ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)

UDC 004.8
DOI: https://doi.org/10.30837/ITSSI.2023.24.054

N. Huliiev, M. Peretiaha, A. Khovrat, D. Teslenko, A. Nazarov

STUDY OF PREDICTION AND CLASSIFICATION MODELS IN THE PROBLEMS OF DIABETES AMONG PATIENTS WITH A STROKE IN DIFFERENT LIVING CONDITIONS

The subject of the study in the article is the methods of predicting the development of diabetes. Diabetes mellitus is a non-communicable disease that has affected 425 million people, and by 2045 the number will only increase by 1.5 times. It has been proven to be an independent contributing factor to stroke development. When there is too much sugar in the blood, it negatively affects the arteries and blood vessels. People with this disease are more likely to develop atherosclerotic plaques and blood clots, which can lead to heart blockage and ischemic stroke. Having diabetes increases the risk and worsens the course of a stroke. According to the Framingham Study, the number of recurrent cases doubles. **The aim of the study** is to investigate methods of predicting and classifying the development of diabetes among people, in particular stroke patients, to prevent the development of other diseases. The complexity of the problem lies in the fact that there are as many undiagnosed cases as diagnosed ones, so about half of people suffer from the disease and the resulting complications due to improper or delayed diagnosis. Therefore, timely diagnosis of a disease that is difficult to detect is important in order to prevent the development of further complications. **The article solves the problem** of a multi-criteria task of choosing the best algorithm for predicting the occurrence of a disease. **The following methods are used in this paper**: multilayer perceptron, k-nearest neighbors method, decision tree, and logistic regression. Nowadays, machine learning has begun to apply to similar problems. In the 1950s and 1960s, there were attempts to combine the approaches to creating neural networks that existed at the time, which made it possible to calculate quantitative descriptions of human intelligence, and memorize, analyze, and process information, which resembled the work of the human brain. Medicine is one of the main areas of human activity where various classifier and neural network algorithms are gaining popularity yearly. They are trendy in disease diagnostics. **Results**: the initial conditions for choosing the best model are met by logistic regression. **Conclusions**: as a result of the study, the optimal model for predicting the development of the disease was selected.

Keywords: multilayer perceptron; neural network; prediction; stroke; diabetes mellitus.

Introduction

Diabetes mellitus (DM) is one of the most common diseases of the endocrine system. Today, diabetes is considered a pandemic, as the number of patients worldwide is increasing by 5–7% annually. It is predicted that by 2030 there may be approximately 360 million people with diabetes, and by 2040 the number will double to 640 million.

Diabetes mellitus is a metabolic group of diseases characterized by impaired insulin secretion and action, as well as hyperglycemia.

According to the 1999 classification, revised and improved in 2019, diabetes has the following types:

– type I – occurs as a result of the destruction of pancreatic β-cells during an autoimmune or undetermined process that causes insulin deficiency;

– type II – the most common type, which develops in case of impaired insulin secretion due to insulin resistance;

– hybrid forms of DM;

– gestational DM;

– DM of known etiology;

– unclassified DM.

The World Health Organization defines DM as a non-communicable disease that is one of the ten possible causes of death, as life expectancy in diabetics is reduced by 25%. Almost 80% of deaths are caused by cardiovascular complications, namely heart attack or stroke, which also lead to disability.

A stroke causes damage to the blood vessels in the brain because they are blocked by a blood clot or rupture, and therefore cannot carry oxygen and nutrients.

Causes of stroke:

– high blood pressure;

– high cholesterol;

– abnormal heart rhythm;

– overweight;

– diabetes mellitus;

– excessive stress.

Analysis of the problem and existing methods

Canadian experts conducted a study: they collected data on 12.200 patients over 30 diagnosed with type II diabetes. Of these, 9.1% were diagnosed with various types of strokes within five years. Studies have shown that long-term diabetes leads to macrovascular problems.

It is known that stroke due to diabetes is diagnosed in people under the age of 40 3–4 times more often, and after 40 years – 1.5–2 times more often than in nondiabetics. Mostly, ischemic variants occur (in 65% of cases, the atherothrombotic subtype), and mortality is 40.3–59.3%. Also, the likelihood of strokes increases due to high blood pressure, which can be accompanied by cerebrovascular syndrome, impaired consciousness, and pneumonia. Thus, the brain is affected, and neurological deficits slowly begin, but the lost functions are not fully restored. It is worth noting that in 46% of cases there are signs of leukoaraiosis in the periventricular zone, which shows the extent of cerebral vascular damage.

More than 80% of people with stroke in the setting of diabetes mellitus may have movement disorders.

To date, the causes of stroke caused by diabetes mellitus have not been fully identified. Researchers believe that in this case, stroke is a clinical syndrome of macroangiopathy caused by a disorder of carbohydrate metabolism.

