2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579869

Diabetes Prediction in Healthcare at Early Stage Using Machine Learning Approach

Md. Mehedi Hassan
Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
mh_ashiq@yahoo.com

Zahrul Jannat Peya
Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
jannat.kuet@gmail.com

Swarnali Mollick
Department of Computer Science and Engineering
Northern University Of Business & Technology Khulna, Bangladesh
swarnalimollick.07@gmail.com

Md. Al-Mamun Billah
Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
almamunnwu@gmail.com

Md. Mehadi Hasan Shakil
Department of Computer Science and Engineering
Northern University Of Business & Technology Khulna, Bangladesh
mhshakil.cse@gmail.com

Asaf Ud Dulla
Department of Electrical and Electronics Engineering
University of Dhaka
Dhaka, Bangladesh
kakon53212@gmail.com

***Abstract*—Diabetes mellitus is a perdurable hyperglycemic disease. Various complications can be caused by this disease. In line with the growing morbidness in the last few years, 642 million people can be infected with diabetes within 2040 which is one among 10 individuals. So undoubtedly this malady needs more attention. Nowadays the usage of machine learning is increasing. So, in many medical perspectives, this technique has been utilized. We have chosen methodologies that give the best performances for independent testing to confirm the universal applicability of the techniques. We have focused on early detecting this disease. We have collected data from Khulna Diabetes Center at Khulna where the instances is 289 and 13 features. In our study, we use the Logistic Regression model with 88%, XGboost 86.36% and Random Forest with 86.36% accuracy. We found that the random forest model performs the best output for diabetics detection.**

Keywords-*Diagnostics, Diabetes, Diabetes Prediction, Data Mining, Predictive Analysis*

I. Introduction

Diabetes is a state of the body, or it is basically known as an incurable and deadliest disease that is caused by an imbalance of the sugar level. It is a hereditary disease or can be infected by environmental factors. Diabetes occurs when the glucose or sugar equality of the body is higher than it actually should be, which is caused when beta cells fail to work appropriately to adequate insulin secretion. A variety of tissues, particularly the kidneys, eyes, blood vessels, nerves, and heart are disordered because of diabetes. Besides, the risk of pancreatic issues, ketoacidosis, hypertension, foot issues, and visual unsettling influences are increased. Diabetes mellitus is now thought to play a significant role in the process of aging. It is classified into two categories. Type-1 diabetes's bearers are mostly aged less than 40 years. Clinical symptoms for type-1 include increased thirst and frequent urination. It is the critical level so it can't be completely incurable by oral medicine resulting in insulin therapy being required. So it is generally called insulin-dependent or juvenile diabetes. Type-2 diabetes's bearers are middle-aged. They are not capable of taking insulin because their beta cells can produce but the insulin can't work efficiently. Arteriosclerosis, obesity, dyslipidemia, hypertension, dyslipidemia, and other diseases are frequently associated with type-2 diabetes. [1].

The pancreas is a diaphragm area's limbs having endocrine and exocrine abilities to collaborate for digestion and to maintain the level of sugar in the circulatory system stable respectively. If the sugar level goes down then the production of insulin has been stopped and alpha cells start working to produce glucagon to maintain the sugar level stable [2]. In the line with the World Health Organization's annual report, 422 million people are living with diabetes. The International Diabetes Federation announced, there were 382 million people with diabetes mellitus in 2013, accounting for 6.6 percent of the overall adult population of the planet and this number is going to be 490 million by 2030 [3]. In total diabetes patients in the world, the SEA Region contains almost 82 million people and this number will rise to 151 million in 2045 [4].

The immunity of diabetic patients is decreased slowly so they can be affected easily by any kinds of diseases and the leading objective for death is cardiovascular disease. If the early prediction of this disease can be done, it will save several human lives.

For this purpose, using machine learning algorithms named Logistic Regression, XGBoost and Random Forest we have built a model by considering some dangerous factors of diabetes to predict it early so that it can help people to survive

against diabetes. The main strength is we have used real time dataset but this is not so much big dataset. By creating predicting models from medical diagnostic datasets gathered from individuals with diabetes, machine learning techniques deliver effective performance for knowledge extraction.

