OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

IMPROVE CLASSIFICATION PERFORMANCE IN DIABETES PREDICTION

Shraddha C Chauhan[1],V S Karvande[2]

P.G. Student, Computer Science & Engineering, Everest Educational Society's Group of Institutions, Maharashtra, India[1]

HOD, Computer Science & Engineering, Everest Educational Society's Group of Institutions, Maharashtra, India[2]

Abstract: The system allows the user to make a use of algorithms to predict the risk of diabetes mellitus in human body. The various classification models such as Decision Tree, artificial Neural Networks, Logistic Regression, Association rules and Naive Bayes are used in this system. Then the Random Forest technique is used to find the accuracy of each model in the project. The dataset used is the Pima Indians Diabetes Data Set, which has the information of patients, some of them have developing diabetes therefore, this project is aimed to create a mobile application for predicting persons class whether present in of the diabetes risk or not.

Keywords: Diabetes Mellitus, Diabetes Prediction, International Diabetes Federa tion, Electronic Medical Records

----------∴∴∴----------

I INTRODUCTION

Diabetes mellitus (DM) is a chronic non-communicable disease. The disease has been closely followed by World Health Organization (WHO) and International Diabetes Federation (IDF) since worldwide number of diabetes increase continuously. It was found that there were 387 million people with diabetes in year 2014 and have a tendency to be 592 million patients in the next 20 years. IDF also found that almost half of diabetes in South East Asia is undiagnosed. According to these amounts, the disease should be controlled and properly maintained for efficient and sustainable prevention. The annual report 2013 of Department of Disease Control, Ministry of Public Health, reports that diabetes is the top three of chronic non-communicable diseases in Thailand.

The statistics shows that 1 of 13 adults Thais had diabetes and the total number of people with diabetes is not less than 3 million. In the future, there will be more than 7 million people are at risk of diabetes. The report also indicates that the number of diabetes is likely to be increased every year. Consider diabetes death rate, there are about 12 dead with diabetes in every 100 thousand people. This can be seen that the rate is a small number however this amount is only the dead with diabetes. In fact, diabetes is an important cause of other diseases such as stroke and heart diseases which are the top three of chronic non communicable diseases and have high death rates. It also leads to the destruction of cells in the body such as nerves, blood vessels, heart, eyes and kidneys.

Health-care information systems tend to capture data in databases for research and analysis in order to assist in making medical decisions. As a result, medical information systems in hospitals and medical institutions become larger and larger and the process of extracting useful information becomes more difficult. Traditional manual data analysis has become inefficient and methods for efficient computer based analysis are very needed. To this aim, many approaches to computerized data analysis have been considered and examined. Data mining represents a significant advance in the type of analytical tools. It has been proven that the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to save human resources.

II LITERATURE SURVEY

There are many destructive diseases in the world which cause rapid death by taking time to affect such as cancer and diabetes. They take a lot of time to spread, thus they are curable or somewhat scalable to a great extent if they are diagnosed soon after introduced into the human body. Research shows that almost all type of cancer can be cured if they are detected in the early stage. It is also true for diabetes as they can be controlled if they are detected at the right time. So, a prediction technique that takes help from the computer and processes data from affected user to detect possible contamination can be a great tool for assisting both the doctors and patients with these diseases. A challenge in the process is that the detection accuracy has to be acceptable in order to make the system a reliable one. In this study, we have analyzed medical data using several classification algorithms in order to optimize classifier performance for cancer and diabetes prediction. [1]

Diabetes Mellitus (DM) is the group of diseases where the patient suffers from higher levels of sugar in blood over a prolonged time. Machine learning classifier helps to predict the disease based on the condition of the symptom suffered by the patient. The aim of this paper is to compare the performance of the machine learning tree classifiers in predicting Diabetes Mellitus (DM). Machine learning tree classifiers such as Random Forest, C4.5, Random Tree, REPTree, and Logistic Model Tree (LMT) were analyzed based on their accuracy and True Positive Rate (TPR). In this analysis of predicting diabetes mellitus Logistic Model Tree (LMT) machine learning tree classifier achieved higher accuracy of 79.31%, True Positive Rate (TPR) 0.739 and an execution time of 1.09 sec than other classifiers under study. [2]

