Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm

Md. Tanvir Islam[1], M. Raihan[2], Fahmida Farzana[3], Nasrin Aktar[4], Promila Ghosh[5] and Sajib Kabiraj[6]
Department of Computer Science and Engineering, North Western University, Khulna, Bangladesh[1-6]
Khulna University of Engineering & Technology, Khulna, Bangladesh[1,3]
Emails: tanvirislamnwu@gmail.com[1], raihanbme@gmail.com[2], mraihan@nwu.edu.bd[2], tanni.dorsonindrio@gmail.com[3]
nasrinlipinwu@gmail.com[4], me@promila.info[5] and kabirajsajib@gmail.com[6]

***Abstract*—A non-communicable disease Diabetes is increasing day by day at an alarming rate all over the world and it may cause some long-term issues such as affecting the eyes, heart, kidneys, brain, feet and nerves. It is really important to find an effective way of predicting diabetes before it turns into one of the major problems for the human being. If we take proper precautions on the early stage, it is possible to take control of diabetes disease. In this analysis, 340 instances have been collected with 26 features of patients who have already been affected by diabetes with various symptoms categorized by two types namely Typical symptoms and Non-typical symptoms. The purpose of this study is to identify the Diabetes Mellitus type accurately using Random Forest algorithm which is an Ensemble Machine Learning technique and we obtained 98.24% accuracy for seed 2 and 97.94% for seed 1 and 3.**

***Keywords*—Machine Learning, Random Forest, Ensemble Learning, Diabetes, Prediction, Typical, Non-typical.**

I. Introduction

Diabetes Mellitus (DM) is one of the major diseases among non-communicable diseases (NCDs) which makes a huge contribution to morbidity and mortality. Moreover, DM is known as Diabetes by which a group of metabolic disorders characterized by high blood sugar levels over a prolonged period. Insulin controls the level of glucose in the blood as a major hormone of the human body. At the time of the generation of insulin is diminished from islets of Langerhans in the pancreas than the Glucose level increment gradually and it causes diabetes [1]. Besides, DM is a condition that occurs when the body cannot utilize glucose while glucose is the main source of energy in the body cells. One of the diabetes types, when the pancreas doesnt make enough insulin is called Type1 Diabetes (T1D). On the other hand, when the body cannot respond to the insulin that is called Type-2 Diabetes (T2D). As a result, it continuously increases the level of glucose in the blood and leading to symptoms such as increased urination, extreme thirst and unexplained weight loss and many more [2]. As stated in statistics, an estimated 1.5 million deaths caused by diabetes and 2.2 million deaths in 2012 directly caused due to high blood sugar [3]. According to the International Diabetes Federation (IDF), the number of people having diabetes has increased from 30 million in 1985 to 150 million in 2000 and continuously 246 million in 2007 [4]. It seems that the expects number will be 380 million by 2025 of people with diabetes. In November 2013, to the aspect of the report presented by the World Health Organization (WHO) mentioned that the village people are less probable to be affected by diabetes than the urban people [5]. In 2016, statistics of WHO showed that about 402 million people the most living in low and middle-class economical countries globally had affected by diabetes. The estimated result according to IDF, 7.1 million people with diabetes and the equal number of people with undetected diabetes [6]. So, it is time to find out the reliable solutions to solve this issue. In Data Mining, Machine Learning (ML) is one of the sections of algorithms which enables the application to predict effectively and efficiently.

The goal of our analysis is to classify the types of DM accurately using the ML approach named Random Forest (RF).

Another part of the manuscript is arranged as follows: in section II and section III, the related works and methodology have been elaborated. In section IV the outcome of this analysis has been discussed with the impulsion to justify the significance of this exploration. Finally, this research paper is resolved with section V.

