treat this condition. A simple blood test can have a lot of implications of the health and it is easily accessible and relatively cheap for public to get diagnose. So, I decided to base my model on different parameters result that can actually be implemented in real life considering its low cost and easy access. If successful the application built in the project could turnover many pharmaceutical companies and most importantly, it can save millions of lives with easy process of diagnosis.

1.2 Background

Medical Services is a big commercial aspect in every time. Revenue stream always running in this fields, so many dissatisfied clients and patients and health hanging in the balance, it looks there is an excessive need for one platform for problem solves in healthcare. Here is our main idea for improving healthcare services is to place more highlighting on prevention and less on treatment. If we think about every health issues, not every health issue is preventable, but in many cases, early observation and detection can pointer to good health outcomes and reduce costs. In addition, the key elements of preventive healthcare are the disease monitor. Early monitoring of diseases can be diagnosed while still in the early, curable phase. However, it is not viable for every patient to get monitor for every probable disease. A better arrangement is than have a reasonable, open, and solid methods for ascertaining sickness chance and anticipating illness presence.

1.3 Machine Learning on Health Care

Disease prediction is very important for medical and health care center in order to make the best possible accurate decisions. Machine learning is tied in with creating scientific, computational, and factual procedures for discovering designs in and separating understanding from information. Information, thusly, are the solid signs of structures and procedures that shape the world. Machine learning research intends to open innovations that can tackle unmanageable issues and change human life in a wide range of territories [3]. Therefore, Computational Intelligence techniques has been effectively related to resolve various important medical also with biological problems. In the first place, the issue is defined as a grouping based issues. Whereas, the undertaking is to take in classifiers from preparing information.. And training classifiers can be predicting the disease. Machine learning algorithms built on regularly collected medical data have the prospective to build just such an application. In this work, I discuss and test several computational methods to this end. In this present study, I focus on chronic diseases, especially Chronic Diabetes Mellitus.

1.4 Research Question

The final set of articles was used to answer the following questions: 1. which is the best Machine Learning techniques for diabetes prediction?

1.5 Objectives

To aim of the study is performance measurements of various Machine Learning algorithms to reduce the high budget of diabetes disease diagnosis by prediction.

1.6 Report Layout

There are six chapters in this thesis paper: Introduction, Literature Review, Materials & Methodology, Description of the Classification Techniques, Analysis & Discussion and Conclusion & Future Work.

Chapter 1: Objectives of this thesis, motivation behind this thesis, background of this thesis, machine learning on health care and report layout.

Chapter 2: Describes the thesis works which are related in the field of question assessment and other important works done by prominent researchers. Comparative study of analysis section discuss about the key of points each article.

Chapter 3: Describes the materials and system architecture which I used for data analysis.

Chapter 4: Describes the different classification of techniques.

Chapter 5: Describes the prediction experiences of four machine learning techniques were investigated for the diabetes prediction.

Chapter 6: The conclusion section enlighten shortly on the analysis and experiments that has been done throughout all the previous chapters, future scope are the discussion section where future possibilities and potential of this thesis has been highlighted.

CHAPTER 2
LITERATURE REVIEW

2.1 Chronic Diabetes Mellitus (CDM)

Chronic diseases are the prominent causes of death and disability worldwide. Disease rates from these conditions are quickening worldwide, advancing crossways every country and permeating all socioeconomic peoples.

Diabetes is an interminable, metabolic illness portrayed by raised dimensions of glucose, which alludes to genuine damage to the heart, veins, eyes, kidneys, and nerve. In diabetes, Age do not depend on this occurrence. There are three types of diabetes [2]. The highly popular is type 2 diabetes, usually it happens on adults. Hence, the body becomes impervious to insulin or doesn't generate enough amount of insulin. Furthermore, In the history of past decades occurrence of type II diabetes has increased theatrically in developing nations of all levels. Type 1 diabetes, when known as adolescent diabetes or youth diabetes, is a ceaseless condition in which is because of the absence of insulin generation. Therefore, Type II Diabetes, is a very popular form of diabetes and it contains huge amount of people in the whole world. And type III is Gestational diabetes. It happens, because of changes about hormones when patients absorbed pregnancy. But early detection and prediction can be prevented it.

2.2 Article Searching Procedure

I used a systematic searching procedure to identify all of the available articles that discuss the disease prediction and especially diabetes prediction using machine learning techniques. In my systematic procedure, I search two keywords from Science Direct databases in order to access the article. I used the keywords, “disease prediction using machine learning” and “diabetes prediction using machine learning” to find journal articles published in English Language between years 2016 to 2018.

2.3 Article Inclusion and Exclusion Benchmark

I used some benchmark to include and exclude articles from the set of articles that were selected through the search of Science Direct databases. To include and exclude articles from the set of articles found through our systematic searching technique, I read the title, abstract, methodology and results of each article. I considered only those article that were written in English and that used machine learning. The exclusion criteria were the following: 1) Article that applied Machine learning for disease prediction, 2) Diabetes prediction using Machine Learning.

2.4 Data Extraction

I carefully read and analyzed all of include articles to collect the key information. I followed the standard data extraction from for the particular analysis of each article. Each article was evaluated for the following information: 1) Performance comparison for different Machine Learning algorithms, 2) Disease prediction using Machine Learning technique, 3) Working with diabetes datasets.

2.5 Article Search Result

I utilized my methodical article seeking system and discovered 12 article that have distributed in presumed diaries and gatherings. Therefore, I skimmed all the selected articles particularly and identified the main aspects. Therefore, article search result is summarized in figure 2.1.