In this paper, we revisit the data of the San Antonio Heart Study, and employ machine learning to predict the future development of type-2 diabetes. To build the prediction model, we use the support vector machines and ten features that are well- known in the literature as strong predictors of future diabetes. Due to the unbalanced nature of the dataset in terms of the class labels, we use 10-fold cross-validation to train the model and a hold-out set to validate it. The results of this study show a validation accuracy of 84.1% with a recall rate of 81.1% averaged over 100 iterations. The outcomes of this study can help in identifying the population that is at high risk of developing type-2 diabetes in the future. [3]

Early diseases prediction plays an important role for improving healthcare quality and can help individuals avoid dangerous health situations before it is too late. This paper proposes a disease prediction model (DPM) to provide an early prediction for type 2 diabetes and hypertension based on individual's risk factors data. The proposed DPM consists of isolation forest (iForest) based outlier detection method to remove outlier data, synthetic minority oversampling technique tomek link (SMOTETomek) to balance data distribution, and ensemble approach to predict the diseases. Four datasets were utilized to build the model and extract the most significant risks factors. The results showed that the proposed DPM achieved highest accuracy when compared to other models and previous studies. We also developed a mobile application to provide the practical application of the proposed DPM. The developed mobile application gathers risk factor data and send it to a remote server, so that an individual's current condition can be diagnosed with the proposed DPM. The prediction result is then sent back to the mobile application; thus, immediate and appropriate action can be taken to reduce and prevent individual's risks once unexpected healthsituations occur (i.e., type 2 diabetes and/or hypertension) at early stages. [4]

Non-invasive diabetes prediction has been gaining prominence over the last decade. Among many human serums evaluated, human breath emerges as a promising option with acetone levels in breath exhibiting a good correlation to blood glucose levels. Such correlation establishes acetone as an acceptable biomarker for diabetes. The most common data analysis strategies to analyze the biomarkers in breath for disease detection use feature extraction and classification algorithms. However, snags such as computational cost and lack of optimal feature selection on application to realtime signals reduce the efficiency of such analysis. This paper explores the use of a one-dimensional (1-D) modified convolution neural network (CNN) algorithm that combines feature extraction and classification techniques. The approach proposed in this paper is found to significantly reduce the limitations associated with using these techniques individually and thereby improving the classifier's performance further. This paper proposes to apply a modified 1-D CNN on real-time breath signals obtained from an array of gas sensors. The experimentation and the performance of the system is carried out and evaluated. [5]

III PROBLEM STATEMENT

To design an algorithm for classifiers Decision tree, Artificial Neural Network, Naive bayes and Association Rule for prediction of diabetes using available Dataset collected from PIMA Indian UCI library.To produce the results from weka tool using same dataset and compare accuracy of both results.

3.1 Scope

The scope of the project is presents four stages of the process of conceptual framework in the study. The process starts with data manipulation. Next, four models will be investigated for finding a prediction model. Then, accuracy of each model will be calculated and compared for seeking the best model. Lastly it provides the result to user whether he will be prone to have diabetes in future. The study ends up with creating a web application.

IV SYSTEM DESIGN

The proposed system described in fig is comprised of four stages in overall framework in the study. The very first step of the framework is data manipulation. Next, there are four models will be examined for determining a prediction model. Then, accuracy of each and every model will be calculated and compared with each other for getting the best model from them so that accurate result will be predicted to user. Lastly it provides the result to user whether he will be prone to have diabetes in future. The study ends up with creating a mobile application. Diabetes Mellitus is a one of the most common and growing disease which occurs due to high glucose level in blood of human being. It is growing in many countries all over the world and it is necessary to prevent this disease at early stage by identifying the symptoms of diabetes using several method and by taking precautions to not to happen. However, the application in disease prediction and medical data analysis still has room for improvement. For example, every hospital possesses a plethora of patient's basic andmedical information, and it is essential to revise, supplement, and extract meaningful knowledge from these data to support clinical analysis and diagnosis. It is reasonable to believe that there are various valuable patterns and waiting for researchers to explore them.

