Известия вузов России. Радиоэлектроника. 2022. Т. 25, № 4. С. 116–122
Journal of the Russian Universities. Radioelectronics. 2022, vol. 25, no. 4, pp. 116–122

Medical Devices, Environment, Substances, Material and Product

UDC 006.21:006.29 Original article

https://doi.org/10.32603/1993-8985-2022-25-4-116-122

Machine Learning System for Predicting Cardiovascular Disorders in Diabetic Patients

Ali Mayya[1,2] ✉, Hanadi Solieman[1,2]

[1] Saint Petersburg Electrotechnical University, St Petersburg, Russia

[2] Tishreen University, Latakia, Syria

✉ alimayya1357@gmail.com

Abstract

Introduction. Patients with diabetes are exposed to various cardiovascular risk factors, which lead to an increased risk of cardiac complications. Therefore, the development of a diagnostic system for diabetes and cardiovascular disease (CVD) is a relevant research task. In addition, the identification of the most significant indicators of both diseases may help physicians improve treatment, speed the diagnosis, and decrease its computational costs.

Aim. To classify subjects with different diabetes types, predict the risk of cardiovascular diseases in diabetic patients using machine learning methods by finding the correlational indicators.

Materials and methods. The NHANES database was used following preprocessing and balancing its data. Machine learning methods were used to classify diabetes based on physical examination data and laboratory data. Feature selection methods were used to derive the most significant indicators for predicting CVD risk in diabetic patients. Performance optimization of the developed classification and prediction models was carried out based on different evaluation metrics.

Results. The developed model (Random Forest) achieved the accuracy of 93.1 % (based on laboratory data) and 88 % (based on pysicical examination plus laboratory data). The top five most common predictors in diabetes and prediabetes were found to be glycohemoglobin, basophil count, triglyceride level, waist size, and body mass index (BMI). These results seem logical, since glycohemoglobin is commonly used to check the amount of glucose (sugar) bound to the hemoglobin in the red blood cells. For CVD patients, the most common predictors inlcude eosinophil count (indicative of blood diseases), gamma-glutamyl transferase (GGT), glycohemoglobin, overall oral health, and hand stiffness.

Conclusion. Balancing the dataset and deleting NaN values improved the performance of the developed models. The RFC and XGBoost models achieved higher accuracy using gradient descending order to minimize the loss function. The final prediction is made using a weighted majority vote of all the decisions. The result was an automated system for predicting CVD risk in diabetic patients.

Keywords: cardiovascular disorders, diabetes, machine learning, preprocessing, feature selection, methods evaluation, correlational analysis

For citation: Mayya A., Solieman H. Machine Learning System for Predicting Cardiovascular Disorders in Diabetic Patients. Journal of the Russian Universities. Radioelectronics. 2022, vol. 25, no. 4, pp. 116–122. doi: 10.32603/1993-8985-2022-25-4-116-122

Conflict of interest. The authors declare no conflicts of interest.

Submitted 20.01.2022; accepted 11.03.2022; published online 28.09.2022

Introduction. Diabetes and cardiovascular diseases (CVD) remain to be among the most dangerous diseases that lead to death. According to the World Health Organization (WHO), an estimated 17.9 million people died from CVD in 2019, accounting for about 32 % of all deaths globally. Out of these cases, 85 % were related to strokes and heart attacks [1, 2]. Most adults diagnosed with diabetes and prediabetes were found to have been unaware of their health condition [3], which might lead to fatal cases if not treated in the early stages [4].

It is widely recognized that there is a close link between diabetes and CVD. Cardiovascular risk factors, such as obesity, hypertension, dyslipidemia, etc., are common in patients with diabetes, exposing them to increased risk of cardiac events. In addition, previous studies have found biological mechanisms associated with diabetes that independently increase the risk of CVD in diabetic patients [5].

At present, machine learning models are popular methods for analyzing and processing biomedical data. Such models have been successfully used for predicting common diseases, including diabetes [6], hypertension in diabetic patients [7], and CVD among diabetic patients [8]. Machine learning methods can reveal hidden factors and determine the common indicators between diabetes and CVD, which is important for early diagnosis and prediction of CVD in patients with diabetes.