II. Related Works

An analysis was conducted by a research team on big data of health-care to predict diabetes disease accurately. The dataset they used contains 9 features with having numerical and nominal attributes. They aimed to build a classifier model by using ML techniques. In the analysis, Support Vector Machine (SVM) gave 79.13% accuracy [7]. Similarly, another research group used ML techniques to do diabetes classification. They used several ML algorithms such as Naive Bayes, SVM, RF and Simple CART. Among those Random Forest gave 76.5% accuracy in terms of classifying diabetes disease [8]. Debadri Dutta, et al. have analyzed on critical features for predicting diabetes. The dataset they used has 9 variables. In the study, they found RF to give 84% accuracy [9]. Sreekanth Rallapalli et al. have proposed a predictive model using CART model & scalable RF to classify diabetes based on various factors. They used a dataset contains 1500 instances with having 6 attributes. In this research, they found that the Scalable RF algorithm gave 87.5% accuracy whereas the normal RF algorithm gave 75% accuracy [10]. Soumayadeep Manna and his cooperators have proposed a system to predict important factors that cause diabetes. They used a dataset which has 3075 instances and each instance has 8 features. They have used Logistic Regression (LR) and Random Forest whereas RF gave 86.70% accuracy and LR gave 89.17% accuracy [11]. In the same manner, research has been conducted based on ML algorithms where researchers used SVM, AdaBoost, Bagging,

K-NN and RF algorithms with a dataset of 506 instances that has 30 features for each instance. They got 75.49% accuracy for AdaBoost, 76.28% for Bagging, 72.33% for K-NN, 75.30% for RF, 72.72% for SVM [12]. Likewise, another research team proposed a predictive model using the RF algorithm based on some variables like age, weight, hip, waist, height, etc. In the study, they performed the analysis based on 4 groups of datasets and for group number 4, RF gave the highest accuracy which is 84.19% [13].

From the above discussion, it is clear that diabetes is becoming a major issue in health-care. So, it is the high time to find out proper solutions to get rid of this serious disease. Since it is proved that using Data Mining and ML techniques we can predict and classify human diseases effectively we decided to work on one of the biggest diseases called diabetes. We studied on the background of Data Mining, ML and DM which motivated us to work and contribute something for people.

III. Methodology

The overall work-flow of the analysis has been shown in Fig. 1 where we can divide the whole process into four main segments. They are:

- Data collection
- Data preprocessing
- Data training
- Application of Random Forest

A. Data collection

We have collected the dataset from Khulna Diabetes Center, Khulna, Bangladesh. The dataset has 340 instances and each instance has 26 unique features. The features of the dataset have shown in Table I and the dataset has two types of symptoms Typical and Non-typical. The symptoms have shown in Table II. The dataset also has a feature except the 26 features named as **Outcome** which has three subcategories have also been presented in Table I.

B. Data preprocessing

There were several missing information in the dataset. So, it was essential to fill up the missing information before performing the analysis on the dataset. We have used a function named *ReplaceMissingValues* in **Waikato Environment for Knowledge Analysis (WEKA)** version 3.8 to fill up the missing data which can replace missing information based on mean, median and mode [14]. We have also used two other functions, namely, **Randomize** and **Resample** in WEKA 3.8.

C. Data training

We have used the 10-Fold Cross-validation technique to train the dataset. In this way, data divided into ten parts. It is a validation method that came from K-Fold Cross-validation where K is a specific parameter which can be assigned with numeric values, for example, K=5,10,15,20n. Different techniques on different dataset have proved that the 10-Fold Cross-validation is the best option to apply on dataset to get estimate error and it also proves that the stratification improves outcomes slightly. So, it is enough to divide data into ten parts [14]. That is the reason behind we used K=10.