4.1 Association Rule

Association rules gives a strong relationship between attribute- value pairs (or items) that occur frequently in a given data set. Association rules are commonly used to analyse the patterns of two or more frequent things that

will happen together and two things which are frequently used together. E.g. purchase pattern of customer in store. Such analysis is useful in many decision-making processes, such as product placement, catalogue design, and cross- marketing. Discovery of association rules are based on associative classification where association rules are generated and analyzed for classification and prediction of diabetes. Association rule mining is two-step process. First step is that it searches for attribute-value pair that occurs repeatedly in data-set. Each pair in dataset is called as item. Group of these items is called as frequent item-set. Next step is to analyse frequent item sets to generate association rules.

V SYSTEM ARCHITECTURE

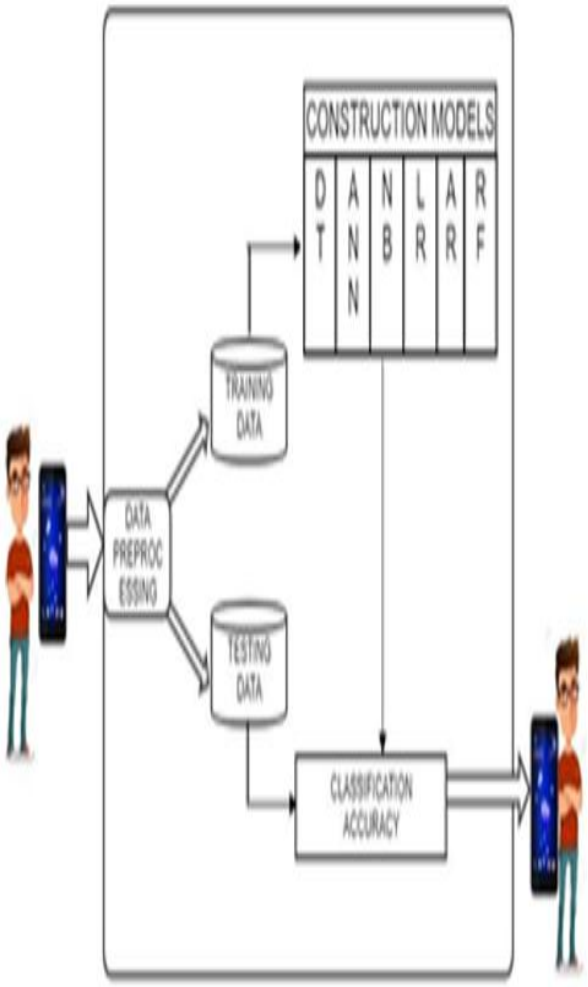


Figure 5.1: Architecture Diagram

VI ALGORITHM DETAILS

1. Generate decision tree
2. Convolution Neural Network Algorithm
3. Logistic Regression Algorithm
4 Random Forest Algorithm