II. Related Work

Aishwarya Mujumdar et al. 2019 [5] used K-means clustering on the attributes - which are highly correlated - Glucose and age, for the diabetic or non-diabetic classification of each patient to get labels of class for the record. To build the models they have applied several algorithms. The accuracy of those algorithm are for Support Vector Classifier - 60%, Random Forest Classifier - 91%, Decision Tree Classifier - 86%, Extra Tree Classifier - 91%, AdaBoost algorithm - 93%, Perceptron - 76%, Linear Discriminant Analysis algorithm - 94%, K-Nearest Neighbour - 90%, Gaussian Naïve Bayes - 93%, Bagging algorithm - 90%, Gradient Boost Classifier - 93% and Logistic Regression - 96%, is the best result. They also applied to pipeline to build the model and the accuracy for pipelining is Gradient Boost Classifier - 98.1%, AdaBoost Classifier - 98.8%, Random Forest Classifier - 98.1%, Extra Trees Classifier - 96.3%, Logistic Regression - 97.5%, Linear Discriminant Analysis - 95% is the best result. Their next step of this work is to find the cause of how non-diabetics can be affected by diabetes.

Muhammad Daniyal Baig et al. 2020 [6] built a system to detect diabetes using four machine learning classifier algorithms. The accuracy for the algorithms are Random forest - 98%, Logistic regression - 84 %, KNN - 88%, Gradient Descent - 90%. And they also generate a ROC curve. For ROC the percentage are for Random forest -99%, Logistic regression - 76%, KNN - 85%, Gradient Descent - 87%. Random forest gave the best result for both cases and Logistic regression gave the worst result for both cases. They will focus on young people's diabetes diagnosis in the future.

Deepti Sisodia et al. 2018 [7] sketched a model by WEKA to prophesy that can give the maximum accuracy of the probability of diabetes in people. They have used three machine-learning algorithms and ROC to confirm the results. The accuracy for different algorithms are Naive Bayes - 76.30%, SVM - 65.10%, Decision Tree - 73.82% and ROC are Naive Bayes - 81%, SVM - 50%, Decision Tree - 75%. Naive Bayes gave the best result in both cases. The tasks of diabetes analysis automation include some other algorithms of machine learning that can be expanded and enhanced.

Muhammad Azeem Sarwar et al. 2018 [8] developed a model to prognosticate the exactness of whether people are infected of diabetes or not. Six machine learning algorithms are applied and 70% of data for training and 30% of data for testing are kept. The accuracy of the model is 74% for LR, 77% for SVM and KNN, 74% for NB, and 71% for DT and RF. SVM and KNN have given the best accuracy. In the future, their plan is to integrate different methods into their existing model to get improved accuracy.

B. M. Patil et al. 2010 [9] used the Apriori algorithm to generate association rules among the factors which contribute to diabetes. They continue their work with type-2 diabetes which bearers mostly are pregnant women under 21 years. For rule-1 74 is the coverage and 100% is the confidence. By taking into account factors that influence diabetes, the statement of rules can be further enhanced.

Talha Mahboob Alam et al. 2019 [10] used the Apriori algorithm to show the association of diabetes with BMI and glucose level and also built a model with ANN, RF, and K-means clustering classification algorithm for the prediction of diabetes. The AUROC curve and accuracy are considered for the results and a confusion matrix is used to evaluate the accuracy. RF has given the accuracy of 74.7% and 0.806 is the AUROC curve result. ANN has given the accuracy of 75.7% and 0.816 is the AUROC curve result. And k-means clustering has given the accuracy of 73.6%. They selected a structured dataset for this study. In their future work, they will use an unstructured dataset. Their plan for the future is to apply the methods to different kinds of chronic diseases and to use various attributes for diabetes prediction.

III. Methodology

To develop the study we have followed a procedure that helped for developing this research work. Overall the process of our work is shown in Figure-1.

A. Dataset and Features Description
B. Dataset Preparation
C. Applying Correlation
D. Applying Machine Learning Techniques
E. Developing Classification Model
F. Model Performance Analysis