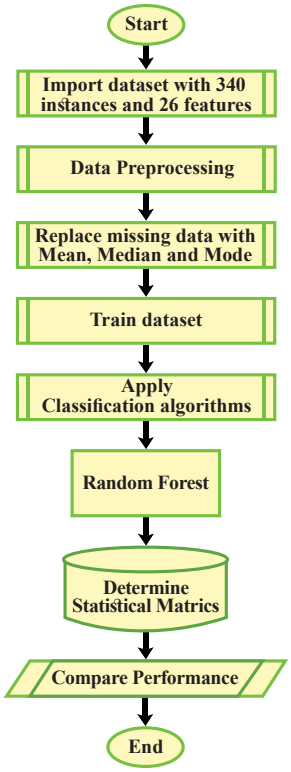

Fig. 1: Work-flow of the analysis

D. Application of Random Forest:

Random Forest is an ensemble ML algorithm which works based on a decision tree and it is a reliable way to enhance the performance of a system. It works basis on several learning algorithms. It merges the results of the learning algorithms to produce an optimal outcome. Therefore, Ensemble ML algorithms are efficient to produce an accurate result. RF uses C4.5 or J48 as its classifier. In 2001 RF was introduced by Breiman, which combines Bagging with random feature selection for decision trees. RF is a supervised classification algorithm. In this technique, each member of ensemble trained on a bootstrap replicate as in Bagging. Then by selecting the features decision trees are grown up and slit on at each node from randomly selected features can be defined by F [15]. So,

$$F = b * log_2 (k + 1) * c \cdots (1)$$

Where, k = total number of features

We do not perform any pruning on these random trees. As the RF works based on several decision trees so it can easily overcome the overfitting issues and it also has less variance than a single decision tree. Additionally, it is highly flexible and able to produce high accuracy in fact, when the dataset contains a large number of missing information [15].

IV. Outcomes

The outcomes from the study have been analyzed based on several performance parameters which have given below:

TABLE I: Feature List

Features Name	Sub Category	Data Distribution
Sex	Male	52.941%
	Female	47.059%
Age	Minimum: 22 years	$Mean \pm Std$
	Maximum: 75 years	50.189 ± 14.933
Profession	Govt	0.294%
	Retired	13.529%
	Farmer	5%
	Housewife	30%
	Business	14.412%
	Engineer	5.882%
	Journalist	1.176%
	Driver	1.471%
	Police	18.529%
	Day Labour	3.235%
	Clerk	9.824%
	Executive Officer	0.882%
	Teacher	0.588%
	Miscellaneous	0.294%
Height	Minimum: 140 cm	158.078 ± 7.334
	Maximum: 176 cm	
Weight	Minimum: 43 kg	61.982 ± 8.965
	Maximum: 88 kg	
BMI	Minimum: 19.2 kg/m-2	24.793 ± 2.877
	Maximum: 33.3 kg/m-2	
Heart rate	Minimum: 60 BPM	75.723 ± 4.819
	Maximum: 80 BPM	
Systolic BP	Minimum: 100 mmHg	124.588 ± 11.11
	Maximum: 160 mmHg	
Diastolic BP	Minimum:66 mmHg	80.48 ± 3.838
	Maximum: 100 mmHg	
Blood sugar before meal	Minimum: 5.1 mmol/L	12.272 ± 3.913
	Maximum: 22 mmol/L	
Blood sugar after meal	Minimum: 6.7 mmol/L	17.407 ± 4.551
	Maximum: 28.8 mmol/L	
Urine color before meal	Green	77.647%
	Yellow	5.588%
	Blue	14.706%
	Red	0.294%
Urine color after meal	Green	84.411%
	Yellow	2.647%
	Orange	11.471%
	Light-green	0.588%
	Blue	0.822%
Drug history	Yes	99.118%
	No	0.882%
Weight loss	Yes	82.647%
	No	17.353%
Thirst	Yes	94.411%
	No	5.588%
Hunger	Yes	82.647%
	No	17.353%
Relatives	Yes	75.294%
	No	24.706%
Physical activity	Yes	96.764%
	No	3.235%
Smoking	Yes	6.176%
	No	93.824%
Tabaco chewing	Yes	75.294%
	No	24.706%
Headache for high blood pressure	Yes	96.764%
	No	3.235%
Burning extremities	Yes	93.824%
	No	6.176%
Weakness	Yes	75.294%
	No	24.706%
Symptom duration	Minimum: 1 days	193.185 ± 187.226
	Maximum: 1460 days	
Diabetes mellitus	Yes	100%
	No	0%
Outcome	Typical	58.824%
	Non-typical	13.529%
	Both	27.647%
Standard Deviation = Std		

TABLE II: Symptom Names and Types

Symptom Types	Symptom Names
Typical	Thirst
	Hunger
	Weight Loss
	Sexual Weakness
Non-typical	Headache for High Blood Pressure
	Burning Extremities
	Physical Weakness

A. Seed

A seed is a random number in WEKA which can be changed randomly to observe the results for different seed numbers.