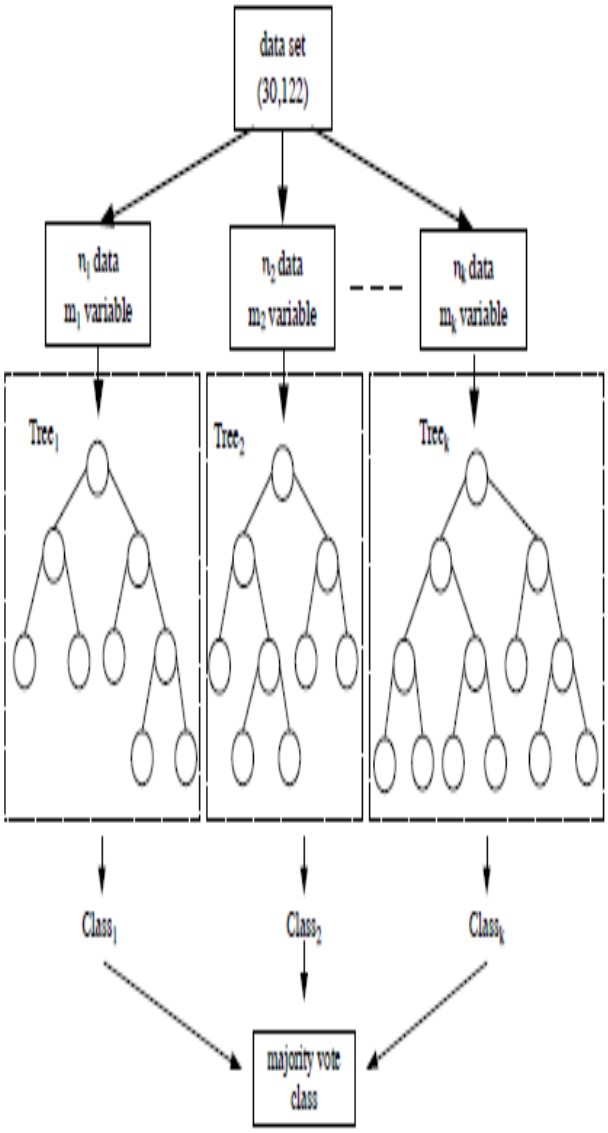


Figure 6.1: Random Forest Model

6.2 Test Cases & Test Results

Table 6.1: Test Cases

No.	Test Case	User Input	Expected Result	Actual Result	Status
1.	Register User	User gives all the credentials asked by the system	Registration successful	Registration successful	Pass
2.	Register User	If User miss any information to enter	Registration Failed	Registration Unsuccessful please Try again	Pass
3.	Login	System takes the username and password	Login successful	Successful	Pass
4.	Login	If incorrect information is entered	Login failed	Login failed, please give correct password or username	Pass

