Machine Learning Based Unified Framework for Diabetes Prediction

S M Hasan Mahmud
Department of Software Eng.
Daffodil International University
Dhaka-1207, Bangladesh
UESTC, Chengdu, China
hasan.swe@daffodilvarsity.edu.bd

Md Altab Hossin
Dept. of Management Science & Engineering
University of Electronic Science and Technology of China
Chengdu- 611731, China
altabbd@163.com

Md. Razu Ahmed
Department of Software Engineering, Daffodil International University
Dhaka-1207, Bangladesh
razu35-1072@diu.edu.bd

Sheak Rashed Haider Noori
Department of Computer Science and Eng.
Daffodil International University, Dhaka-1207, Bangladesh
drnoori@daffodilvarsity.edu.bd

Md Nazirul Islam Sarkar
School of Public Administration
Sichuan University, Chengdu- 610065, China
sarker.scu@yahoo.com

ABSTRACT
Machine learning gained a significant position in healthcare services (HCS) due to its ability to improve the disease prediction in HCS. Machine learning techniques and artificial intelligence have already been worked in the HCS area. Recently, diabetes is a notable public chronic disease worldwide. It is growing rapidly because of bad lifestyles, taking more junk food and also lake of health awareness. Therefore, there is a need of framework that can effectively track and monitor people's diabetes and health condition within an application view. In this study, we proposed a framework for real time diabetes prediction, monitoring and application (DPMA). Our objective is to develop an optimized and efficient machine learning (ML) application which can effectually recognize and predict the condition of the diabetes. In this work, five most important machine learning classification techniques were considered for predicting diabetes. However, we use different evaluation criteria to investigate the performance of these classification techniques. In addition, performance measurement of the classification techniques was evaluated by applying the 10-fold cross validation method. The analysis results show that Naïve Bayes achieved highest performance than the other classifiers, obtaining the F1 measure of 0.74.

CCS Concepts
• **Computing methodologies→Machine learning algorithms**

Keywords
Machine Learning; Classification; Supervised Learning; Diabetes Prediction; Disease Prediction.

1. INTRODUCTION
Diabetes Mellitus (DM) is defined as a group of metabolic diseases in which humans have towering blood sugar levels.

Diabetes is a prolonged disease that happens when the body cannot efficiently use the insulin it generates. As a result, the disease increase the risk of malfunction of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels [1]. According to the report of World Health Organization (WHO) diabetes will be the seventh prominent cause of death by 2030 [2]. About 642 million adults (1 in 10 adults) are projected to have diabetes in 2040 [3]. The deaths of around 1.6 million people were completely affected by diabetes in 2015 and 2.2 million deaths due to high blood glucose in 2012 [4]. Diabetes Mellitus do not depend on the age, it can happen with people anytime. There are three types of diabetes [4]: i) Juvenile or childhood diabetes (type 1 diabetes), ii) Type 2 or adult diabetes iii) Gestational or type 3 diabetes. Gestational diabetes is hyperglycemia which occurs because of the change in hormones during pregnancy. Generally, type 1 diabetes happens due to the lack of insulin production and it is diagnosed in people of young age [4]. Type 2 is a very familiar form of diabetes, and it contains a huge volume of people from around the world [5]. Type 2 mostly causes surplus body weight and physical disuse. Whatsoever, type 1 and type 2 diabetes cannot be cured properly. But, early diagnosis and simple lifestyle can prevent it. Moreover, there are different new cases of diabetes arises from the developing countries [5] where shocking amounts of diabetes affected people are from Bangladesh which is projected to climb up to more than 16 million by 2020 [6].

In last few decades, data has been elevated in a vast scale in diverse arenas [7] [8] including medical fields. Machine Learning is a discipline that aims to solve different important biomedical problems [9] [24]. The machine learning based classification techniques are the most operative methods for both real-life and scientific problems [10]. The use of these classification based approaches in the diagnosis and cure of diseases can significantly decrease medical errors and human costs. As described in the study [11], machine learning based classification techniques have prospective performance in prediction accuracy as compared to other algorithms for data classification. Data classification accuracy may vary conditionally on different machine learning techniques. Many of the researchers have been focused on diabetes from various perspectives of their works where most of the study discussed the classification techniques for diabetes

BDET 2018, August 25–27, 2018, Chengdu, China.