In this paper, machine learning models are used to predict diabetes and CVD. Prediction models were developed separately to maximize their benefit for targeting a larger range of patients, despite the known association between these diseases.

Feature selection methods, such as Random Forest Classifier (RFC) and Extreme Gradient Boosting (XGB), are widely used to identify the most significant features common between diseases, thus facilitating disease classification and prediction [9]. Prediction models, including training and testing processes, were developed based on the National Health and Nutrition Examination Survey (NHANES) dataset.

Methods. Several machine learning models were utilized in this research. First, the model is fed with training data containing the recorded observations and the corresponding labels for the observations category in supervised learning. After that, the model predicts which output label should be associated with the new given observation. The approach is depicted in Fig. 1, starting from raw data through the development of the classification models ending their evaluation in predicting diabetes or cardiovascular disease. The flow diagram consists of four stages: data mining and modeling, model development, model evaluation, and classification and correlation.

Data Mining and Modeling. The research methodology is pipeline data mining and modeling, including preprocessing, normalization and standardization. The first step in preprocessing involves converting the NHANES data (raw patient records) into an acceptable and suitable for-

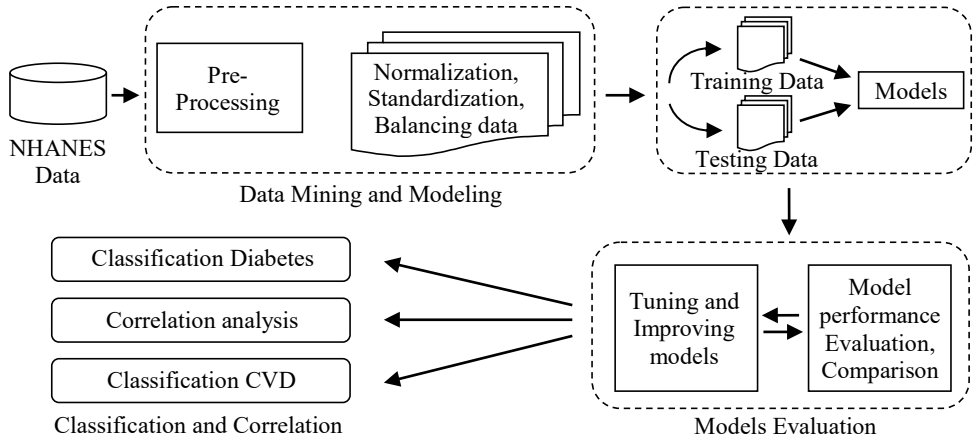

Fig. 1. A flow chart visualizing the main stages in the model approach

mat for the machine learning algorithms, cleaning NaN values, removing redundant rows and columns, and replacing the mean value. The second step of preprocessing involves labeling the data and the output to split the data into training and testing sets. This paper includes physical examinations and laboratory incidents in the database. Binary labels are generated on the survey to facilitate the classification task, as shown in Tab. 1.

Tab. 1. Data labels conversion

Code in NHANES dataset	Diagnosing (label)
DIQ010	Diabetes (diagnosed with diabetes Yes = 1, No = 0)
DIQ160	Prediabetes (diagnosed with prediabetes Yes = 1, No = 0)
MCQ160F	CVD (stroke, heart failure)

After preprocessing the data separately for each disease category, the resulting patients' distribution is shown in Fig. 2. The histograms show the proportion of diagnosed patients with one of the diseases under study to the proportion of healthy individuals. The percentages were taken depending on the NHANES data after preprocessing and removing redundant and NaN values. The data distribution is as follows: 8.5 % were diagnosed with diabetes, 4.8 % were diagnosed with prediabetes, and 3.6 % had one form of CVD. The dataset was normalized and standardized using machine learning tools in Python. The majority of medical databases face the imbalance problem, leading to errors in classification. This problem was solved using the Synthetic Minority Oversampling Technique (SMOTE) for oversampling imbalanced classification datasets.

