Savitribai Phule Pune University
Modern Education Society's College of Engineering, Pune
19, Bund Garden, V.K. Joag Path, Pune – 411001.

ACCREDITED BY NAAC WITH "A" GRADE (CGPA – 3.13)

DEPARTMENT OF COMPUTER ENGINEERING

A SEMINAR REPORT

ON

"Diagnosis of Diabetes using an Machine Learning Algorithm based Model"

T.E. (COMPUTER)

SUBMITTED BY

Mr. GHANSHYAM PATIL

Exam seat No:71818487E

UNDER THE GUIDANCE OF

Dr. J.R.PANSARE

(Academic Year: 2019-2020)

Savitribai Phule Pune University
Modern Education Society's College of Engineering, Pune
19, Bund Garden, V.K. Joag Path, Pune – 411001.

ACCREDITED BY NAAC WITH "A" GRADE (CGPA – 3.13)

DEPARTMENT OF COMPUTER ENGINEERING

Certificate

This is to certify that seminar entitled

"Diagnosis of Diabetes using an Machine Learning Algorithm based Model"

has been completed by Mr. **GHANSHYAM PATIL** (Roll No. 06) of TE COMP I in the Semester - II of academic year 2019-2020 in partial fulfillment of the Third Year of Bachelor degree in "Computer Engineering" as prescribed by the Savitribai Phule Pune University.

Dr. J.R.Pansare
Seminar Guide

(Dr.(Mrs.) N. F. Shaikh)
H.O.D

Place: MESCOE, Pune.
Date: / /2020

ACKNOWLEDGEMENT

It gives me great pleasure and satisfaction in presenting this seminar on "Diagnosis of Diabetes using an Machine Learning Algorithm based Model".

I would like to express my deep sense of gratitude towards my seminar guide **Dr. J.R. Pansare** *for her support, continuous guidance and the tremendous contribution of being so understanding and helpful throughout the seminar.I would also like to express my gratitude to our honourable HOD ma'am* **Dr.(Mrs.) N. F. Shaikh**.

I would like to thank all those, who have directly or indirectly helped me for the completion of the work during this mini project.

GHANSHYAM PATIL
T.E. Computer
Roll no. : 06

Contents

Abstract

Data in electronic medical records (EMRs) have been widely employed owing to rapid advances in disease assessment technologies.

Diabetes is a metabolic disease affecting a multitude of people worldwide. Its incidence rates are increasing alarmingly every year. If untreated, diabetes-related complications in many vital organs of the body may turn fatal. With the emerging increase of diabetes, that recently affects around 346 million people, of which more than one-third go undetected in early stage, a strong need for supporting the medical decision-making process is generated. A number of researches have focused either in using one of the algorithms or in the comparisons of the performances of algorithms on a given, usually predefined and static datasets that are accessible through the Internet. This paper focuses on the implementation of the support vector machine (SVM),Decision Tree and Naïve Bayes statistical modeling, in order to improve the computer-supported diagnosis reliability.Results on a real-life diabetes dataset show that SVMs provide a promising tool for the prediction of diabetes. Furthermore, the extracted rules agree with the outcome of relevant medical studies.

Keywords-*Decision Tree, Support Vector Machine(SVM), Naive Bayes.Machine Learning*

List of Figures

List of Tables

LIST OF ABBREVIATIONS

Abbreviation	ILLUSTRATION
SVM	SUPPORT VECTOR MACHINE
KNN	K Nearest Neighbour
RF	Random Forest
ANN	ARTIFICIAL NEURAL NETWORK
MLP	Multilayer Perceptron

Chapter 1

INTRODUCTION

1.1 About Diabetes

Diabetes is a chronic disease caused by the increase in blood sugar, mainly either due to the less production or no production of insulin in body (type 1diabetes), or due to the fact that cells do not respond to the produced insulin (type 2 diabetes). In recent years, the number of diabetic patients has increased drastically, as noted in [1], mainly due to the aging population and irregular western food habits. According to the World Health Organization, diabetes affects around 346 million people in the world, with the prevalence of diabetes type 2. Moreover, diabetes is the major cause for heart stroke, kidney failure, lower-limb amputations and blindness. As presented in [2], the absence of symptoms, or the absence of recognition of the indicators in the patient's data, may lead to the pre-diabetes or diabetes condition that goes undetected even in more than one-third of people that are later diagnosed with diabetes.