ACM ISBN 978-1-4503-6582-6/18/08…$15.00
DOI: https://doi.org/10.1145/3297730.3297737

prediction and its accuracy [11] [12]. However, the author's acknowledgement, no one has referred to the real time diabetes prediction with an application using various algorithms and techniques. This paper main aspect is to obtain more efficient results and reduce the cost of diagnosis in the Health Care Services. Therefore, the aim of this study is to evaluate the performance of different machine learning (supervised learning) classification techniques for the classification of diabetic affected or not. We use six classification techniques, Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistics Regression (LR) and Naïve Bayes (NB). We explore the performance result of different techniques where the performance is evaluated by various standards, such as accuracy, precision, true positive rate (TPR), true negative rate (TNR), F-score. Moreover, the most accurate classification technique is donated for diagnosis of such disease with proposed unified framework.

The rest of the paper is organized as follows, section 2 presents the details about data set, proposed methodology and describes the workflow of proposed system. Section 3 describes the classification algorithms. Section 4 illustrates the analysis results and evaluation technique on performance measurement. Finally, the conclusions are made in section 5.

2. MATERIALS AND METHODOLOGY

2.1 Datasets

The Prima Indian Diabetes Dataset has been used in this study, provided by the UCI Machine Learning Repository. The dataset has been originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases [13]. The dataset consists of some medical distinct variables, such as pregnancy record, BMI, insulin level, age, glucose concentration, diastolic blood pressure, triceps skin fold thickness, diabetes pedigree function [14] etc. This dataset has 768 patient's data where all the patients are female and at least 21 years old. The number of true cases are 268 (34.90%) and the number of false cases are 500 (65.10%), respectively, in the dataset. In the following we chose eight distinct parameters for data prepossessing such as,

I) Pregnancies: Pregnancy records

II) Glucose: Plasma glucose concentration (2hrs in OGTT)

III) Blood Pressure (mm Hg)

IV) Skin Thickness: Triceps with skin fold thickness (mm)

V) Insulin: Patients, 2-Hour level of serum insulin record

VI) BMI

VII) Diabetes pedigree function

VIII) Age: Age (years)

2.2 Proposed Framework for Diabetes Prediction

In this section, the proposed unified framework has been presented. The framework is then further introduced by focusing on the machine learning based prediction. In addition, this section provides a closer look into the real time prediction for diabetes patients. The proposed machine learning based unified architecture for diabetes prediction is shown in Fig. 1. Motivated by the significance of future machine learning based disease predictions, this paper considers classification techniques as well as pre-trained model and real time data prediction service to develop an efficient solution for real time prediction, monitoring, and application of machine learning based diabetes prediction. The author's contributions of the paper are summarized as follows:

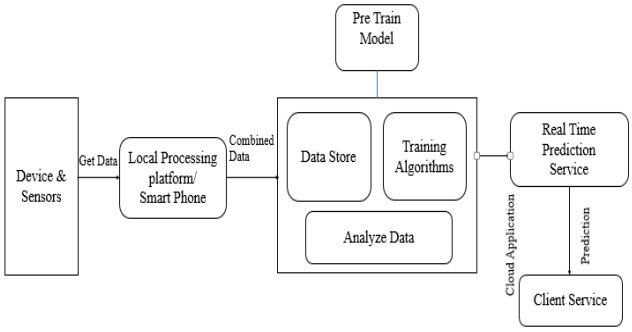

Figure 1. Proposed framework for diabetes prediction, monitoring and application of machine learning.

- A framework for diabetes prediction, monitoring and application (DPMA) of machine learning is proposed. 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.
- Most of the study do not consider the F-score, precision, and recall. However, our study provide average prediction of classification model by considering the F-score, recall and precision.