Model Development. The dataset resulting from the previous stage was split into training and testing sets. First, downsampling was used to produce a balanced 80/20 train/test split. Next, the training set was used to teach the models; then, the models were tested by the testing sets to evaluate the prediction models in the validation stage. The testing set will be fed to the models as unseen data, then the performance of the models will be evaluated with the testing set.

Model evaluation. All the models were tested by metric parameters that were based on the performance statistics in terms of accuracy, precision, recall, and F1 score. The latter sums up the prediction efficiency of a model by combining the recall and precision metrics. Depending on the value of accuracy and F1 score, the models were evaluated. Although multiple methods were used for the purposes of this research, such as logistic regression, SVM, decision trees, adaptive boosting, gradient boosting, random forests (RFC), and XGboost (XGB), we selected RFC and XGB due to their protential of feature selection and correlational analysis between diabetes and CVD.

Classification and Correlation. This pipeline output is a well-tuned classifier that can predict which output label should be associated with a new observation for diabetes and an individual CVD form.

In addition to the correlational analysis, which shows the common most significant indicators between diabetes and CVD and provides a medical interpretation for these indicators.

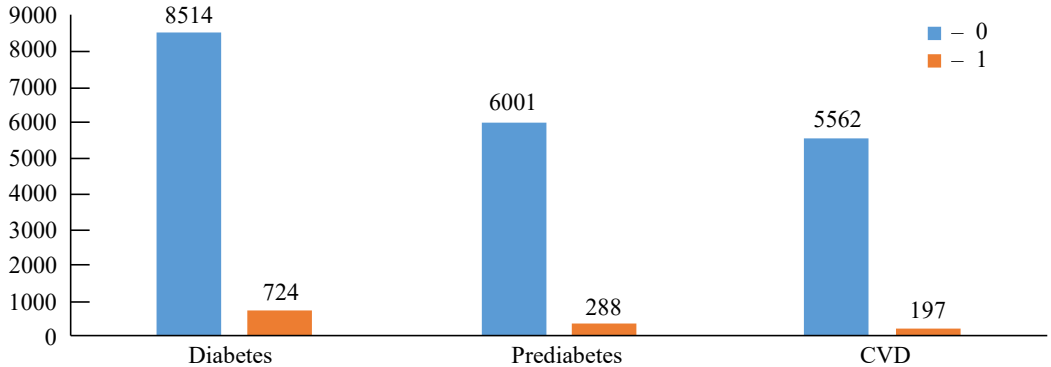

Fig. 2. Diabetes, prediabetes and CVD patients' proportion to the healthy people in NHANES dataset

Tab. 2. Evaluation parameters for the XGB and RFC models with different data cases

Method	Diagnosing	Dataset	Precision	Recall	F1-score	Accuracy
XGB	Prediabetes	Lab	0.71	0.99	0.83	0.79
RFC	Prediabetes	Lab	0.71	0.99	0.83	0.79
XGB	Prediabetes	Exam + Lab	0.69	0.97	0.81	0.77
RFC	Prediabetes	Exam + Lab	0.63	0.98	0.77	0.70
XGB	Diabetes	Lab	0.85	0.96	0.91	0.90
RFC	Diabetes	Lab	0.90	0.97	**0.94**	**0.93**
XGB	Diabetes	Exam + Lab	0.77	0.93	0.84	0.83
RFC	Diabetes	Exam + Lab	0.83	0.94	**0.88**	**0.87**

Results. Tab. 2 describes the evaluation metrics that were used to evaluate the diabetes and prediabetes classification produced by XGB and RFC classifiers based on the data used: only laboratory data (Lab), both physical examination and laboratory data (Exam + Lab). Tab. 2 shows a close convergence between the parameters. However, it can be seen that RFC demonstrates higher accuracy and F1 score (93 and 94 %, respectively) levels for diabetes based on both laboratory data only and physical examination plus laboratory data. Therefore, RFC can be considered to be the best method. Fig. 3 visualizes the performance of the models in terms of accuracy when classifying CVD based on physical examination plus laboratory data. Fig. 4 shows that RFC produces the best accuracy when classifying using laboratory data only.