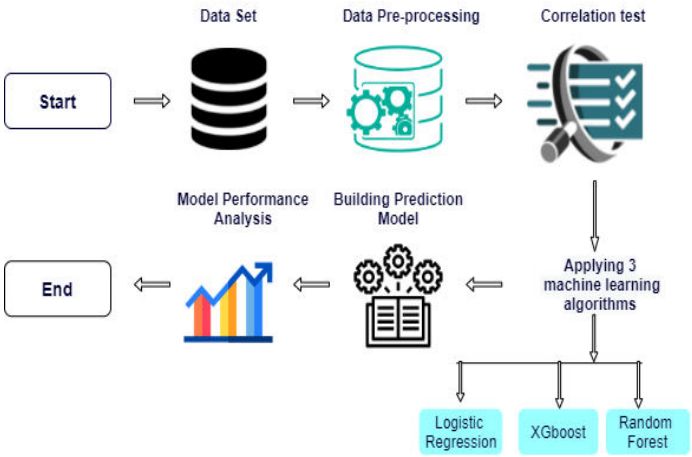

Fig. 1. System Architecture of this study

A. Dataset and Features Description

Our data have been gathered from Khulna Diabetes Center, Khulna. There are 289 people in the sample, and 13 diabetes-related characteristics are obtained from them. 10 of the 11 variables being gathered to identify the types of diabetes, and Table-I displays the participant's present diabetes type. The dataset's description has been shown in the table below.

TABLE I
DATASET DESCRIPTION

SL No.	Attributes	Type	Values
1.	Age	Continuous	{18 to 85}
2.	Height	Continuous	{130 to 180}
3.	Weight	Continuous	{32 to 100}
4.	BMI	Continuous	{14.4 to 42.2}
5.	Duration of Diabetes	Continuous	{0 to 20}
6.	Diastolic Blood Pressure	Continuous	{90 to 200}
7.	Systolic Blood Pressure	Continuous	{60 to 130}
8.	FBS	Continuous	{4.3 to 24}
9.	PPBS	Continuous	{ 5 to 28.5}
10.	Urine color of FBS	Nominal	{Blue, Green, Green Yellow, Lemon Green, Orange, Red, Yellow}
11.	Urine color of PPBS	Nominal	{Blue, Brick Red, Green, Green Yellow, Lemon Green, Orange, Red, Yellow}
12.	Type of medicine	Nominal	{Insulin, Tablet}
13.	Class	Nominal	{1,2}
14.	Gender	Nominal	{Male, Female}

B. Dataset Preparation

Amongst the most important stages mostly in the data mining process is data pre-processing. We have focused on the dataset for preparing the training dataset. For developing models we have transformed the dataset categorical to numerical. For females, we have replaced 0 and for males, we have replaced 1. Then we have transformed properly and we have made ready of this dataset. Finally, we have taken the decision for keeping 75% of train data and 25% is for testing.

C. Applying Correlation

Correlation is a most common and important technique for researchers, it helps to find a dynamic relationship between two variables from a dataset. This relationship finds that those variables are related to each other positively or negatively. Also, this technique gives results if there is no relation at all. The correlation test of this dataset is shown in Figure-2, where we have used Pearson's correlation [11]. Mainly we can know a relationship between those two variables if one variable is affected by another variable.

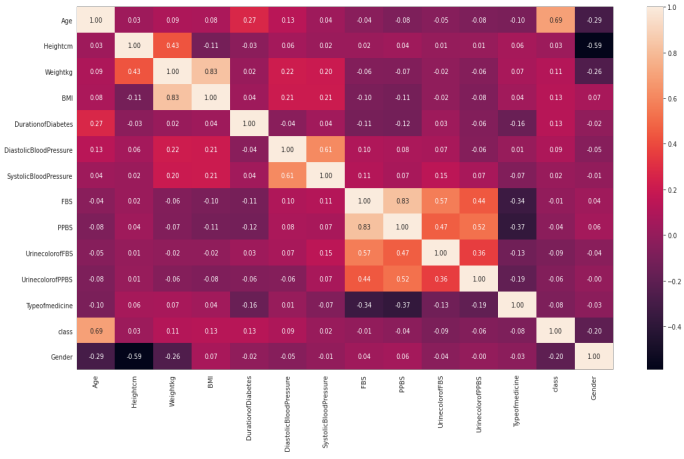

Fig. 2. Correlation test of our dataset

D. Applying Machine Learning Techniques

For the building model, we have used four classification algorithms named Logistic Regression, XGBoost, Random Forest, Decision Tree. We have analyzed the prediction of diabetes and why humans are being attacked by diabetes.

1) Logistic Regression: Logistic regression is a machine learning strategy that has taken advantage of statistics. The logistic function is the root of logistic regression so the main concept of this algorithm comes from this function. This function is additionally referred to as the sigmoid function. Among all algorithms of machine learning, Logistic regression is the most popular and comes after linear regression. They can be compared in diverse manners but their usage is different. The requirement of linear regression comes when to predict values and at the time of classification, logistic regression is used [12].