The proposed DPMA framework for the prediction system is a part of real Health Care Services (HCS). We decided to interpret the diabetes prediction results into an application level view. As shown in Figure 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 processing platform (LPP) or smart phone, which are able to process data from the sensing devices [15]. Once the data is processed by an LPP or a smart phone, it can be sent to the cloud application phase. In the cloud application phase, the processed data from LPP 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.

3. DISCRIPTION OF THE CLASSIFICATION ALGORITHMS

3.1 Artificial Neural Network (ANN)

Artificial neural network (ANN) is an important machine learning technique for biological research. In machine learning, ANN is a convenient computational model which works similar to biological neurons [16]. 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. ANN mainly organized into three layers; i) Input layer (nodes can take input data), ii) Hidden layer or processing stage (processes the input data from input layer) and iii) output layer (results are sent from the hidden layer). In

addition, the result of output layer in each node is called its activation or node value [17].

3.2 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning algorithm which is based on linear classification. SVM work well for many health care problems and can solve both linear and non-linear problems. In order to solve the regression and classification problems efficiently SVM perform better than other classification techniques. Therefore, Vladimir Vapnik and Alexey Chervonenkis [18] [19] introduced the support vector machine classification technique which is attempt to pass a linearly separable hyperplane to classify the dataset into two classes. Finally, the model can undoubtedly estimate the target groups (labels) for new cases.

3.3 Logistics Regression (LR)

Logistic Regression was mostly used in the biological research and applications in the early 20th century [20]. Logistic Regression (LR) is one of the most used machine learning algorithm that is used where the target variable is categorical. Recently, LR is a popular method for binary classification problems. Moreover, it presents a discrete binary product between 0 and 1. Logistic Regression computes the relationship between the feature variables by assessing probabilities (p) using underlying logistic function.

3.4 Decision Tree (DT)

Decision tree (DT) is one of the popular supervised learning-based classification algorithms in Machine Learning. DT can be used for both classification and regression problems. Moreover, DT is a classification technique which breaks a dataset into smaller subsets or a composite decision into a union of several easier decisions, at the meantime, the final solution with associated decision tree is incrementally developed [21].

3.5 Random Forest (RF)

Random Forest (RF) is a well-known supervised classification algorithm which is able to perform both regression and classification. RF has been first proposed by Leo Breiman [22]. In general, 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 preferably at the node where the trees splits.

3.6 Naive Bayes (NB)

Naive Bayes classifier is a simple but most operative algorithm for the classification problems. Naive Bayes are statistical classifiers that works by making a hypothesis of conditional independence with the training datasets [23]. Henceforth, Naive Bayes classifier is the appropriate classification technique that verdicts best solution for a dataset from a pool of different objects.

4. RESULTS AND DISCUSSION

4.1 Measurement of Classification Techniques

In this study, we used 10-fold validation technique to measure the performance of each classification algorithm. Performance of all the classification algorithms are assessed by different statistical measurement aspects such as accuracy, sensitivity, specificity, NPV, PPV etc. These classification measurement factors are calculated by the terms: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Here,

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

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

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

False Negative (FN): 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.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

True positive rate, 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).

$$\text{TPR= Sensitivity = Recall} = \frac{TP}{(TP+FN)}$$

True negative rate or specificity 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.

$$\text{Specificity = TNR} = \frac{TN}{(TN+FP)}$$

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

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

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

$$\text{F1=} \frac{2*(Recall*Precision)}{Recall+Precision}$$

4.2 Analysis of the Results

In this experiment, we conducted different analysis to evaluate the 6 machine learning classification algorithms for diabetes prediction.