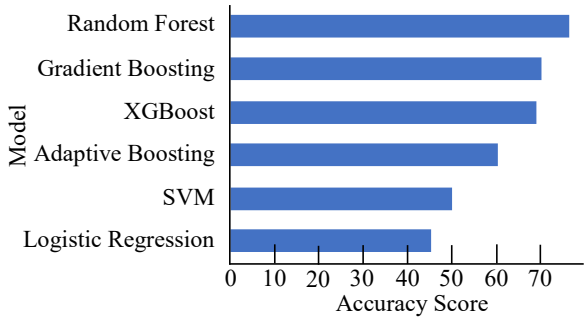

Fig. 3. Accuracy comparison for the models based on physical examination plus laboratory data

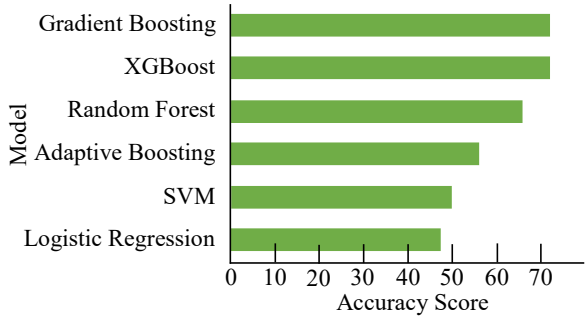

Fig. 4. Accuracy comparison for the models based on physical examination plus laboratory data

Hence, for the purposes of correlational analysis based on the selected features, XGB and RFC were used to select 24 most significant features from the laboratory dataset and from the physical examination plus laboratory dataset, respectively. Fig. 5 and 6 show the most significant features in each case.

Discussion. Diabetic and prediabetes prediction. The ensemble models trained on diabetic patients were shown to have a higher predictive power in terms of accuracy. This particularly concerns RFC (93 and 87 %), as can be seen from Tab. 2. For comparison, XGB showed the highest accuracy of 79 % (Lab) and 77 % (Exam + Lab) in predicting prediabetes. This decrease in detection performance between RFC and XGB can be explained by two factors: 1) a decreased number of observations; 2) boundary conditions for the recorded observations. The prediabetes dataset includes about 6300 available observations, compared to 9400 observations for diabetes. The difference in the size of the observation datasets was connected with the preprocessing stage, which aimed to delete NaN values, redundant observations, and those without corresponding diagnoses. Moreover, it can be seen from Tab. 2 that F1 scores are high in general, which suggest that the models are stable and that F1 scores increased gradually and achieved their maximal value in diabetic prediction by RFC based on the laboratory dataset.

Cardiovascular patient prediction. Fig. 3 demonstrates the performance of the developed models in terms of their accuracy in classifying CVD using physical examination plus laboratory data. Fig. 4 shows that RFC has the highest accuracy of 77 %