2) XGBoost: Basically, XGboost is a boosting algorithm. It is also a framework that runs in multiple languages. It is employed in regression as well as classification. It belongs to supervised machine learning. This algorithm is the advanced version of gradient boost. It is portable and language-independent. It is an effective and excellent algorithm which incorporates many weak classifiers into a solid classifier. It is also useful for solving different multileveled classification problems. This algorithm is very popular and widely used for its better speed and performance[13]. The speed of the algorithm includes parallelization, cache optimization, and out-of-memory computation. The performance includes automatic regularization - preventing the model from overfitting, auto pruning of the tree - the tree cannot make a group beyond a certain level and handling the missing value. Adding trees constantly and spitting the features constantly to grow a tree is the main concept behind this algorithm. Several steps are in the workflow of this algorithm - construct the tree, calculate similarity weight, calculate information gain, and for avoiding overfitting a new tree is added to the model. Whenever a tree is constructed a binary classifier is needed[14].

3) Random Forest: Random Forest is an ensemble classifier that uses decision tree algorithms in a randomized way. This algorithm is employed in regression as well as classification. It belongs to supervised machine learning. Leo Breiman was the developer of Random Forest which is considered the greatest classifier algorithm for a wide range of data. Any kind of pruning is not used here to grow the trees. This algorithm demonstrates randomness in two particular cases, to make a bootstrap dataset and to make decision trees from this dataset. This algorithm generates the result very fast and the accuracy of the prediction is very high. A wide range of input can be handled by this algorithm easily. This subsurface randomization scheme is combined with the bagging to prove each new tree by replacing the training data set [15]. In Figure-3 we have showed the box plot of dataset between all features and gender.

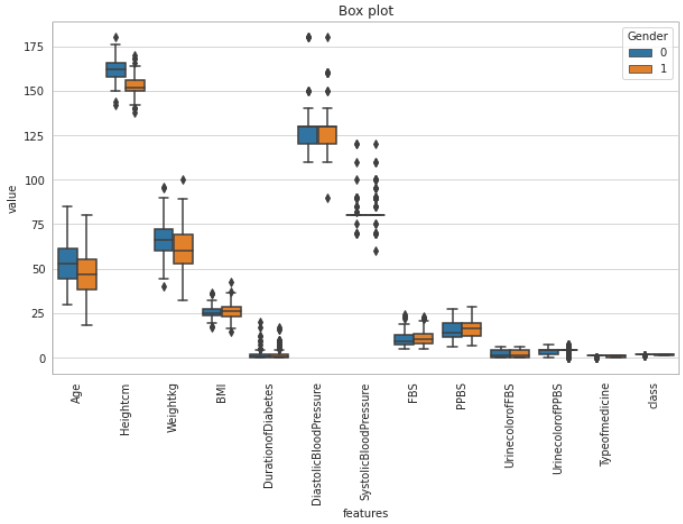

Fig. 3. Box Plot of this dataset.

E. Developing Classification Model

For measuring accuracy value there is a procedure. In our study, we have calculated accuracy, Precession, Recall, and F1 score.

TP - Rate: This True-Positive rate is measured by the total number of positive predicted numbers and the total actual number of positive cases [16].

$$TP - Rate = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

FP - Rate: This False Positive rate is always measured by calculating the total number of negative predicted numbers to the total number of negative numbers [17].

$$FP - Rate = \frac{False\ Positive}{FalsePositive + True\ Negative} \quad (2)$$

Precision: This value is determined from the total number of predicted positive values to all of the possible positive cases.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (3)$$

Recall: Recall is the value of the total number of predicted positive results to the total number of actual positives values.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

F- Measure: This value is used to present the overall statistics. This value is the weighted harmonic mean value of the recall and precision [18].

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

We have gotten accuracy from these four different algorithms. The performance of those algorithms is shown in Table-II.

TABLE II
MODELS PERFORMANCE OF ALL ALGORITHMS.

Algorithm Name	Accuracy	Precision	Recall	F1-Score
Logistic Regression	88%	0.81	0.95	0.88
XGBoost	86.36%	0.82	0.93	0.87
Random Forest	86.36%	0.80	0.98	0.88

We have gotten best accuracy rate from Logistic Regression and all of performance of all models like accuracy, Precision, Recall and F1 Score is shown to Figure-4.