Chronic hyperglycaemia in people with diabetes increases the risk of microvascular damage, which leads to retinopathy, nephropathy, and neuropathy. Therefore, diabetes is the leading cause of blindness and visual impairment in adults in developed countries [3] and is responsible for over one million lower limb amputations each year. Diabetic people are also exposed to an elevated risk of macrovascular complications, where they are two to four times more likely to get cardiovascular disease (CVD) than people without diabetes. Due to these these complications, diabetes is found to be the fourth leading cause of global death by disease.

The prevalence of type 2 diabetes is increasing at a fast pace due to obesity, in particular, central obesity, physical inactivity, and unhealthy dietary habits [4]. Early detection of diabetes would be of great value given the fact that at least 50% and 80% in some countries, of all people with diabetes are unaware of their condition and will remain unaware until complications appear [3], [5].

Recent studies have shown that 80% of type 2 diabetes complications can be prevented or delayed by early identification and intervention in people at risk [3], [5], for example, by changing their lifestyle [4] and/or by therapeutic methods. Intelligent data analysis, such as data mining and machine learning techniques are, therefore, valuable for identifying those people

Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms [6],[10],[8] works better in diagnosing different diseases. Data Mining [11], [7] and Machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study [9]. .

In this work, Naive Bayes, SVM, and Decision Tree machine learning classification algorithms are used and evaluated on the dataset to find the prediction of diabetes in a patient. Experimental performance of all the three algorithms are compared on various measures and achieved good accuracy . The developed diagnostic tool enables the computer-based prediction on diabetes, based on the previously acquired values. The statistical analysis shows the high accuracy of data classification. Also, the proposed implementation of algorithms aims to improve the reliability of the decision by using the power of algorithms in minimizing their individual weakness.

Chapter 2

Literature Survey

2.1 Related Work on Diagnosis of Diabetes

Different classification and clustering algorithms are used for prediction and diagnosis of diabetes. Support vector machine was used to diagnose diabetes on Pima Indian diabetes dataset [10]. [7] suggest that using Adaptive Neuroo Fuzzy Inference System shows a better accuracy for the diagnosis of diabetes and prediction of cancer. [7] also gives the information about the accuracy of Naive Bayes classifier and K –means algorithm. The accuracy obtained by these methods is around 80

In [11] Expectation–Maximization, Loyyd's H means+ algorithm and genetic algorithm were suggessted. EM algorithm becomes inaccurate when the dimension of dataset d is high [11]. Decision tree is also used for prediction of diabetes which shows an accuracy of 78.17% [12]. ANN andd back propagation models are used for pattern identificaation and binary classification in [12].

A recent study in this direction is given in [2], where two machine learning techniques, namely SVM and ANN (Artificial Neural Networks), are used to predict pre-diabetes in Korean population. A similar approach is used in studying the correlation for hematological parameters and glucose level for identification of diabetes . In general, the research focus is to conduct some of the supervised learning algorithms on the given dataset and extract the knowledge about the prediction of diabetes based on given values of the appropriate attributes. In [7], the SVM implementation gives the prediction accuracy of 94%. Another implementation of the SVM in detecting the diabetes is given in [8]. Here, the SVM classifier, however, performs only 78 of accuracy. A method for prediction of diabetes by using Bayesian network is given in [9] while the authors in [10] separately use Naïve Bayes and k-nearest neighbor algorithm.

Most of the mentioned researches, rely on Pima Indian database of diabetic, and therefore have the same attributes and similar conclusions. Furthermore, they treat one or two algorithms independently to compare the efficiency of the algorithms between each other. Some studies, however, recommend the hybrid use of a distance-based algorithm and a statistical based method [11] or the combination of classification and clustering [1].