VII RESULTS

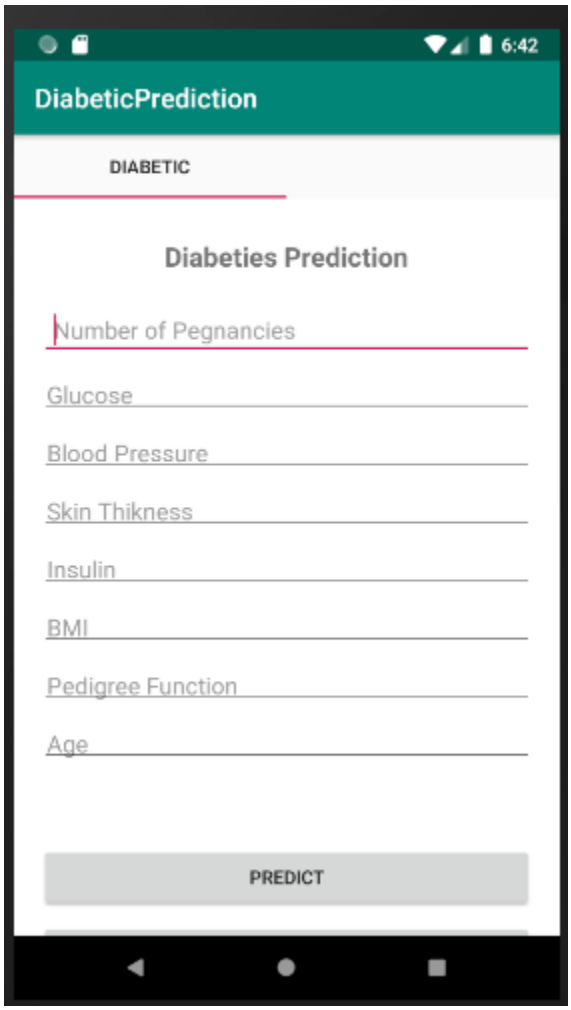


Figure 7.1: Diabetes Prediction Information

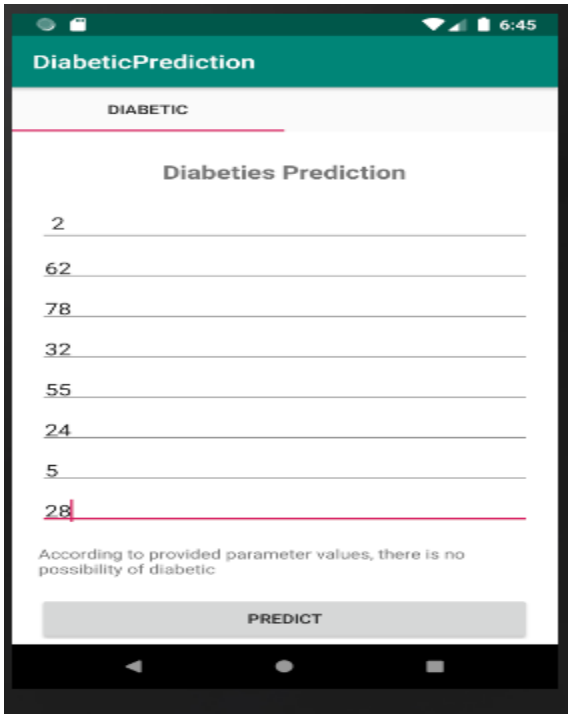
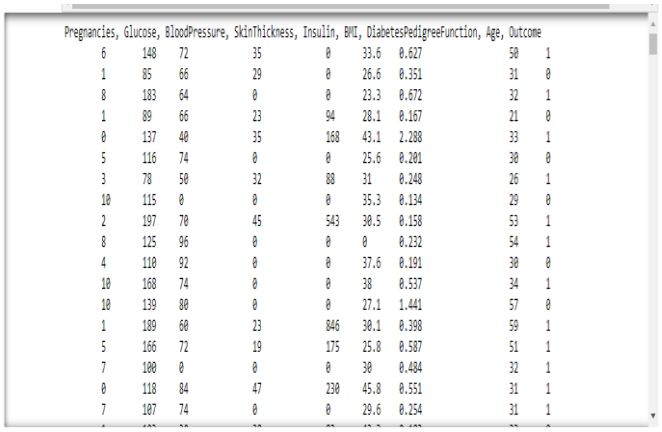