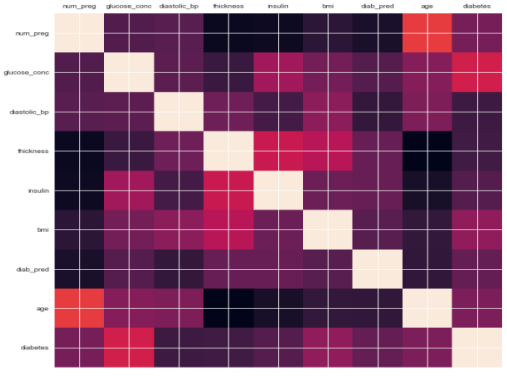

Figure 2. Heat map for checking correlated columns.

From the Prima Indian Data set, 268 true samples and 500 negative samples were taken into analysis. We split the diabetes data set into two parts where the training set contains 70% and the test set contains the remaining 30% of the data, where, training true: 188 (35.01%), training false: 349 (64.99%), test true: 80 (34.63%) and test false: 151 (65.37%). Moreover, the dataset was also checked to verify the correlated features in order to drop the redundant columns. We found that the skin and thickness columns are correlated with 1 to 1. Therefore, we dropped the skin column. The heatmap shown in Figure 2 appear to have no correlated columns.

Figure 3 shows the performance of 6 supervised machine learning techniques for diabetes prediction. Here, NB and SVM outperformed the other classification techniques in terms of accuracy by obtaining the highest accuracy as 74% and 73%, respectively. However, the artificial neural network exhibits lowest performance than the other classification algorithms.

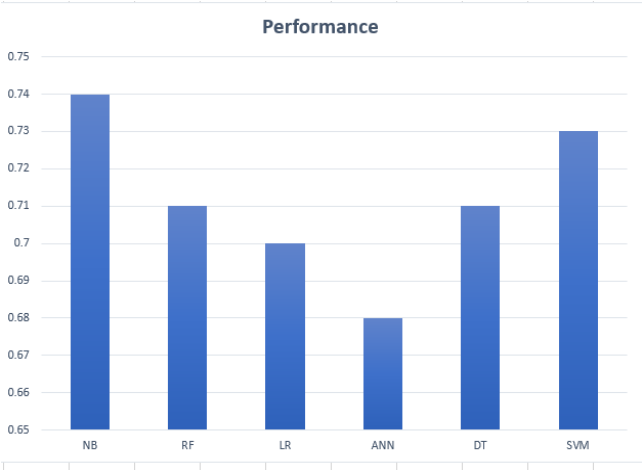

Figure 3. Classification accuracy using ML algorithms.

4.3 Performance Evaluation

The 10-fold cross validation approach is used to evaluate the performance of the prediction model. Predictions of all the machine learning classification algorithms are presented in figure 4, which clearly indicates that Naïve Bayes and SVM exhibits the highest performance and ANN shows the lowest performance than the other 5 classification algorithms in terms of the four measurement factors: specificity, recall, precision and f1 measure.

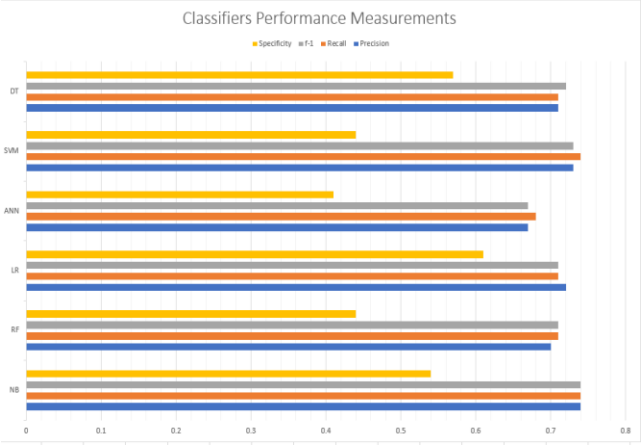

Figure 4. The figure shows the performance of classification techniques on specificity, precision, recall and f1 measure.

Table 1 illustrates the different classification measures. Such as accuracy, sensitivity, specificity, precision, and f1 measure.