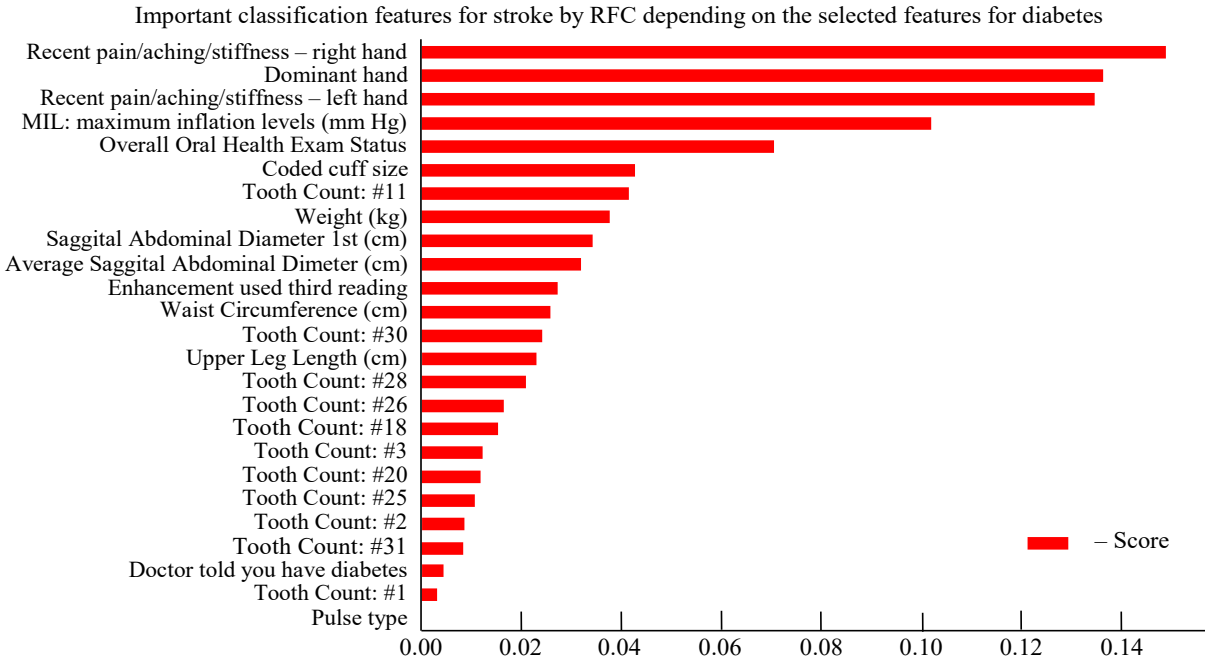

Fig. 5. Feature importance for diabetes classifiers (examination + laboratory). The most important features results for predicting CVD from the selected features for diabetes are shown

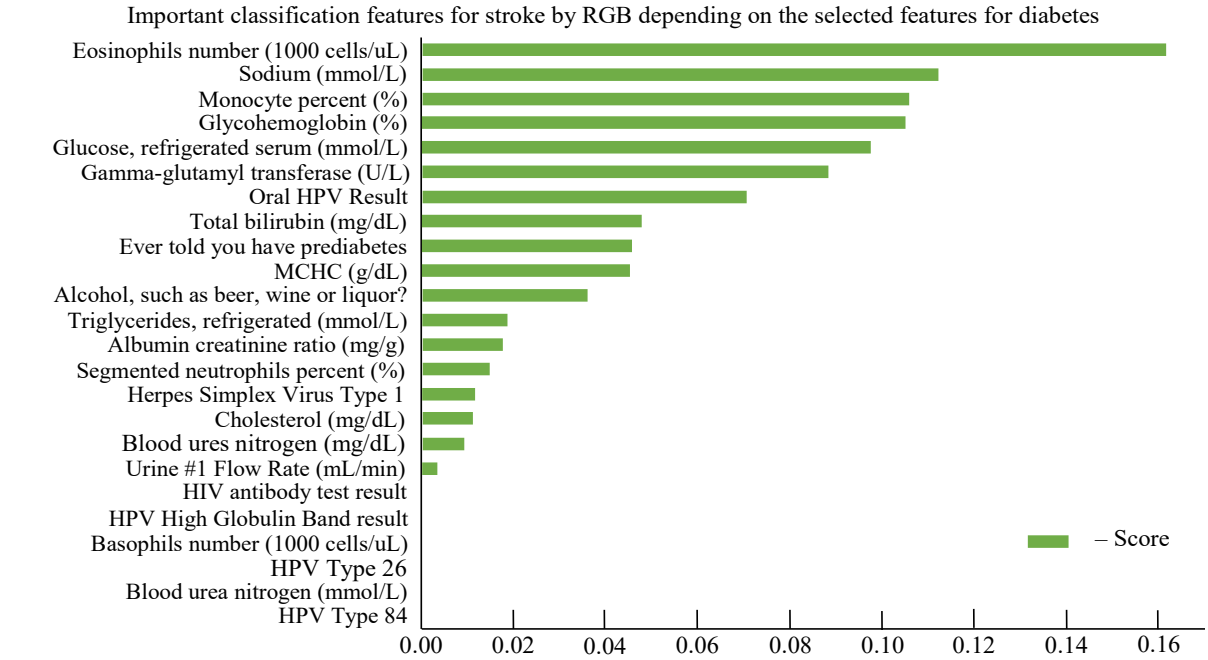

Fig. 6. Feature importance for diabetes classifiers (laboratory). The most important features results for predicting CVD from the selected features for diabetes are shown

when using the laboratory dataset only, compared to Gradient Boosting with an accuracy of 74 %. The size of observations, in this case, is about 5600, which explains a decrease in accuracy compared with diabetes prediction.