B. Correctly Classified Instances:

It defines the accuracy of any proposed model or algorithm [14],[16].

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \cdots (2)$$

C. Kappa Statistics (KS)

KS is used to evaluate the statement among foreseen and watched arrangements of a dataset [14], [17].

$$KS = \frac{R_0 - R_e}{1 - R_e} \cdots (3)$$

Where, R_0 = Relative watched understanding among raters, R_e = Theoretical likelihood of chance statement.

D. Mean Absolute Error (MAE)

It is an estimation that defines the difference between two continuous variables [17].

$$MAE = \frac{\mid p_1 - b_1 \mid + ... + \mid p_n - b_n \mid}{n} \cdots (4)$$

Where, p = Predicted value b = Actual value

E. Relative Absolute Error (RAE)

It takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor [16].

$$RAE = \frac{\mid p_1 - b_1 \mid + \cdots + \mid p_n - b_n \mid}{\mid b_1 - \bar{b} \mid + \cdots + \mid b_n - \bar{b} \mid} \cdots (5)$$

F. Sensitivity/True Positive (TP) Rate

It estimates the proportion of genuine positives that are accurately distinguished [17].

$$TPR = \frac{T_p}{P} \cdots (6)$$

G. Precision (PRE)

It can be defined as [17]:

$$PRE = \frac{T_p}{T_p + F_p} \cdots (7)$$

H. Recall (REC)

It can be defined as [18]:

$$REC = \frac{T_p}{T_p + F_n} \cdots (8)$$

I. F-Measure

If we denote FM as F-Measure then [18],

$$FM = 2 \times \frac{PRE \times REC}{PRE + REC} = \frac{2 \times T_p}{2 \times T_p + F_p + F_n} \cdots (9)$$

J. MCC

The full meaning of MCC is Matthews Correlation Coefficient which is the cohesion between PRE and REC [17].

K. ROC Area

It is the probability that a randomly chosen positive instance in the test data is ranked above a randomly chosen negative instance, based on the ranking produced by the classier [17].

L. PRC

It is an elective summary measurement that is favored by a few specialists, especially in the data recovery zone [17].

M. Explanation of the Analysis

We have performed the analysis in three different seeds, seed 1, seed 1 and seed 3, where the values of the seeds vary from 1 to 3. The outcomes for the various seeds have shown in Table III. RF algorithm gives the highest accuracy at seed 2. At seed 2, it provides 98.24% accurate result to classify diabetes while it gives 97.94% accuracy for both seed 1 and seed 3. In addition, the values of KS are 0.963, 0.9682 and 0.963 while the values of MAE are 0.0309, 0.0316 and 0.313, respectively. Moreover, the outcomes of Root Mean Squared Error are 0.104 for both seed 1 and 3 while it is 0.1053 for seed 2. Although at seed 1 and 3 its TP Rate is 0.979, its 0.982 in seed 2. Also, it produces a better FP rate better at seed 2. Furthermore, the values of Precision, Recall and F-Measure are 0.979 for both seed 1 and seed 3 while at seed 2 it is around 0.982. The values of MCC is 0.962, 0.967 and 0.962 for seed 1, 2 and 3, respectively. Nevertheless, ROC Area and PRC Area are the same in terms of seed 1, seed 2 and seed 3.

Fig. 2 is showing the ROC Curve for seed 1, where it has been plotted by True Positive Rate with respect to False Positive Rate.