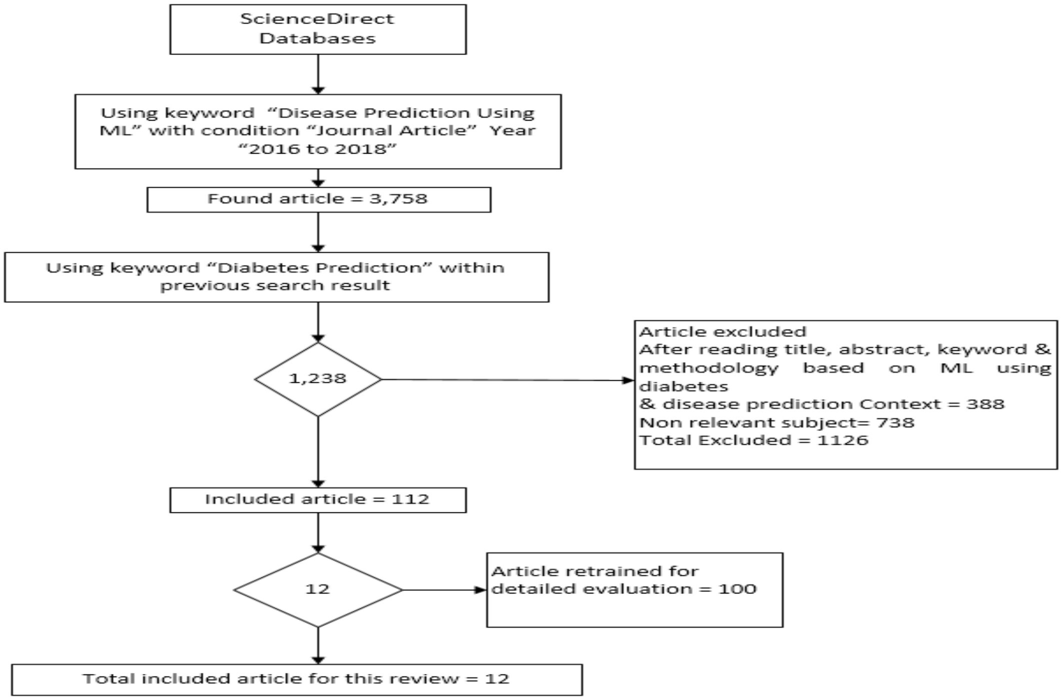

ScienceDirect Databases
Using keyword "Disease Prediction Using ML" with condition "Journal Article" Year "2016 to 2018"
Found article = 3,758
Using keyword "Diabetes Prediction" within previous search result
1,238
Article excluded After reading title, abstract, keyword & methodology based on ML using diabetes & disease prediction Context = 388 Non relevant subject= 738 Total Excluded = 1126
Included article = 112
12
Article retrained for detailed evaluation = 100
Total included article for this review = 12

Figure 2.1: Article Search Result

2.6 Article Key points Discussion

From the selected of 12 articles, I discovered 8 articles that were identified with diabetes expectation using machine learning techniques and 4 articles that discussed disease prediction using machine learning. However, the outcomes of the 12 articles on Machine Learning used in disease prediction are summarized in Table 2.1

Table 2.1: Key points of each article

Ref	Approach	Study
[11]	Methodical survey of the uses of machine learning, information mining systems and devices in the field of diabetes inquire about.	The investigation portrays the precision 85% of those utilized were described by directed learning approaches and 15% by unsupervised. SVM emerge as the best and generally utilized calculation
[12]	The intention of this investigation is to plan a model which can guess the probability of	Near examination on three sorts ML calculation. What's more, Guileless

	diabetes in patients with most extreme precision.	Bayes outflanks with the most astounding precision of 76.30% similarly different calculations
[13]	Their proposed model is an endeavor to assess demonstrative legitimacy of an old correlative and elective drug procedure, iridology for finding of sort 2 diabetes utilizing delicate processing	The outcomes indicate best grouping exactness of 89.63% determined from RF classifier
[14]	Specifically, the point is to assist doctors with identifying the important SNPs identified with Sort 2 diabetes, and to assemble a choice help apparatus for hazard expectation.	The Arbitrary Woods is a helpful technique for learning prescient models and the significance of SNPs with no fundamental presumption
[15]	In this paper creators are proposed a strategy ready to arrange patients influenced by diabetes and so as to help and to quicken the determination of diabetes	They demonstrated the estimation of exactness equivalent to 0.770 and a review equivalent to 0.775 utilizing the Hoeffding Tree calculation
[16]	They proposed another learning based framework for illnesses expectation utilizing bunching, commotion evacuation, and forecast techniques.	The contemplate demonstrated that the mix of fluffy guideline based, Truck with clamor expulsion and grouping procedures can be compelling in ailments expectation from certifiable medicinal datasets.
[17]	To characterize diabetes, a few order strategies are utilized, for example, direct discriminant investigation (LDA), quadratic discriminant examination (QDA), and Innocent Bayes (NB)	Their machine learning framework demonstrates the execution of GP-based model as: ACC 81.97%, SE 91.79%, SP 63.33%, PPV 84.91% and NPV 62.50% which are bigger contrasted with different techniques.
[18]	In this investigation, A propose structure for distinguishing subjects with and without T2DM from EHR by means of highlight building and machine learning.	Their proposed system shows a progressively precise and proficient methodology for recognizing subjects with and without T2DM from EHR.
[19]	In this investigation, they are proposed a novel model dependent on information digging procedures for anticipating type 2 diabetes mellitus (T2DM).	The fundamental issues that they are endeavoring to unravel are to enhance the exactness of the expectation show, and to make the model versatile to more than one dataset.
[20]	This paper introduces a study on the use of highlight choice and order procedures for the conclusion and forecast of unending diseases	This work exhibits a far reaching diagram of different component determination strategies and their characteristic upsides and downsides

[21]	They utilized diverse characterization strategies for the expectation of bosom malignant growth survival and metastasis.	Their finding demonstrated that the SVM beat other machine learning techniques in forecast of survival of the patients regarding a few criteria
[22]	In this examination audit, creators are assessed the capability of ML for neurosurgical result prediction	In the exploration setting, ML has been considered broadly, exhibiting an astounding execution in result expectation for an extensive variety of neurosurgical conditions

I studied particularly the selected 12 articles on “Diabetes & Disease prediction with machine learning algorithms” for the analysis of the present processes to identify future research in the area. The current study present a summary of the data found in the literature that focused on performance evaluation on machine learning classification techniques. In this present study, I provide evaluation of different classification algorithms on machine learning. L Kavakiotis et al. [11] showed their study, for machine learning data mining techniques and tools in the field of diabetes research the experiments describes the accuracy 72% of those used were characterized by supervised learning approaches and 15% by unsupervised. SVM arise as the most successful and widely used algorithm. Two studies show that the comparative analysis on different machine learning algorithm [12, 13]. The result show best accuracy of 89.63% calculated from RF classifier than other algorithms. B. Lopez et al. [14] conduct an experiment on types II diabetes and they showed RF is a useful method for learning predictive models. Two of the studies provide a comparison between different classification algorithms [15, 17]. They showed the value of precision equal to 0.77 and recall equal to 0.78 using the Hoeffing Tree algorithm and their [17] machine learning system shows the performance of GP-based model as: ACC 81.97%, SE 91.79%, SP 63.33%, PPV 84.91% and NPV 62.50% which are larger compared to other methods. Three of studies [16,18, 19], they proposed a new knowledge-based system for diseases prediction using clustering, noise removal, and prediction techniques. Moreover, other three studies of selected literature [20-22], they presented survey, comparison between different classification algorithms. This work [20] presents a comprehensive overview of various feature selection methods and their inherent pros and cons. Their finding showed that the SVM outperformed other machine learning methods in prediction of survival of the patients in terms of several criteria [21].