Correlational analysis. For the purposes of correlational analysis, five most significant indicators from each of the examination and laboratory datasets were selected for interpretation based on medical information in order to verify whether these indicators might place diabetic patients at risk of CVD. The most significant indicators were revealved through classification based on laboratory data: eosinophil count, sodium intake, monocyte percent, glycohemoglobin, and gamma-glutamyl transferace (GGT). Based on physical

examination data, the top five indicators comprised hand stiffness, overall oral health, weight, waist circumference, and upper leg length [10, 11].

Eosinophils are a kind of white blood cell, which, upon activation, release enzymes to fight foreign substances and infections [12]. Disbalances in the normal eosinophil count can represent an unhealthy blood status. A high level of sodium intake can raise blood pressure, thus leading to the risk of a heart disease or stroke [13]. Most of the sodium consumed with food comes in the form of salt.

Studies showed that an increased intermediate monocyte count is independently associated with CVD incidence [14]. Glycohemoglobin is a blood test that determines the level of glucose (sugar) bound to the hemoglobin in the red blood cells. Glycohemoglobin levels in the blood can predict CVD risk in people with diabetes. Moreover, it is suggested that good blood glucose control plays a significant role in reducing CVD risk [15]. Gamma-glutamyl transferase (GGT) is a special enzyme on the external surface of cellular membranes. GGT levels can be elevated under many pathophysiological conditions, and elevated GGT activity is related to CVD risk, such as coronary heart disease [16]. Body mass index and waist circumference were found to be associated with coronary heart disease [17].

The developed models also revealed other indicators that may contribute to CVD risk in patients with diabetes, including alcohol consumption, triglyceride level, and basophil count.

Conclusion. The preprocessing stage dealt with the problem of a highly uneven (imbalanced) dataset in terms of sample size and NaN values. Solving these problems improved the performance of the models. XGB and RFC models have demonstrated the highest accuracy, since these models minimize the loss function by building decision trees using gradient descent and making the final decision as a weighted vote of all the decision trees. XGB showed a relatively better results in terms of speed and performance compared to RFC, which is sensitive to the kernel type and slow with extensive datasets. However, in general, RFC showed the highest performance and, therefore, was used to determine diagnostic indicators. Among the top five most common predictors in diabetes and prediabetes patients were found to be glycohemoglobin, basophils, triglycerides, waist size, and body mass index (BMI). These analytical results agree well with published literature. The models identified eosinophils number (which is indicative of blood diseases), GGT, glycohemoglobin, overall oral health, and hand stiffness as predictors for cardiovascular diseases.

References

1. Benjamin E. J., Blaha M. J., Chiuve S. E. et al. Heart Disease and Stroke Statistics – 2017 Update. Circulation. 2017, vol. 135, no. 10, pp. e146–e603. doi: 10.1161/CIR.0000000000000485

2. Dinh A., Miertschin S., Young A., Mohanty S. D. A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning. BMC Medical Informatics and Decision Making. 2019, vol. 19, no. 1, p. 211.

3. Dounias G., Vemmos K., Alexopoulos E. Medical Diagnosis Of Stroke Using Inductive Machine Learning. Machine Learning and Applications. 1999, 4 p.