Fig. 3 and 4 is representing the ROC Curve for seed 2 and 3 respectively. The graphs have also been plotted by False Positive Rate with respect to True Positive Rate.

TABLE III: Outcomes of 3 seed of Random Forest algorithm

Evaluation Metrics	Random Forest Classifier		
	Seed 1	Seed 2	Seed 3
Accuracy	97.9412%	98.2353%	97.9412%
Incorrectly Classified Instances	2.0588%	1.7647%	2.0588%
Kappa Statistic	0.963	0.9682	0.963
Mean Absolute Error	0.0309	0.0316	0.0313
Root Mean Squared Error	0.1046	0.1053	0.104
Relative Absolute Error	8.2711%	8.4711%	8.3704%
TP Rate (Weighted Avg.)	0.979	0.982	0.979
FP Rate (Weighted Avg.)	0.023	0.022	0.023
Precision (Weighted Avg.)	0.979	0.983	0.979
Recall (Weighted Avg.)	0.979	0.982	0.979
F-Measure (Weighted Avg.)	0.979	0.982	0.979
MCC (Weighted Avg.)	0.962	0.967	0.962
ROC Area (Weighted Avg.)	0.999	0.999	0.999
PRC Area(Weighted Avg.)	0.998	0.998	0.998

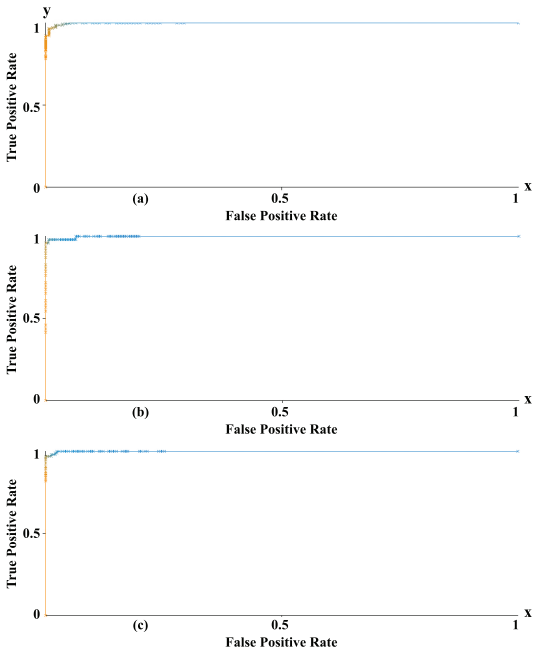

Fig. 2: ROC curves of Random Forest at seed 1

Fig. 5 illustrates the percentages of males and females in the dataset. It also represents the percentage of Yes and No for different nominal attributes of the dataset where yes means a person has the issue and no means the person doesn't have the problem. For example, if a person losses weight then the attribute 'weight'would be labeled by 'yes'whereas if the person doesn't loss weigh then for the person it would be labeled by 'no'. Similarly, all the attributes in the dataset have been labeled.

Fig. 6 illustrates the lowest and highest values of several numerical attributes including age, height, weight, BMI, heart rate etc.

Table IV shows the comparison between our proposed system and several existing systems based on accuracy as well as the number of instances and attributes have been used. The table clarify that our proposed system with Random Forest algorithm is better than the existing systems in terms of performance which gives the highest accuracy 98.24%.