In the examination setting, ML has been contemplated broadly, exhibiting a superb execution in result forecast for an extensive variety of neurosurgical conditions [22]. Three examinations [9-11] gave a suggestion on the reasonable Machine learning calculations for diabetes forecast. I. Kavakiotis et al [9] experimented between various regulated learning and utilizing diabetes datasets. They indicated SVM emerge as the best and best entertainer calculation. Other two investigations [10-11] portrayed into their article, Naive Bayes outflanks with the most noteworthy exactness of 76.30% relatively different calculations and Random backwoods acquired best precision is 89.63% than others, individually.

CHAPTER 3
MATERIALS & METHODOLOGY

3.1 Data Collection and Feature Selection

In this literature, I use the patient data from Prima Indian Diabetes Datasets provided by the University of California, Irvine (also known as UCI Machine Learning Repository). In addition, this dataset is initially from the National Institute of Diabetes and Digestive and Kidney Diseases. The goal of the dataset is to indicatively anticipate regardless of whether a patient has diabetes, in light of certain demonstrative parameters incorporated into the dataset [4] . Moreover, the datasets contain of some particular medical variables. For precedents, pregnancies record, BMI, insulin level, age, glucose focus, diastolic circulatory strain, triceps skin overlay thickness, diabetes family work This dataset has 768 female patient's data at least 21 years old. Hence, There are number of true cases 268 (34.90%) and number of false cases 500 (65.10%). In the following, I choose eight particular parameters for data analysis such as,

i) Pregnancies: available pregnancy records.

ii) Glucose: Plasma glucose concentration with 2hrs in OGTT

iii) Patient's BP

iv) Skin Thickness

v) Insulin: each patient, 2-Hour level of serum insulin record (mu U/ml)

vi) BMI (Body Mass Calculation)

vii) Diabetes pedigree function

viii) Age: Mostly adults (years)

3.2 System Architecture

The proposed system architecture machine learning based for diabetes prediction is shown in Figure 3.1

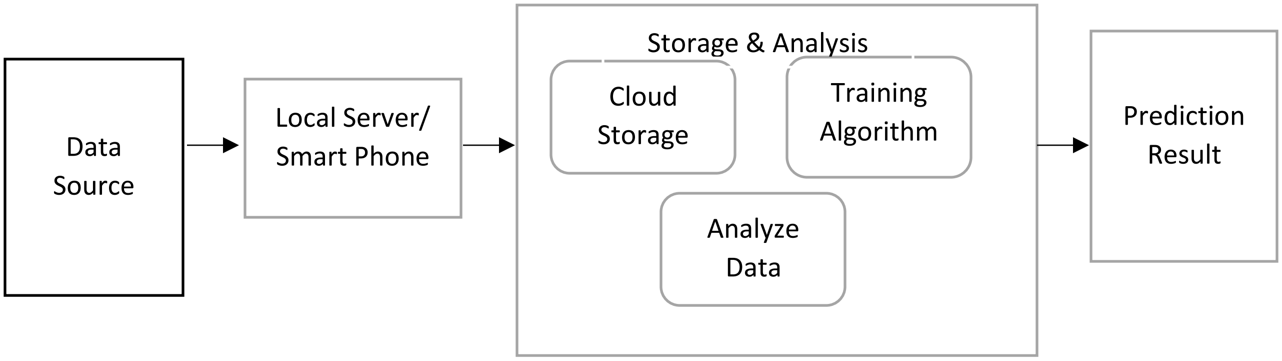


Figure 3.1: The proposed system architecture

The proposed framework helps in efficient decision making process and provides an effective solution for diabetes prediction and monitoring. Considering the enormous growth of the disease data, the proposed model aims to handle this issue effectively by cloud application.

The proposed system architecture for the prediction system is a part of real Health Care Services (HCS). I decided to interpret the diabetes prediction results into prediction result. As shown in Figure 3.1, the health tracking devices and sensors are used in order to generate different types of health data such as blood pressure, step count, checkup history etc. The health sensing and tracking devices are connected to each local server or smart phone, which are able to process data from the sensing devices. Once the data is processed by local server or a smart phone, it can be sent to the cloud application phase. In the cloud application phase, the processed data from local server are stored and analyzed with ML kit (training algorithm), and evaluated by the pre-trained model (trained data) for more accurate real time diabetes prediction. Moreover, the real time prediction service users can see and be notified with their daily health activities and even record them among their smart devices or mobile phones.

CHAPTER 4
DESCRIPTION OF THE MACHINE LEARNING TECHNIQUES

4.1 Artificial Neural Network

Artificial neural networks (ANN) is an important machine learning technique for biological research. In machine learning, ANN is a convenient computational models which works similar to biological neurons [5]. Elementary structure of ANN is a collection of linked nodes. Moreover, these nodes help to perform as neurons in ANN. Considering the nodes are connected by a link and each link has some weight. Moreover, ANN as like to brain, learn through samples and experiences not from already defined instructions through programs. ANN manually learns from the instances and experiences it-self and then apply the learnings on analyses.

ANN mainly organized into three layers; i) Input layer (Nodes can take input data), ii) Hidden layer or processing stage (transferred input data from input layer) and iii) output layer (results are sent from the hidden layer). In addition, the result of output layer at each node is called its activation or node value [6].

4.2 Random Forest (RF)

Random Forest is an assembling method and one of the most popular and powerful algorithm in Machine Learning era. There is an immediate connection between the quantity of trees in the timberland and the outcomes it can get: the bigger the quantity of trees, the more exact the outcome. Random Forest (RF) is a well-known supervised classification algorithm which is able to performing both regression and classification problems. RF has been first proposed by Leo Bierman [7]. In generally, RF constructs several decision trees and combines them together to acquire more accurate and efficient prediction. These techniques add an extra layer of randomness to bagging. Moreover, the random-forest algorithm fetches a subset of predictors randomly preferred at that node when the trees split.