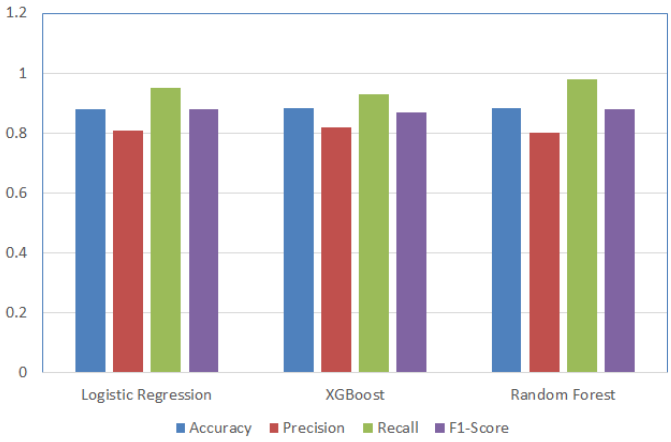

Fig. 4. Performance of all algorithms.

F. Model Performance Analysis

From those 3 machine learning algorithms, we have gotten various results. For the Logistic Regression algorithm we have gotten the best accuracy which is 88% and for XGBoost, Random Forest we have gotten 88.36% accuracy and for XGBoost the Precision is 82%, recall is 93% and F1-Score is 87%. Also for Random Forest the precision is 80%, Recall is 98% and F1-Score is 88% and last of all for Logistic Regression the Precision is 81%, Recall is 95% and F1-Score is 88%. We have also shown our model performance by ROC curve. We have compared our True Positive and False Positive rate in ROC, which is shown in Figure- 5, 6, 7 .

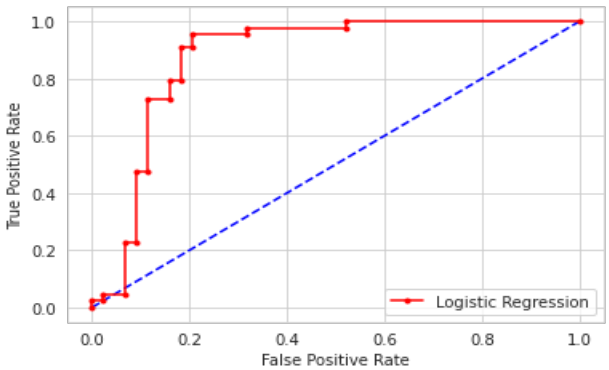

Fig. 5. ROC Curve of Logistic Regression

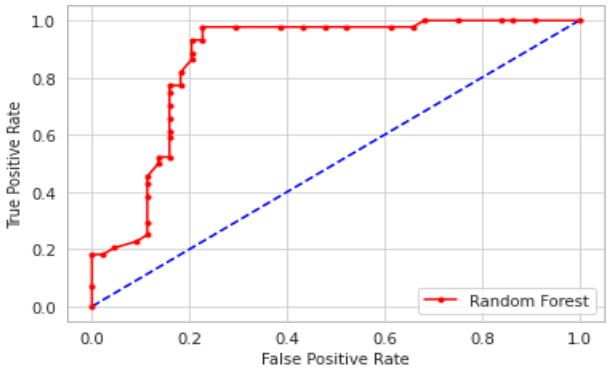

Fig. 6. ROC Curve of Random Forest

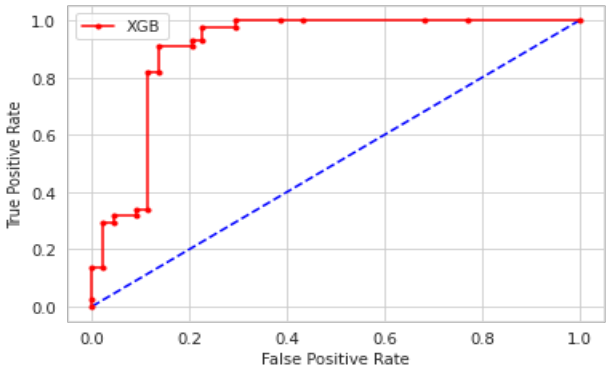

Fig. 7. ROC Curve of XGBoost.