Table 1. Classification performance measurements

Measurement Techniques	NB	RF	LR	ANN	SVM	DT
Accuracy	.74	.71	.70	.68	.73	.71
Precision	.74	.70	.70	.67	.73	.71
Sensitivity	.74	.71	.71	.68	.74	.71
F-1	.74	.71	.71	.67	.73	.72
Specificity	.54	.44	.61	.41	.44	.57

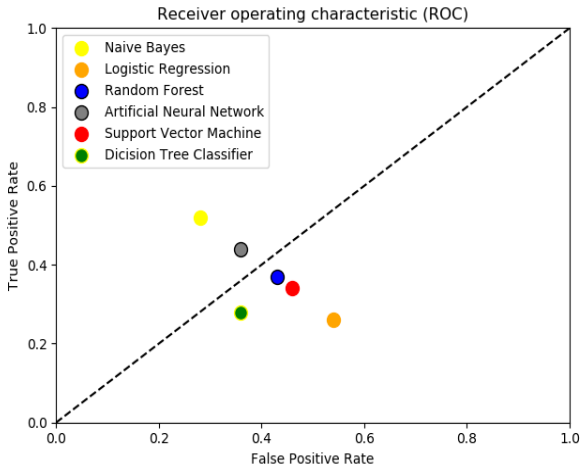

Figure 5. ROC for diabetes prediction of machine learning classification techniques.

All the machine learning classifiers show the accuracy level of nearly 75%, which indicates that the performance of these techniques are pretty well. F-1 measure indicates (NB, SVM, DT, LR and RF) that the five-classification techniques mostly predict accurate results. From the above discussion, it is important to know about the Receiver Operating Characteristics (ROC) curve, which is based on the true positive rate (TPR) and false positive rate (FPR) of these classification results. The ROC curve is presented in Figure 5.

In summary, we highlight the research directions and scope in relation to Health Care Services (HCS) and Bio-medical fields by machine learning classification techniques, which has emerging impact in medical sector. Hence, disease prediction by machine learning classification algorithms should be improved. We describe the most popular machine learning classification techniques and proposed a unified framework for diabetes prediction that require further research in terms of machine learning based disease prediction.

5. CONCLUSION

The main contribution of this study are as follows; first, we compare the performance of the six-machine learning classification techniques and evaluated their performance using the 10-fold validation technique. Secondly, we proposed a framework for diabetes prediction, monitoring and application (DPMA). In general, multiple machine learning classifiers should perform better than a single machine learning classifier. The

experimental results show that the highest classification accuracy is 74% and highest F1 measure is 0.74, respectively. In addition, this application is able to classify the patients based on their diabetes level by collecting real time data from various Health Care Services, such as medical diagnose center, hospital, health tracking devices and sensors etc. We are currently developing a mobile application for predicting and monitoring diabetes for new and old patients. Hence, this application represents a promising tool to aid the stratification of diabetes patients.

6. ACKNOWLEDGMENTS

The authors are grateful who have participated in this research work.

7. REFERENCES

[1] American Diabetes Association. 2010. Diagnosis and classification of diabetes mellitus. Diabetes care. 33 Suppl 1, (Jan. 2010), S62-9. DOI= 10.2337/dc10-S062.

[2] Mathers, C. D., and Loncar, D. 2006. Projections of Global Mortality and Burden of Disease from 2002 to 2030. PLoS Medicine. 3, 11: e442. (Nov. 2006). DOI= https://doi.org/10.1371/journal.pmed.0030442.

[3] How Many People Have Diabetes? https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-many-people-have-diabetes/. Accessed: 2018-06-08.

[4] Diabetes: 2017. http://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed: 2018-06-08.

[5] Samant, P., and Agarwal, R. 2018. Machine learning techniques for medical diagnosis of diabetes using iris images. Computer Methods and Programs in Biomedicine. 157, (Apr. 2018), 121–128. DOI: https://doi.org/10.1016/J.CMPB.2018.01.004.