4.3 Naive Bayes (NB)

Naive Bayes algorithm can be defined as a supervised classification algorithm in machine learning which is based on Bayes theorem with a hypothesis of individuality among features. Naive Bayes classifier is simple but most operative algorithm for the classification problem analysis. Naive Bayes are statistical classifiers does that by making an hypothesis of conditional independence with the training datasets [8]. Henceforth, Naive Bayes classifier is the appropriate classification technique for verdict best solution from a dataset whereas given different object into predefined groups

4.4 Support Vector Machine (SVM)

Support vector machine (SVM) are managed realizing which depends on straight order. SVM function admirably for some, human services issues and can take care of straight and non-direct issues moreover. For taking care of relapse and grouping issue effectively SVM perform superior to other characterization strategies. Along these lines, Vladimir Vapnik and Alexey Chervonenkis [9] [10] presented the help vector machine order method which is endeavor to pass a straightly divisible hyper plane to group the datasets into two classes. Support Vector Machines is basic: The calculation makes a line which isolates the classes on the off chance that e.g. in an arrangement issue. The objective of the line is to boosting the edge between the focuses on either side of the alleged choice line. At long last, the model can without a doubt gauge the objective gatherings (names) for new cases.

CHAPTER 5
ANALYSIS & DISCUSSION

5.1 Measurement of Classification Techniques

In this study, I used 10-fold validation technique to measure the execution of each classification algorithm. Presentation of all ML algorithm are considered by different statistical measurement aspects such as accuracy, sensitivity, specificity, NPV, PPV etc. These classification measurement factors are calculated by the terms: True Positive , False Positive , True Negative and False Negative. Here,

True Positive: Prediction results are yes and the patient have diabetes.

True Negative : Prediction results are no and the patient do not have diabetes.

False Positive: Prediction results are yes but the patient do not actually have the diabetes (Also known as a “Type 1 error”).

False Negative: Prediction results are no but the patient have diabetes (Also known as a “Type 2 error”).

The computation formula of the measurement factors are as follows,

Accuracy in classification problems is the ratio of correct predictions made by the model over all kinds of suitable predictions completed.

Accuracy = (TP+TN) / (TP+FP+FN+TN) (1)

TPR, sensitivity, or recall defined here is a measure that tells us what ratio of positive instances that actually have diabetes with the actual positive instances (patient having diabetes are TP and FN).

TPR = = TP / (TP+FN) (2)

True negative rate is a measure which defines the ratio of the patients that do not have diabetes, and also predicted by the model as non-diabetes. In addition, specificity is the suitable opposite of recall.

TNR = TN / (TN+FP) (3)

Positive predictive value or precision is the number of accurate positive scores divided by the number of positive scores predicted by the classification algorithm.

Precision = TP / (TP+FP)

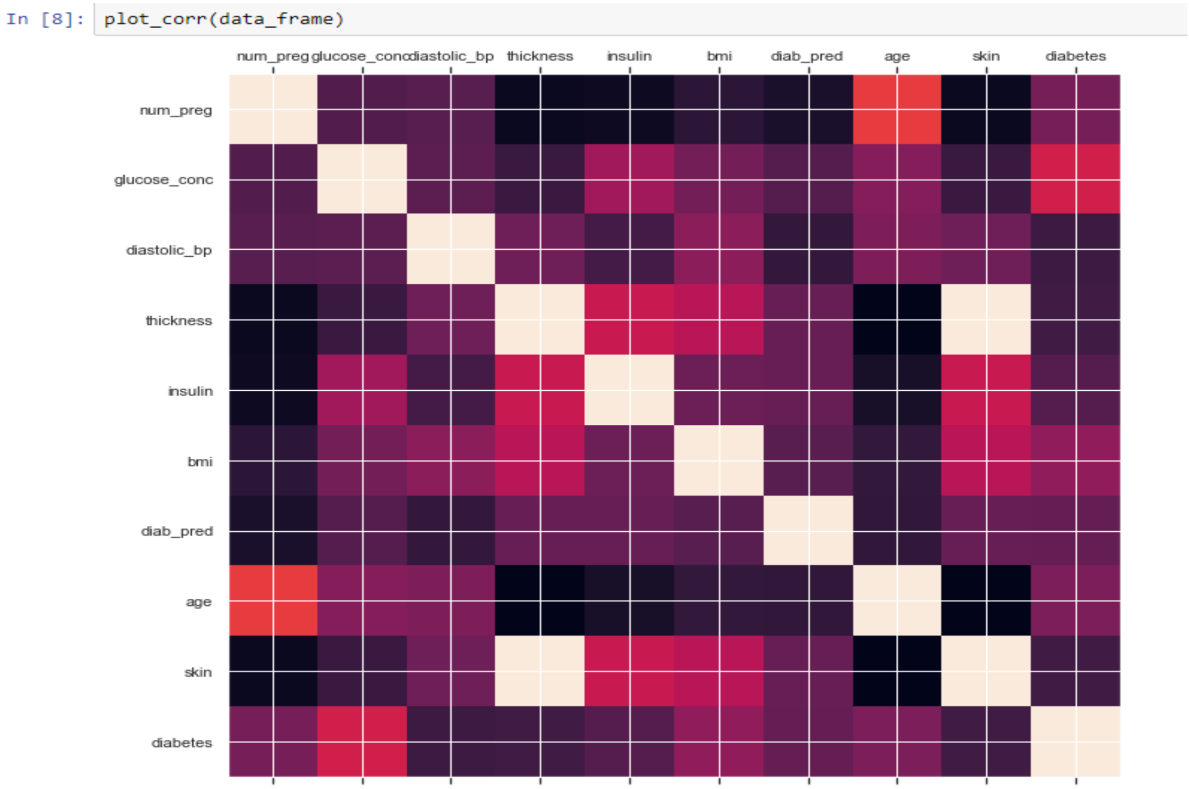
F1 measure is a weighted norm of the recall & precision. For good performance of the classification algorithm, it must be one and for the bad performance, it must be zero.

F1= 2* (Recall * Precision) / (Recall + Precision)

5.2 The performance result

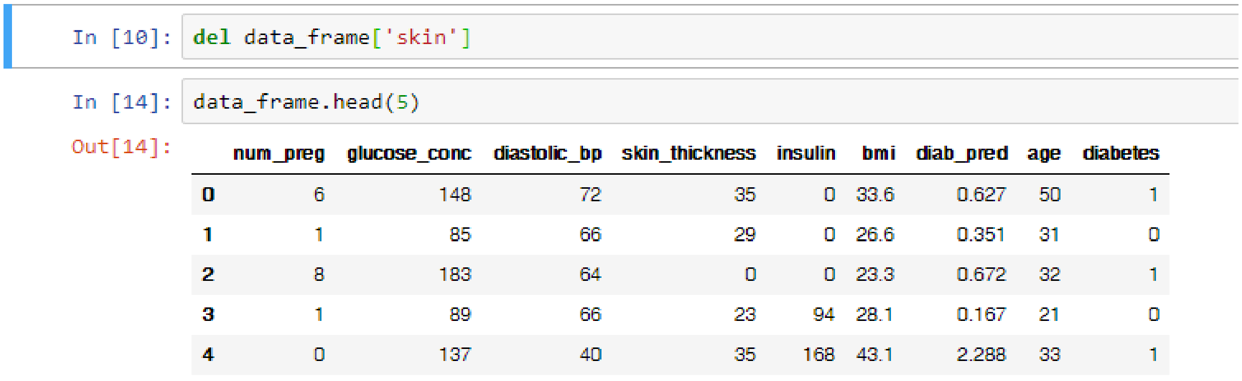
The prediction experiences of four machine learning techniques were investigated for the diabetes prediction. I analyzed data from 768 samples with 80% for training and 20% for testing. In the study, from this datasets containing the original true case 34.90% , original false case 65.10% , training true and false are 34.69% , 65.31% and test true and false case are 35.71% & 64.29% , respectively. Moreover, the datasets evaluated by the different statistical methods. Such as mean, median, standard deviation etc. The statistical evaluation is presented in figure 5.3. Furthermore, the datasets was also checked to verify the correlated values in order to drop the duplicate values.