2.2 Methodology used

Proposed procedure is summarized in figure-1 below in the form of model diagram. The figure shows the flow of the research conducted in constructing the model.

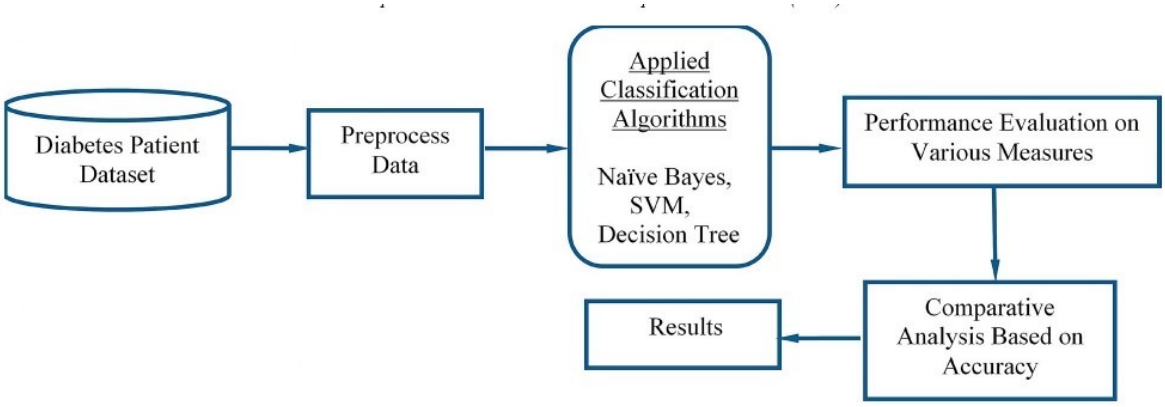

Figure 2.1: Proposed Model Diagram

Chapter 3

Analytical Work

3.1 Support Vector Machine(SVM)

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes[26]. For better generalization hyperplane should not lies closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors [27].

The Accuracy of the experiment is evaluated using WEKA interface. The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyperplane which is defined by w^T x + b = 1 and the hyperplane defined by w^T x + b = 1 This distance is equal to 2/w. This means we want to solve max 2/ w. Equivalently we want min w/ 2. The SVM should also correctly classify all x(i), which means $y^i(w^T x^i + b) > = 1,1, N$. The evaluated performance of SVM algorithm for prediction of Diabetes [16], [30] using Confusion Matrix is as follows:

	A	B
A-Tested Negative	500	0
B-Tested Positive	268	0

Table 3.1: Confusion matrix of SVM

3.2 Naive Bayes Classifier

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with imbalancing problems and missing values. Naive Bayes [24] is a machine

learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability P(C—X) can be calculated from P(C),P(X) and P(X—C) [23]. Therefore, P(C—X) = (P(X—C) P(C))/P(X) Where, P(C—X) = target class's posterior probability . P(X—C) = predictor class's probability. P(C) = class C's probability being true. P(X) = predictor's prior probability. The evaluated performance of Naive Bayes algorithm using Confusion Matrix is as follows:

	A	B
A-Tested Negative	422	78
B-Tested Positive	104	164

Table 3.2: Confusion matrix of Naive Bayes

3.3 Decision Tree

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes [11]. The evaluated performance of Decision Tree technique using Confusion Matrix is as follows:

	A	B
A-Tested Negative	407	93
B-Tested Positive	108	160

Table 3.3: Confusion matrix of Decision Tree

Chapter 4

Evaluation

4.1 Dataset

In this work WEKA tool [12], [13] is used for performing the experiment. WEKA is a software which is designed in the country New Zealand by University of Waikato, which includes a collection of various machine learning methods for data classification, clustering, regression, visualization etc. One of the biggest advantages of using WEKA is that it can be personalized according to the requirements. The main aim of this study is the prediction of the patient affected by diabetes using the WEKA tool by using the medical database PIDD. Table-4.1 shows a brief description of the dataset.