[6] Shariful Islam, SM., Lechner, A., Ferrari, U., Froeschl, G., Niessen, L.W., Seissler, J., and Alam, D.S. 2013. Social and economic impact of diabetics in Bangladesh: protocol for a case–control study. BMC Public Health. 13, 1 (Dec. 2013), 1217. DOI: https://doi.org/10.1186/1471-2458-13-1217.

[7] Ahmed, M. R., Khatun, M. A., Ali, A., and Sundaraj, K. 2018. A literature review on NoSQL database for big data processing. *International Journal of Engineering & Technology*. 7, 2 (Jun. 2018), 902–906. DOI: https://doi.org/10.14419/ijet.v7i2.12113.

[8] Mahmud, S. M. H., Hossin, M. A., Jahan, H., Noori, S. R. H., and Bhuiyan, T. 2018. CSV-ANNOTATE: Generate Annotated Tables from CSV File. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*. (2018), 71–75. DOI: 10.1109/ICAIBD.2018.8396169.

[9] Tarca, A. L., Carey, V. J., Chen, X., Romero, R., and Drăghici, S. 2007. Machine Learning and Its Applications to Biology. *PLoS Comput.* Biol. 3, 6 (June. 2007), 116. DOI: https://doi.org/10.1371/journal.pcbi.0030116.

[10] Kononenko, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*. 23, 1 (Aug. 2001), 89–109. DOI: https://doi.org/10.1016/S0933-3657(01)00077-X.

[11] Dwivedi, A.K. 2017. Analysis of computational intelligence techniques for diabetes mellitus prediction. *Neural Computing and Applications*. (Apr. 2017), 1–9. DOI: https://doi.org/10.1007/s00521-017-2969-9.

[12] Heydari, M.,Teimouri, M., Heshmati, Z., and Alavinia, S. M. 2016. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. *Int. J. Diabetes Dev. Ctries*. 36, 2 (Jun. 2016), 167–173. DOI: https://doi.org/10.1007/s13410-015-0374-4.

[13] 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].

[14] Smith, J. W., Everhart, J.E., Dickson, W.C., Knowler, W.C., and Johannes, R.S. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proc Annu Symp Comput Appl Med Care*. 261–265.

[15] Mahmud, S. M. H., Kabir, M. A., Salem, O. A. M., and Fernand, K. N. G. 2017. The comparative analysis of online shopping information platform's security based on customer satisfaction. *5th International Conference on Computer Science and Network Technology, ICCSNT*. (2016), 157–161. DOI: https://doi.org/10.1109/ICCSNT.2016.8070139.

[16] Van Gerven, M., and Bohte, S. 2017. Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Front. Comput. Neurosci.* 11 (Dec. 2017), 114.

[17] Hecht-Nielsen. 1989. Theory of the backpropagation neural network. *International 1989 Joint Conference on Neural Networks*. 1,1 (1989), 593-605. DOI: 10.1109/IJCNN.1989.118638.

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

[19] Chervonenkis, A. Y. 2013. Early History of Support Vector Machines. *Empirical Inference*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, 13–20. DOI: https://doi.org/10.1007/978-3-642-41136-6_3.

[20] Berkson, J. 1944. Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*. 39, 227 (Sep. 1944), 357–365. DOI: https://doi.org/10.1080/01621459.1944.10500699.

[21] Safavian, S.R. and Landgrebe, D. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*. 21, 3 (1991), 660–674. DOI: https://doi.org/10.1109/21.97458.

[22] Breiman, L. 2001. Random Forests. Machine Learning. 45, 1 (2001), 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

[23] Leung, K.M. 2007. Naive bayesian classifier. Polytechnic University Department of Computer Science Finance and Risk Engineering. (2007).

[24] Hossain, R., Mahmud, S. M. H., Hossin., M. A., Noori , S. R. H., and Jahan, H. 2018. PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques. Procedia Computer Science. 132, (Jan. 2018), 1068–1076. DOI: https://doi.org/10.1016/J.PROCS.2018.05.022.