I found the two column are correlated thickness and skin whereas 1 to 1. Hence, I dropped by the duplicate skin column by del function. The heatmap shown in figure 5.1 whereas the correlated column were occur. Therefore, figure 5.2 were presented, here is no corelated value.



```
In [8]: plot_corr(data_frame)
```

Figure 5.1: Heat map for checking correlated columns



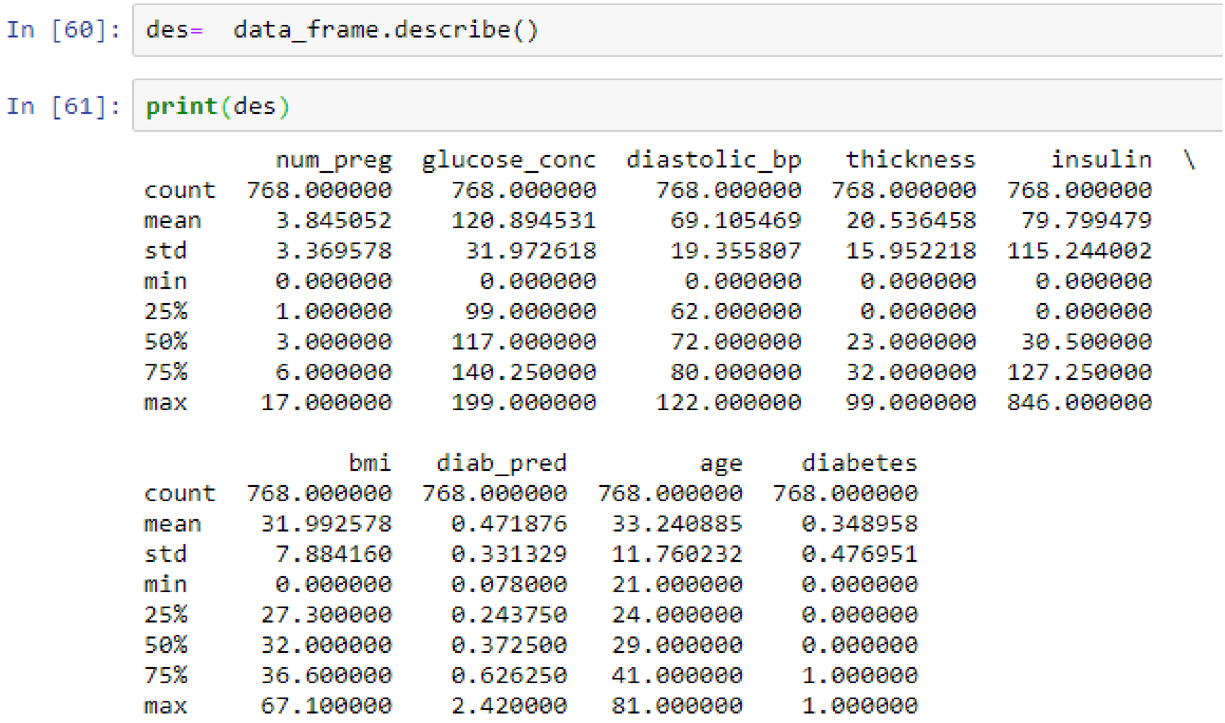
```
In [10]: del data_frame['skin']
```

```
In [14]: data_frame.head(5)
```

Out[14]:

	num_preg	glucose_conc	diastolic_bp	skin_thickness	insulin	bmi	diab_pred	age	diabetes
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 5.2: Dropped redundant column from datasets



```
In [60]: des=  data_frame.describe()
```

```
In [61]: print(des)
```

```
         num_preg  glucose_conc  diastolic_bp   thickness     insulin  \
count  768.000000    768.000000    768.000000  768.000000  768.000000   
mean     3.845052    120.894531     69.105469   20.536458   79.799479   
std      3.369578     31.972618     19.355807   15.952218  115.244002   
min      0.000000      0.000000      0.000000    0.000000    0.000000   
25%      1.000000     99.000000     62.000000    0.000000    0.000000   
50%      3.000000    117.000000     72.000000   23.000000   30.500000   
75%      6.000000    140.250000     80.000000   32.000000  127.250000   
max     17.000000    199.000000    122.000000   99.000000  846.000000   

              bmi   diab_pred         age    diabetes  
count  768.000000  768.000000  768.000000  768.000000  
mean    31.992578    0.471876   33.240885    0.348958  
std      7.884160    0.331329   11.760232    0.476951  
min      0.000000    0.078000   21.000000    0.000000  
25%     27.300000    0.243750   24.000000    0.000000  
50%     32.000000    0.372500   29.000000    0.000000  
75%     36.600000    0.626250   41.000000    1.000000  
max     67.100000    2.420000   81.000000    1.000000  
```

Figure 5.3: Statistical results from diabetes datasets

In this examination, I consider diverse investigation to invesigate the four machine learning calculations for the orders of Diabetes Datasets. Morover, from the diabetes datasets the majority of the examples are assessed by 10 crease cross approval procedures. Figure 5.4 demonstrates the disarray grid (Classification aftereffects of the computational Intelliegence for pediction of diabetes. Here, TP genuine positive, FP false positive, TN genuine negative, FP bogus positive) of the four classificaton calculations. Figure 5.5 present the exactness of four regulated based characterizations methods. Henceforth, SVM accomplished the best accuaracy (i.e. 76%) and ANN performed most exceedingly awful (i.e. 72%) . In addition, NB and Random Forest are nearly accomplished relatively same exactness (i.e. 74% and 73%, individually).