Database	No. of Attributes	No. of Instances
PIDD	8	768

Table 4.1: Confusion matrix of Decision Tree

PIDD-Pima Indians Diabetes Dataset

The proposed methodology is evaluated on Diabetes Dataset namely (PIDD) [14], which is taken from UCI Repository. This dataset comprises of medical detail of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class '0' treated as tested negative for diabetes and value of another class '1' is treated as tested positive for diabetes. Dataset description is defined by Table-4.1 .

4.2 Accuracy Measures

Naive Bayes, SVM and Decision Tree algorithms are used in this research work. Experiments are performed using internal cross-validation 10-folds. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve)measures are used for the classification of this work. Table-4.2 defines accuracy measures below:

Measures	Definitions	Formula
1. Accuracy (A)	Accuracy determines the accuracy of the algorithm in predicting instances.	A=(TP+TN) / (Total no of samples)
2. Precision (P)	Classifiers correctness/accuracy is measured by Precision.	P = TP / (TP+ FP)
3. Recall (R)	To measure the classifiers completeness or sensitivity, Recall is used.	R =TP / (TP+FN)
4. F-Measure	F-Measure is the weighted average of precision and recall.	F=2*(P*R) / (P+R)
5. ROC	ROC(Receiver Operating Curve) curves are used to compare the usefulness of tests.	–

Table 4.2: Accuracy Measures

Classification Algorithms	Precision	Recall	F-Measure	Accuracy %	ROC
Naive Bayes	0.759	0.763	0.760	76.30	0.819
SVM	0.424	0.651	0.513	65.10	0.500
Decision Tree	0.735	0.738	0.736	73.82	0.751

Table 4.3: Comparative Performance of Classification Algorithms on Various Measures

4.3 Few Techniques which can be used for detection

4.3.1 K-Nearest Neighbors

In this classification technique, the anonymous data points are discovered using the familiar data points which are known as nearest neighbors. k-Nearest neighbors (k NN) is conceptually simple and is also called as lazy learning, where “k” is the nearest neighbor. In kNN algorithm, the aim is to vigorously recognize k samples in the training dataset which are identical to a new sample.

Advantages:

• It is easy to implement.

• Training is done in a faster manner

Disadvantages:

• Time becomes prohibitive for finding the nearest neighbor in the training data which is of huge size, thus making it slow.

• It requires large storage space.

• The transparency of knowledge representation is very poor.

4.3.2 Random Forest

Random forest is a classifier that constitutes a few decision trees and considered as one of the dimensionality reduction methods. It is one of the ensemble methods for classification, regression, and other terms. It can be used to rank the importance of variables.
Advantages:
• Random forest improves classification accuracy.
• It works well with the dataset of large number of input variables.
Disadvantages:
• Random Forest is fast to train but once trained, it becomes slow to create predictions.
• It is slow to evaluate.
• Interpretation is very hard.

4.3.3 Artificial Neural Network(ANN)

The artificial neural network is aroused by the neural network of human being, and it is a combination of three layers, i.e., input layer, hidden layer, and output layer, which is also called as MLP (Multilayer Perceptron). The hidden layer is similar to neuron, and each hidden layer consists of probabilistic behavior.
Advantages:
• Ability to learn and model nonlinear and complex relationships.
• Ability to generalize the model and predict the unseen data.
• Resistant to partial damage.
Disadvantages:
• Optimizing the network can be challenging because of the number of parameters to be set in.
• For large neural networks, it requires high processing time.

Chapter 5

Discussions and Results

Table-4.3 represents different performance values of all classification algorithms calculated on various measures. From Table-4.3 it is analyzed that Naive Bayes showing the maximum accuracy. So the Naive Bayes machine learning classifier can predict the chances of diabetes with more accuracy as compared to other classifiers. Performances of all classifier's based on various measures are plotted via a graph in Figure-5.1 Figure-5.2 represents ROC area of all classification algorithms.