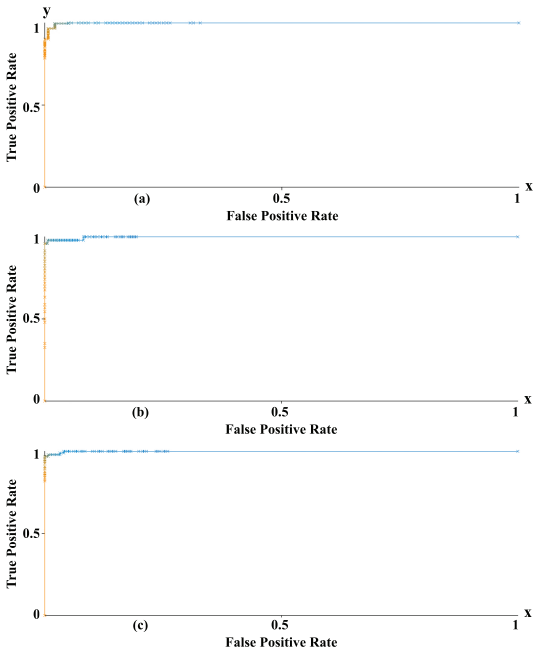

Fig. 3: ROC curves of Random forest at seed 2

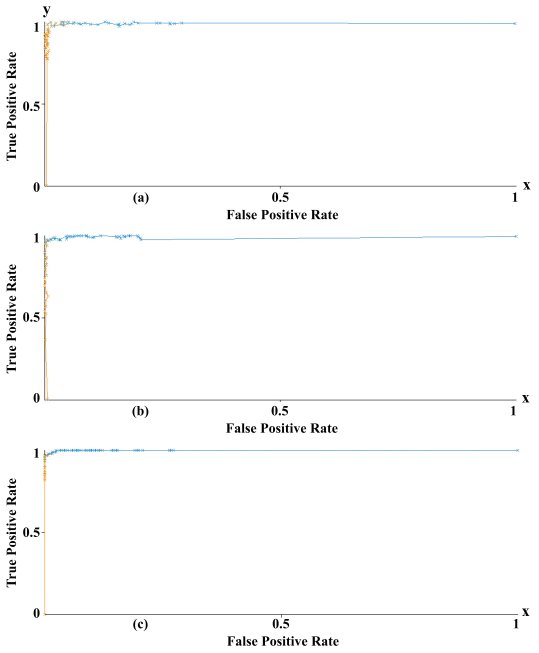

Fig. 4: ROC curves of Random forest at seed 3

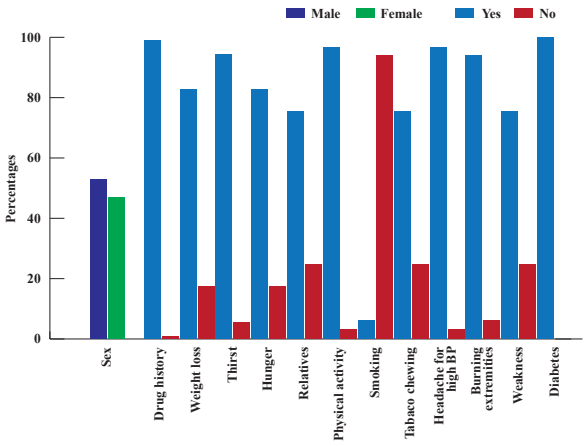

Fig. 5: Percentages of nominal features in terms of yes and no

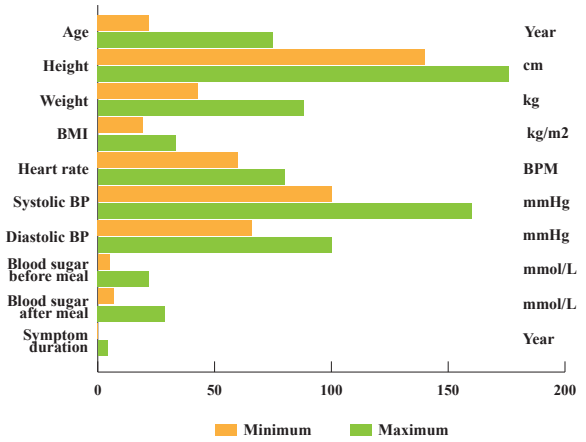

Fig. 6: Minimum and Maximum values of some numeric variables

V. Conclusion

In Bangladesh, a large number of people are suffering from DM and most of them are unaware of it. They don't know that they have diabetes disease. So, if it was possible to predict DM easily, it would be very useful for people. According to the analysis aspects with overall related research, the Random Forest is the best technique in terms of prediction diseases. We have faced several limitations during perform the analysis, for example, the collection of real information from patients was one of the challenges. Furthermore, there were several missing information in the dataset. Although we have faced these issues, we overcome and performed the analysis successfully using Machine Learning techniques. Finally, we have used Random Forest algorithm which has given 98.24% accuracy. Shortly, we would like to develop an intellect system to predict DM accurately using our proposed model.