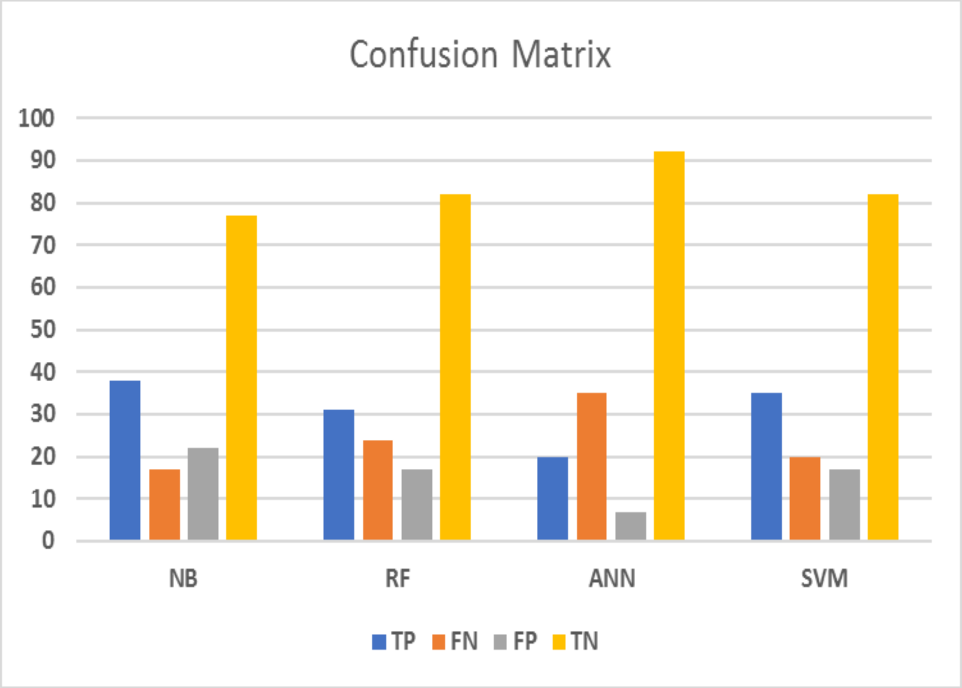


Figure 5.4: Confusion matrix of classification algorithms

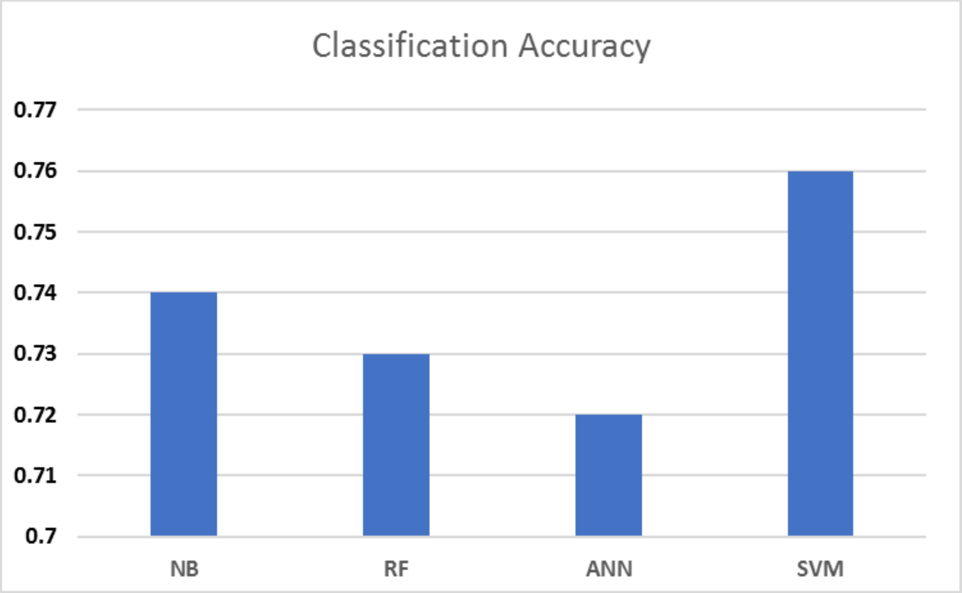


Figure 5.5: Classification accuracy of four classifiers

5.3 Performance Evaluation

Results of all selected classifiers are presented in figure 5.6 and table 5.2, according to their sensitivity, precision, f1 measure and specificity. With respect to the precision, SVM achieved the high performance (it's 0.90). and Random Forest performed poorest (it's 0.77). However, when considered the Sensitivity, ANN achieved the outperform (87%) and SVM showed lowest performance (77%). In addition, ANN is the best performer in terms of f1 measure.

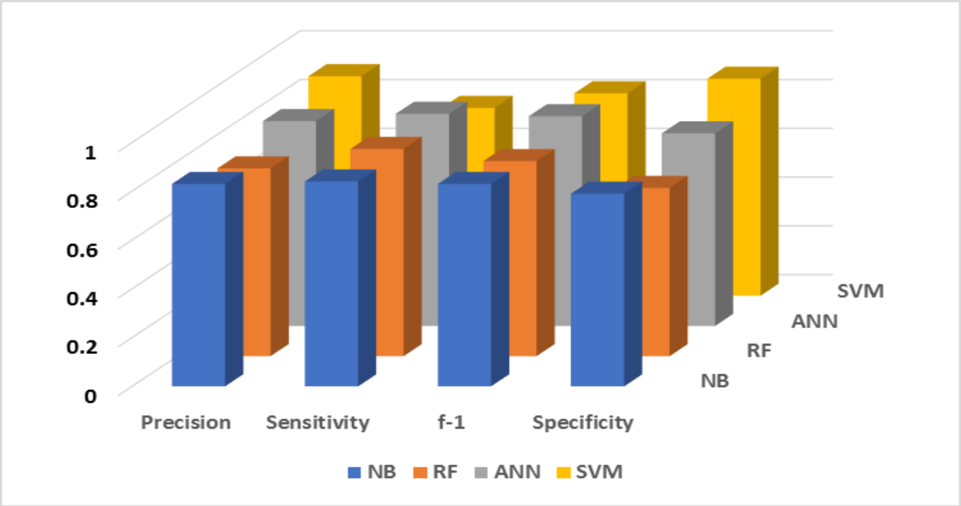


Figure 5.6: Classification Performance of Machine Learning Techniques

Table 5.1: Classification Performance Measurements

Measurement Techniques	NB	RF	ANN	SVM
Precision	0.83	0.77	0.84	0.9
Sensitivity	0.84	0.85	0.87	0.77
f-1	0.83	0.8	0.86	0.83
Specificity	0.79	0.69	0.79	0.89

Furthermore, Random forest showed the worst performance in terms of f1 measure. By taking a gander at RF and SVM groupings strategies, we can see that their Specificity scores the most reduced and most astounding. It's 0.69 and 0.89, independently.