Total instances	Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
	Naive Bayes	586	182
768	SVM	500	268
	Decision Tree	567	201

Table 5.1: Confusion matrix of Decision Tree

Table-5.1 determines classifiers performance on the basis of classified instances. According to these classified instances, accuracy is calculated and analyzed. Performance of individual algorithm is evaluated on the basis of Correctly Classified Instances and Incorrectly Classified Instances out of a total number of instances. Figure-5.3 shows the graphical performance of all classification algorithms on the basis of classified instances. From Table-4.3 and Table-5.1 we can conclude that Naive Bayes classification algorithm outperforms comparatively other algorithms. So, Naive Bayes algorithm is considered as the best supervised machine learning method of this experiment because it gives higher accuracy in respective to other classification algorithms with an accuracy of 76.30 %.

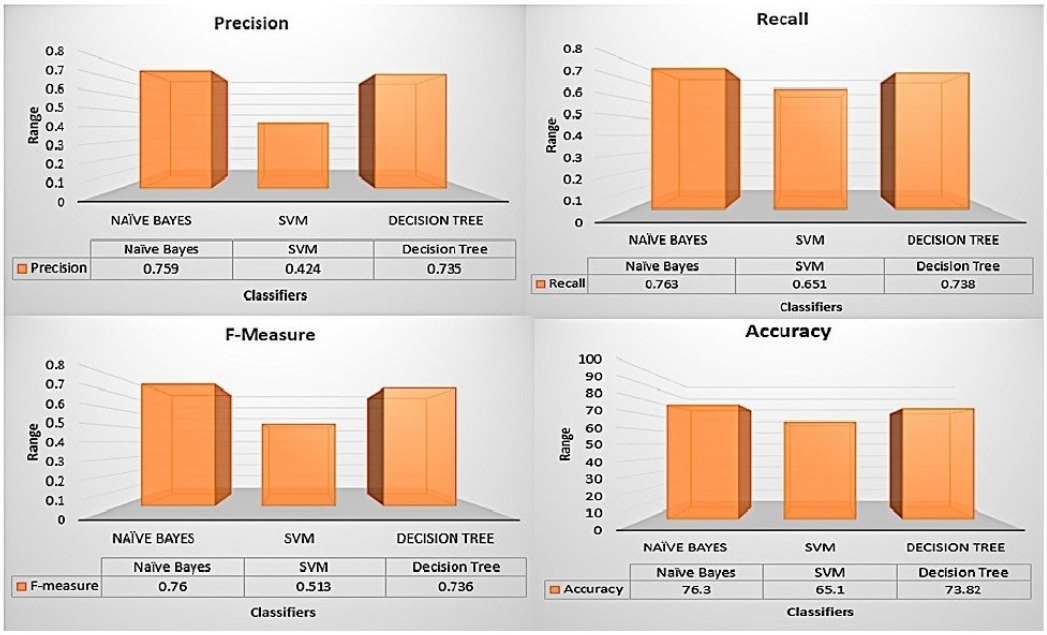

Figure 5.1: Classifiers Performance on Various Measures

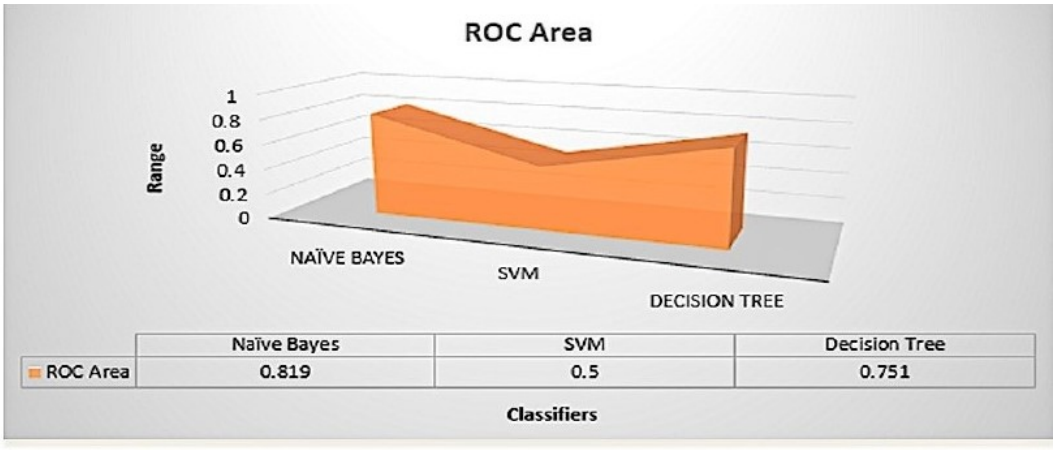

Figure 5.2: ROC Area of all Classification Algorithms

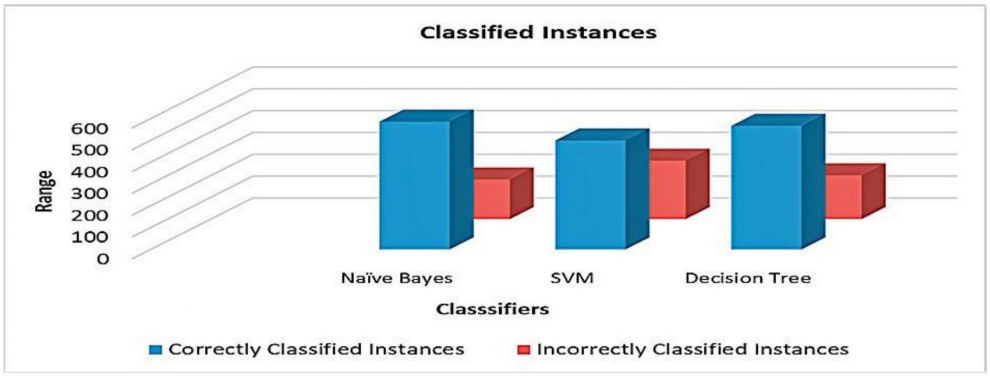

Figure 5.3: Classified Instances

CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms

REFERENCES

[1] S. Peter, "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus," Bonfring International Journal on Data Mining, vol. 4, no.2, pp. 7-11, 2014.