Figure 7.2: Diabetes Prediction



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1

Figure 7.3: Pima Dataset

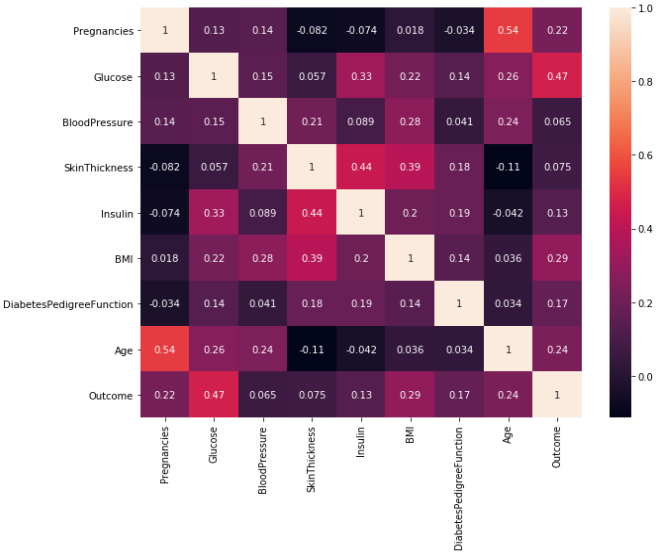


Figure 7.4: Corelation values

VIII CONCLUSIONS

The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in database and generate clear and understandable description of patterns. This study aimed at the discovery of a model for the diagnosis of diabetes. The dataset used was the Pima Indian diabetes dataset. Pre-processing was used to improve the quality of data.In this work, a web application is proposed by using a use of disease classifiers and a real data set. Before creating the web application, three classification models were evaluated for seeking a predicting model. This models consists Decision Tree, Neural Network, Naive Bayes and association rules.To investigate the robustness of each model, accuracy and ROC Curve were calculated and compared with others. Therefore this algorithm was selected to model the diabetes risk prediction and used for creating the application.

REFERENCES

[1] Mustakim Al Helal, Atiqul Islam Chowdhury, Ashraful Islam, Eshtiak Ahmed,Md. Swakshar Mahmud, Sabrina Hossain, "An Optimization Approach to Improve Classification Performance in Cancer and Diabetes

Prediction", International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[2] chine Learning Tree Classifiers in Predicting Diabetes Mellitus", International Conference on Advanced Computing & Communication Systems (ICACCS), 2019.

[3] Lejla Alic, Hasan T. Abbas, Marelyn Rios, Muhammad AbdulGhani, and Khalid Qaraqe, "Predicting Diabetes in Healthy Population through Machine Learning", IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 2019.

[4] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian, Jongtae Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension", Data-Enabled Intelligence for Digital Health, 2019.

[5] S. Lekha, Suchetha M, "Real-Time Non-Invasive Detection and Classification of Diabetes Using Modified Convolution Neural Network", IEEE Journal ofBiomedical and Health Informatics, 2018.

[6] Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, Aziz Guergachi, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques", IEEE Access, 2018.

[7] Longfei Han, Senlin Luo, Jianmin Yu, Limin Pan, Songjing Chen, "Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes", IEEE Journal of Biomedical and Health Informatics, 2014.

[8] Shaker El-Sappagh, Farman Ali, Samir El-Masri, Kyehyun Kim, Amjad Ali, Kyung-Sup Kwak, "Mobile Health Technologies for Diabetes Mellitus: Current State and Future Challenges", IEEE Access, 2018.

[9] Levente Kovcs, Gyrgy Eigner, Mt Siket, Lszl Barkai, "Control of Diabetes Mellitus by Advanced Robust Control Solution", IEEE Access, 2019.

[10] Stefano Bromuri, Serban Puricel, Rene Schumann, Johannes Krampf, Juan Ruiz and Michael Schumacher, "An expert Personal Health System to monitor patients affected by Gestational Diabetes Mellitus: A feasibility study", Journal of Ambient Intelligence and Smart Environments 8(2016) 219237.

[11] Gyorgy J. Simon, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro and Peter W. Li, "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus", IEEE TRANSACTIONS ONKNOWLEDGE AND DATA ENGINEERING.

[12] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", Informatics inMedicine Unlocked 10 (2018) 100107.

[13] Xue-Hui Meng a, Yi-Xiang Huang a, Dong-Ping Rao b, Qiu Zhang a, QingLiu b, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", Kaohsiung Journal of Medical Sciences (2013)29, 93e99.

[14] Kung-Jeng Wang, Angelia Melani Adrian a, Kun-Huang Chen a, Kung-Min Wang b, "An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus", Journal of Biomedical Informatics 54 (2015) 220229.

[15] Jinn-Yi Yeh, Tai-Hsi Wu b, Chuan-Wei Tsao, "Using data mining techniques to predict hospitalization of hemodialysis patients", Decision Support Systems 50 (2011) 439448.

[16] Nongyao Naiarun, Rungruttikarn Moungmai , "Study and implementation of classifiers for the risk of diabetes prediction", 7th International Conference on Advances in Information Technology.