Most of the classification's techniques showed the accuracy level above 70% which is indicates that the performance of our selected algorithms is good. ROC shows, the True Positive Rate and False Positive Rate of the classifiers from the diabetes data analysis. The area under the ROC must be close to one for the best classification techniques. Figure 5.7 represents the ROC for the selected four classification algorithms.

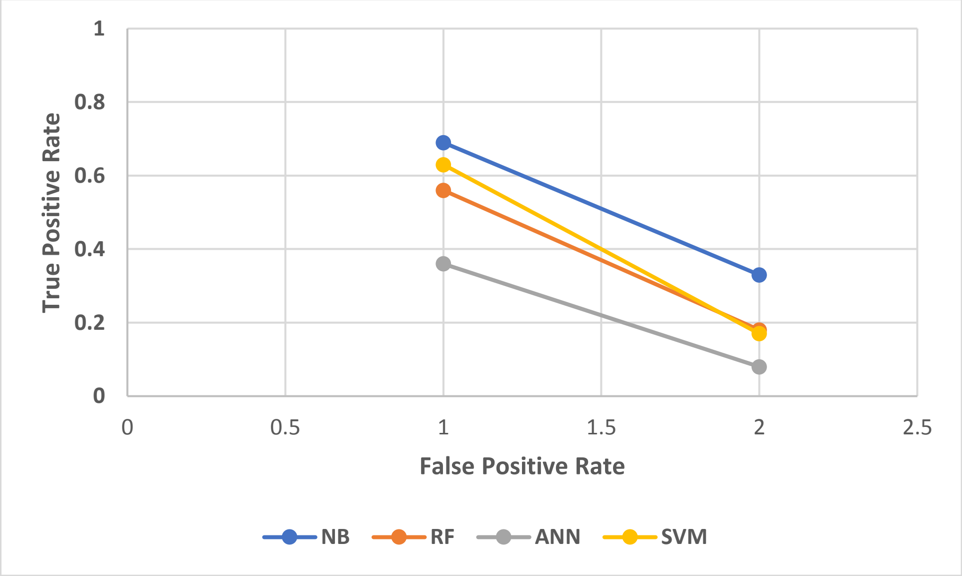


Figure 5.7: Receiver Operating Curve for four classification algorithms

In summary, I highlight research directions and challenges in relation to diabetes prediction and alone with disease prediction through Machine Learning algorithms. Which is emerging impact of disease prediction and health care services. In Machine Learning classifiers, performance and classification issues can be more improved. Here, I describe most popular supervised learning algorithms that require further research in terms of Machine Learning and Health Care fields.

CHAPTER 6
CONCLUSION & FUTURE WORK

6.1 Conclusion

The major contribution of this work are as follows: First, I carefully read and analyzed all of included articles to minutes the key information for literature review. Secondly, I present some experimental comparison between four classification algorithms. Also, this study has involved different classification techniques in the perdition of diabetes based on various medical parameters, specifically it's based on eight parameters. In addition, support vector machine outperformed all other techniques with the high accuracy (76%) and precision (90%).

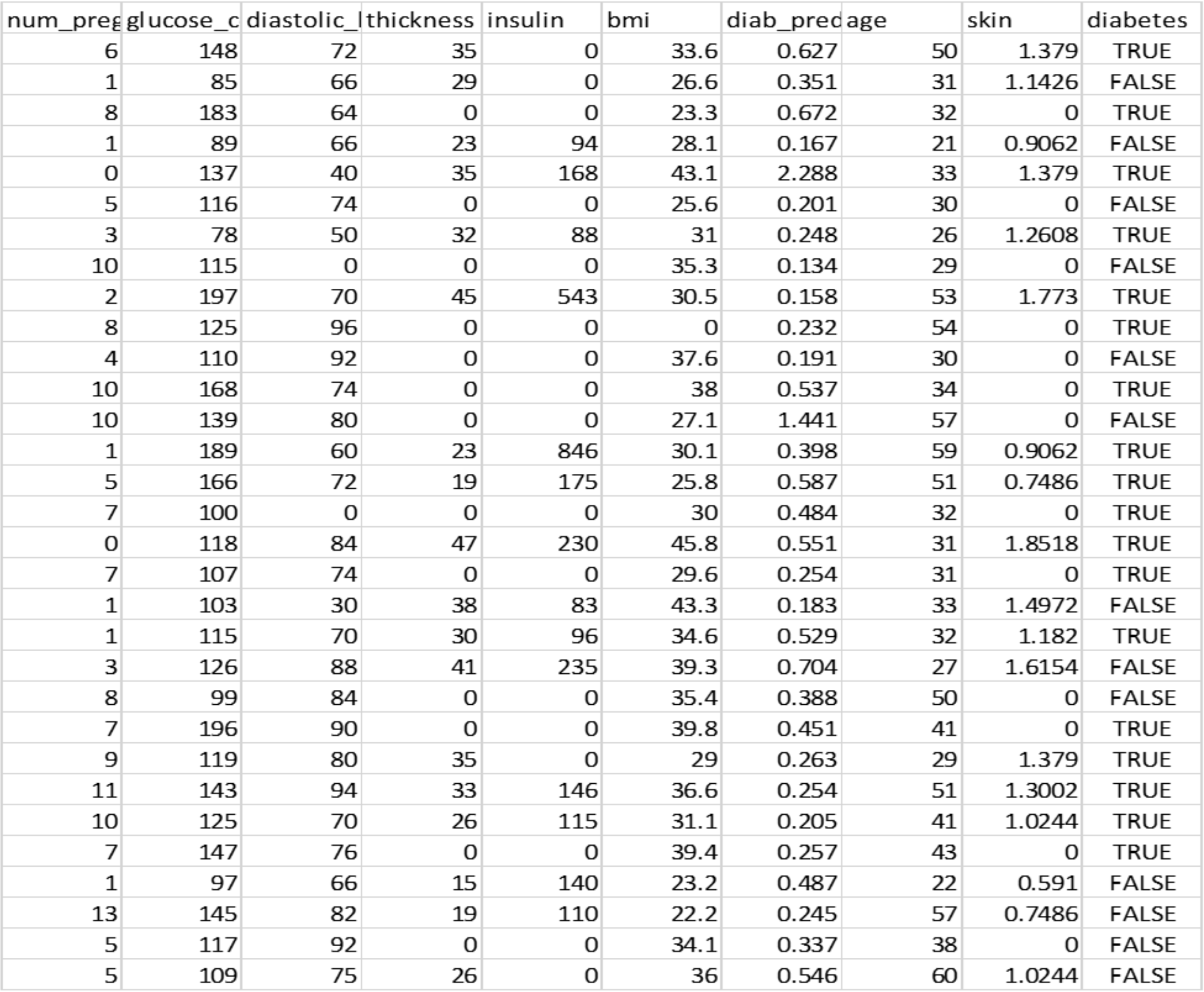
6.2 Future Work

In my examinations, identified with most work in the writing, every twofold classifier was prepared and assessed on a preparation set that incorporates both positive and negative examples. Moreover, my present work can be helpful for diabetes deduction by collecting data from different places and can deliver more accurate results for diabetes prediction. This work can be considered for operation application for diabetes prediction. There are several directions for future work. We only investigated to some machine learning algorithms, it can be choosing more algorithm for build the accurate model of diabetes prediction and performance can be more improved.