IV. Conclusion

Diabetes is a challenging disease in medical science. We have focused on the computer-aided system for challenging this situation. Basically, we have worked on Type-2 diabetes data and we have classified data for developing a model by using three different algorithms named Logistic Regression, XGBoost, and Random Forest. We have gotten a good accuracy from this dataset by using those algorithms. This outcome might help for the medicine suggestion of a patient. Also, this system can predict the early stage of diabetes attack. In our work, we have worked on a dataset which quantity is not so much. For developing our result we have worked on 289 instances. It's possible to built better model from large number of dataset that's why we didn't get best model's performance. In the future, we are focusing on developing an application in which expert systems will help for predicting this disease and also that expert system will give suggestions for medicine and daily activities.

References

[1] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", Frontiers in Genetics, vol. 9, 2018. Available: 10.3389/fgene.2018.00515

[2] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection", Journal of Big Data, vol. 6, no. 1, 2019. Available: 10.1186/s40537-019-0175-6

[3] M. Faruque, Asaduzzaman and I. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019. Available: 10.1109/ecace.2019.8679365

[4] M. Tanvir Islam, M. Raihan, F. Farzana, P. Ghosh and S. Ahmed Shaj, "An Empirical Study on Diabetes Mellitus Prediction Using Apriori Algorithm", Advances in Intelligent Systems and Computing, pp. 539-550, 2020. Available: 10.1007/978-981-15-5148-2-48

[5] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science, vol. 165, pp. 292-299, 2019. Available: 10.1016/j.procs.2020.01.047

[6] M. Daniyal Baig and M. Farrukh Nadeem, "Diabetes prediction using machine learning algorithms", 2020. Available:https://www.researchgate.net/publication/345991601 Diabetes prediction using machine learning algorithms

[7] D. Sisodia and D. Sisodia, "Prediction of Diabetes using Classification Algorithms", Procedia Computer Science, vol. 132, pp. 1578-1585, 2018. Available: 10.1016/j.procs.2018.05.122

[8] M. Sarwar, N. Kamal, W. Hamid and M. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", 2018 24th International Conference on Automation and Computing (ICAC), 2018. Available: 10.23919/iconac.2018.8748992 [Accessed 1 June 2021].

[9] B. Patil, R. Joshi and D. Toshniwal, "Association Rule for Classification of Type-2 Diabetic Patients", 2010 Second International Conference on Machine Learning and Computing, 2010. Available: 10.1109/icmlc.2010.67

[10] T. Mahboob Alam et al., "A model for early prediction of diabetes", Informatics in Medicine Unlocked, vol. 16, p. 100204, 2019. Available: 10.1016/j.imu.2019.100204

[11] P. Schober, C. Boer and L. Schwarte, "Correlation Coefficients", Anesthesia & Analgesia, vol. 126, no. 5, pp. 1763-1768, 2018. Available: 10.1213/ane.0000000000002864.

[12] M. Hassan, Z. Peya, S. Zaman, J. Angon, A. Keya and A. Dulla, "A Machine Learning Approach to Identify the Correlation and Association among the Students' Drug Addict Behavior", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020. Available: 10.1109/icccnt49239.2020.9225355

[13] P. Ghosh, S. Azam, A. Karim, M. Jonkman and M. Hasan, "Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases", 5th ACM International Conference on Information System and Data Mining (ICISDM2021), 2021.

[14] N. Tigga and S. Garg, "Predicting Type 2 Diabetes Using Logistic Regression", Lecture Notes in Electrical Engineering, pp. 491-500, 2020. Available: 10.1007/978-981-15-5546-6_42

[15] L. Wang, X. Wang, A. Chen, X. Jin and H. Che, "Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model", Healthcare, vol. 8, no. 3, p. 247, 2020. Available: 10.3390/healthcare8030247

[16] M. Raihan, M. Islam, P. Ghosh, M. Hassan, J. Angon and S. Kabiraj, "Human Behavior Analysis using Association Rule Mining Techniques", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020. Available: 10.1109/icccnt49239.2020.9225662

[17] P. Sittidech and N. Nai-arun, "Random Forest Analysis on Diabetes Complication Data", Biomedical Engineering / 817: Robotics Applications, 2014. Available: 10.2316/p.2014.818-047

[18] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, M. Jonkman, "A Comparative Study of Different Machine Learning Tools in Detecting Diabetes," 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES-2021), 2021.