TABLE IV: Comparison between Existing System with Proposed Technique

Reference Number	No. of Features	Sample	Algorithm	Accuracy
[8]	9	768	Random Forest	76.5%
[9]	9	768	Random Forest	84%.
[10]	6	1500	Random Forest	87.5
[11]	8	3075	Random Forest	86.7%
[12]	30	506	Random Forest	75.30%
[13]	10	373	Random Forest	84.19%
Our Proposed Systems	10	340	Random Forest	98.24%

REFERENCES

[1] "islets of Langerhans — Definition, Function, Location, & Facts", *Encyclopedia Britannica*, 2019. [Online]. Available: https://www.britannica.com/science/islets-of-Langerhans. [Accessed: 01 April 2020].

[2] L. Méjean, M. Kolopp and P. Drouin, "Chronobiology, Nutrition, and Diabetes Mellitus", *Biologic Rhythms in Clinical and Laboratory Medicine*, pp. 375-385, 1992.

[3] "Global Report on Diabetes", WHO, France, 2019.

[4] K. Ahmed and T. Jesmin, "Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach", *International Journal of Science and Engineering*, vol. 7, no. 2, 2014.

[5] "Diabetes", *WHO*, 2018. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: 5 April 2020].

[6] A. Mohiuddin, "Diabetes Fact: Bangladesh Perspective", *International Journal of Diabetes Research*, vol. 2, no. 1, pp. 14-20, 2019.

[7] A. Alahmar, E. Mohammed and R. Benlamri, "Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes", in *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*, Barcelona, Spain, 2018.

[8] A. Mir and S. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare", in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018.

[9] D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning", in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, Canada, 2018.

[10] S. Rallapalli and T. Suryakanthi, "Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm", in *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, Durban, South Africa, 2016.

[11] S. Manna, S. Maity, S. Munshi and M. Adhikari, "Diabetes Prediction Model Using Cloud Analytics", in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Bangalore, India, 2018.

[12] M. Raihan, Muhammad Muinul Islam, Promila Ghosh, Shakil Ahmed Shaj, Mubtasim Rafid Chowdhury, Saikat Mondal, Arun More, "A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome Using Machine Learning Approaches", in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018, pp. 1 - 6.

[13] Xu, W., Zhang, J., Zhang, Q., & Wei, X.,"Risk prediction of type II diabetes based on random forest model"in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*.

[14] I. Witten, E. Frank and M. Hall, "Data Mining practical Machine Learning Tools and Techniques", 3rd ed. *Morgan Kaufmann*, 2011, pp. 166-580.

[15] "What is Random Forest & Learn Random Forest using Excel", *New Tech Dojo*, 2017. [Online]. Available: www.newtechdojo.com/learn-random-forest-using-excel. [Accessed: 10 April 2020].

[16] "Relative Absolute Error", *Gepsoft.com*. [Online]. Available: https://www.gepsoft.com/gxpt4kb/Chapter10/Section2/SS15.htm. [Accessed: 11 April 2020].

[17] M. Islam, M. Raihan, S. Akash, F. Farzana and N. Aktar, "Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques", *Advances in Computational Intelligence, Security and Internet of Things*, vol. 1192, pp. 453-467, 2020. Available: 10.1007/978-981-15-3666-3_37 [Accessed 12 April 2020].

[18] M. Raihan, S. Mondal, A. More, P. Boni and M. Sagor, "Smartphone Based Heart Attack Risk Prediction System with Statistical Analysis and Data Mining Approaches", *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 3, pp. 1815-1822, 2017.