[2] J. C. Yong, C. K, Hyeon, M. K. Hee, W. P. Seok, J. Jongoh, and J. K. Dae, "Prevalence and Management of Diabetes in Korean Adults: Korea National Health and Nutrition Examination Surveys," Dibetes Care, vol. 32, no. 11, pp. 2016-2020, 2009

[3] International Diabetes Federation, Diabetes Atlas, 3rd ed. Brussels, Belgium: International Diabetes Federation, 2007.

[4] M. Uusitupa, "Lifestylematter in prevention of type 2 diabetes," Diabetes Care, vol. 25, no. 9, pp. 1650–1651, 2002.

[5] M. Franciosi, G. D. Berardis, M. C. E. Rossi, and M. Sacco, "Use of the diabetes risk score for opportunistic screening and impaired glucose tolerance," Diabetes Care, vol. 28, no. 5, pp. 1187–1193, 2005.

[6] Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[7] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[8] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10

[9] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.

[10] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116. doi:10.1016/j.csbj.2016.12.005.

[11]Kumar,P.S.,Umatejaswi,V.,2017.DiagnosingDiabetesusingDataMiningTechniques.InternationalJournal

[12] Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-

[13] Garner, S.R., 1995. Weka: TheWaikato Environment for Knowledge Analysis, in: Proceedings of the New Zealand computer science research students conference, Citeseer. pp. 57–64

[14]Kayaer,K.,Tulay,2003.MedicaldiagnosisonPimaIndiandiabetesusinggeneralregressionneuralnetworks,in tionalconferenceonartificialneuralnetworksandneuralinformationprocessing(ICANN/ICONIP),pp.181–184.