4. Flint A. J., Rexrode K. M., Hu F. B., Glynn R. J., Caspard H., Manson J. E., Willet W. C., Rimm E. B. Body Mass Index, Waist Circumference, and Risk of Coronary Heart Disease: A Prospective Study Among Men and Women. Obes Res Clin Pract. 2010, vol. 4, no. 3, pp. e171–e181. doi: 10.1016/j.orcp.2010.01.001

5. Khaw K.-T., Wareham N. Glycated Hemoglobin as a Marker of Cardiovascular Risk. Curr Opin Lipidol. 2006, vol. 17, no. 6, pp. 637–643. doi: 10.1097/MOL.0b013e3280106b95

6. Leon B. M., Maddox T. M. Diabetes and Cardiovascular Disease: Epidemiology, Biological Mechanisms, Treatment Recommendations and Future Research. World J Diabetes. 2015, vol. 6, no. 13, pp. 1246–1258. doi: 10.4239/wjd.v6.i13.1246

7. National Academies of Sciences, Engineering and Medicine; Health and Medicine Division; Food and Nutrition Board; Committee to Review the Dietary Reference Intakes for Sodium and Potassium. Dietary Reference Intakes for Sodium and Potassium. Ed. by Oria M., Harrison M., Stallings V. A. Washington (DC), National Academies Press, 2019, 594 p. doi: 10.17226/25353

8. Ndrepepa G., Kastrati A. Gamma-Glutamyl Transferase and Cardiovascular Disease. Ann Transl Med. 2016, vol. 4, no. 24, p. 481. doi: 10.21037/atm.2016.12.27

9. Parthiban G., Srivatsa S. Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. Intern. J. of Applied Information Systems. 2012, vol. 3, pp. 25–30.

10. SahBandar I. N., Ndhlovu L. C., Saiki K., Kohorn L. B., Peterson M. M., D'Antoni M. L., Shiramizu B., Shikuma C. M., Chow D. C. Relationship between Circulating Inflammatory Monocytes and Cardiovascular Disease Measures of Carotid Intimal Thickness. J. of Atherosclerosis and Thrombosis. 2019, vol. 27, no. 5, pp. 1–8. doi: 10.5551/jat.49791

11. Semerdjian J., Frank S. An Ensemble Classifier for Predicting the Onset of Type II Diabetes. arXiv:1708.074802017. 2017. doi: 10.48550/arXiv.1708.07480

12. Teimouri M., Ebrahimi E., Alavinia M. Comparison of Various Machine Learning Methods in Diagnosis of Hypertension in Diabetics with/without Consideration of Costs. Iranian J. of Epidemiology. 2016, vol. 11, no. 4, pp. 46–54.

13. National Diabetes Statistics Report. Available at: https://www.cdc.gov/diabetes/data/statistics-report/index.html (accessed 15.01.2022)

14. National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/index.htm (accessed 15.01.2022)

15. Cardiovascular diseases (CVDs). Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed 15.01.2022)

16. Zeya L. T. Essential Things You Need to Know About F1-Score. Towards Data Science. Available at: https://towardsdatascience.com/essential-things-you-need-to-know-about-f1-score-dbd973bf1a3 (accessed 15.01.2022)

17. DerSarkissian C. Eosinophils and Eosinophil Count Test. Available at: https://www.webmd.com/asthma/eosinophil-count-facts#1 (accessed 15.01.2022)

Information about the authors

Ali Mayya, Master student at the Department of Bioengineering Systems of Saint Petersburg Electrotechnical University, Bachelor (2019) in Electromechanics – Mechatronics of Tishreen University. Area of expertise: machine learning; deep learning; biomedical modeling; medical instrumentation; health analytics; digital healthcare.
Address: Tishreen University, Southern Entrance, Latakia, Syria
E-mail: alimayya1357@gmail.ru
https://orcid.org/0000-0002-4806-8587

Hanadi Solieman, Postgraduate student, Assistant at the Department of Bioengineering Systems of Saint Petersburg Electrotechnical University, Assistant at the Mechatronics program for Distinguished of Tishreen University. The author of 9 scientific publications. Area of expertise: medical instrumentation; medical informatics; processing and analysis of biomedical signals and data.
Address: Tishreen University, Southern Entrance, Latakia, Syria
E-mail: khsoliman@stud.etu.ru
https://orcid.org/0000-0002-9868-8960