Finally, just predicting if someone will get a disease within a certain period of time might not be good enough. Hence, predicting when the disease is likely to occur can accelerate more targeted and effective protective care and primary treatment.

APPENDIX

Appendix A: Pima Indian diabetes dataset



num_preg	glucose_c	diastolic_	thickness	insulin	bmi	diab_pred	age	skin	diabetes
6	148	72	35	0	33.6	0.627	50	1.379	TRUE
1	85	66	29	0	26.6	0.351	31	1.1426	FALSE
8	183	64	0	0	23.3	0.672	32	0	TRUE
1	89	66	23	94	28.1	0.167	21	0.9062	FALSE
0	137	40	35	168	43.1	2.288	33	1.379	TRUE
5	116	74	0	0	25.6	0.201	30	0	FALSE
3	78	50	32	88	31	0.248	26	1.2608	TRUE
10	115	0	0	0	35.3	0.134	29	0	FALSE
2	197	70	45	543	30.5	0.158	53	1.773	TRUE
8	125	96	0	0	0	0.232	54	0	TRUE
4	110	92	0	0	37.6	0.191	30	0	FALSE
10	168	74	0	0	38	0.537	34	0	TRUE
10	139	80	0	0	27.1	1.441	57	0	FALSE
1	189	60	23	846	30.1	0.398	59	0.9062	TRUE
5	166	72	19	175	25.8	0.587	51	0.7486	TRUE
7	100	0	0	0	30	0.484	32	0	TRUE
0	118	84	47	230	45.8	0.551	31	1.8518	TRUE
7	107	74	0	0	29.6	0.254	31	0	TRUE
1	103	30	38	83	43.3	0.183	33	1.4972	FALSE
1	115	70	30	96	34.6	0.529	32	1.182	TRUE
3	126	88	41	235	39.3	0.704	27	1.6154	FALSE
8	99	84	0	0	35.4	0.388	50	0	FALSE
7	196	90	0	0	39.8	0.451	41	0	TRUE
9	119	80	35	0	29	0.263	29	1.379	TRUE
11	143	94	33	146	36.6	0.254	51	1.3002	TRUE
10	125	70	26	115	31.1	0.205	41	1.0244	TRUE
7	147	76	0	0	39.4	0.257	43	0	TRUE
1	97	66	15	140	23.2	0.487	22	0.591	FALSE
13	145	82	19	110	22.2	0.245	57	0.7486	FALSE
5	117	92	0	0	34.1	0.337	38	0	FALSE
5	109	75	26	0	36	0.546	60	1.0244	FALSE

REFERENCES

[1] "How Many People Have Diabetes?" [Online]. Available: https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-many-people-have-diabetes/. [Accessed: 08-Jun-2018].

[2] "Diabetes," 2017. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: 08-Jun-2018].

[3] A. L. Tarca, V. J. Carey, X. Chen, R. Romero, and S. Drăghici, "Machine Learning and Its Applications to Biology," PLoS Comput. Biol., vol. 3, no. 6, p. e116, 2007.

[4] "Research Summary | National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)." [Online]. Available: https://www.niddk.nih.gov/about-niddk/staff-directory/intramural/leslie-baier/Pages/research-summary.aspx. [Accessed: 08-Jun-2018].

[5] M. van Gerven and S. Bohte, "Editorial: Artificial Neural Networks as Models of Neural Information Processing," Front. Comput. Neurosci., vol. 11, p. 114, Dec. 2017.

[6] R. HECHT-NIELSEN, "Theory of the Backpropagation Neural Network**Based on 'nonindent' by Robert Hecht-Nielsen, which appeared in Proceedings of the International Joint Conference on Neural Networks 1, 593–611, June 1989. © 1989 IEEE.," in Neural Networks for Perception, Elsevier, 1992, pp. 65–93.

[7] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[8] K. M. Leung, "Naive bayesian classifier," Polytech. Univ. Dep. Comput. Sci. Risk Eng., 2007.

[9] V. Vapnik, I. Guyon, T. H.-M. Learn, and undefined 1995, "Support vector machines," statweb.stanford.edu.

[10] A. Y. Chervonenkis, "Early History of Support Vector Machines," in Empirical Inference, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–20.

[11] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, Jan. 2017.

[12] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, pp. 1578–1585, Jan. 2018.

[13] P. Samant and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images," Comput. Methods Programs Biomed., vol. 157, pp. 121–128, Apr. 2018.

[14] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, "Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction," Artif. Intell. Med., vol. 85, pp. 43–49, Apr. 2018.

[15] F. Mercaldo, V. Nardone, and A. Santone, “Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques,” Procedia Comput. Sci., vol. 112, pp. 2519–2528, Jan. 2017.

[16] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, “An analytical method for diseases prediction using machine learning techniques,” Comput. Chem. Eng., vol. 106, pp. 212–223, Nov. 2017.

[17] M. Maniruzzaman et al., “Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm,” Comput. Methods Programs Biomed., vol. 152, pp. 23–34, Dec. 2017.

[18] T. Zheng et al., “A machine learning-based framework to identify type 2 diabetes through electronic health records,” Int. J. Med. Inform., vol. 97, pp. 120–127, Jan. 2017.

[19] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining,” Informatics Med. Unlocked, vol. 10, pp. 100–107, Jan. 2018.

[20] D. Jain and V. Singh, “Feature selection and classification systems for chronic disease prediction: A review,” Egypt. Informatics J., Apr. 2018.

[21] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, “Prediction of survival and metastasis in breast cancer patients using machine learning classifiers,” Clin. Epidemiol. Glob. Heal., Oct. 2018.

[22] J. T. Senders et al., “Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review,” World Neurosurg., vol. 109, p. 476–486.e1, Jan. 2018.