Machine learning is now being used to solve such problems. In the 1950s and 1960s, there were attempts to combine the approaches to creating neural networks available at the time. They enabled computational capabilities that quantitatively described the features of human intelligence, memorization, analysis, and processing of information that resembled the work of the human brain.

In the modern world, many problems faced by specialists in various fields are solved with the help of machine learning, as there is a need to process a significant amount of information. As you know, there are countless medical records for each disease, so the question arises: "How to correctly predict the possible development of a disease?" The article [1] describes *Big Data* as a set of organized and unorganized data for which conventional methods are not effective, so in this case, machine learning is used as a way to find unnoticeable connections among a large number of queries. This method involves the stages of collecting and preparing information, selecting and training a model that can solve the problem of a particular industry. At the same time, the authors of the article emphasize that the use of machine learning in this case does not require deep knowledge and full immersion in the subject area, which also simplifies the solution of tasks when processing a significant amount of information.

It is worth noting that the first step is to define the criteria for choosing a model. The article [2] describes the construction of the principles of choosing the best model for decision-making, in particular, for hiring. The authors present indicators that provide useful information for a more *accurate* assessment of the algorithm’s effectiveness. In the process of developing classification algorithms, the authors calculate the *accuracy*, *precision*, *recall*, and *f*-measure criteria for each. The authors argue that *accuracy* should be used when the number of positive and negative examples is approximately equal; *precision* calculates the severity of possible consequences in case a negative example is identified as positive; *recall* is appropriate when there is no positive data; and the *f*-measure is effective when the records of one class significantly exceed the data of another. Thus, the study decided to apply the following criteria.

It is well known that many neural network models are used in medicine. Article [3] discusses the construction, analysis, and development of neural networks used to predict the development of endocrine disease. First of all, we are talking about the Kohonen neural network, multilayer perceptron, hybrid neural network, adaptive resonance theory models and its modified model – *Fuzzy-ART*. To choose the most optimal model, the following indicators are taken into account: neural network speed, reliability of results, amount of memory required, training method, and application principles. The best of these neural networks is the multilayer perceptron, which will be discussed in this paper.

Thus, the basis for the use of neural networks in medicine is the construction of a multilayer perceptron. The variable accuracy of the neural network for diagnosing cardiovascular diseases ranged from 64% to 94%. These were models of a multilayer perceptron with two hidden layers with an accuracy of over 90%. Their training was based on genetic algorithms.

The above proves that stroke and diabetes are dangerous not only because of their consequences, but also because they can lead to other serious diseases. Therefore, **the aim of the article** is to investigate methods of predicting the development of diabetes among stroke patients under different conditions of life and to choose the most optimal model for this task.

To select a model among a set of alternatives with different characteristics, a multi-criteria problem of choosing the best option is used.

Solving such problems can be difficult due to the ambiguity of the choice. In such cases, methods from two groups are used: the first is designed to reduce the number of evaluation criteria, in which case assumptions are made to rank the values of characteristics and compare all options; the second group of methods is aimed at removing bad alternatives before the comparison algorithm begins.

For our study, the preferred method is the first group's method – collapsing – a method in which all criteria of alternatives become one common one. The most commonly used methods are additive, multiplicative, and maximin collapsing.

Additive collapsing is presented as follows:

$$K(x)=\sum_{j=1}^{n} a_j K_j(x), \qquad (1)$$

where $K(x)$ – general criterion for the alternative $x \in X$;

$(K_1(x), .. . K_j(x), ... K_n(x))$ – a set of initial criteria;

n – a number that describes its number;

a_j – normalization factor, weight of the alternative's characteristic feature.

The best alternative is calculated as follows:

$$x^* = \arg\max_{x \in x} K(x). \qquad (2)$$

That is, the solution is the largest value calculated by convolution.

The best solutions of multiplicative and maximal convolution are also calculated using formula (2).

We have to choose the right method for the study. We cannot immediately reject all possible alternatives, so the methods of the second group are not suitable.

Let's choose additive convolution, because multiplicative convolution requires normalization of values from 0 to 1, whereby in the case of 0 we will have 0, despite other priorities of the criteria. First, it is necessary to determine the criteria by which all proposed alternatives will be evaluated, and then weights are calculated for each of them, i.e., the most important and effective alternative in decision-making. All criteria are distinguished by their indicators – qualitative and quantitative. This method works with the latter, so if you have the former, you need to replace them with the appropriate values of the latter type.

Once the quantitative indicators are ready, some of the alternatives can be eliminated using the Pareto principle if there are those that are worse than others in all criteria, and then normalization is necessary. The scores for the criteria differ in the scales of their values, i.e., mass is measured in kilograms, speed in meters per second or in seconds, so for a correct assessment of the values, they need to be normalized in the range from 0 to 1. Usually, the higher the value, the better, but it can also be the other way around, depending on the task at hand.

The next step is to determine the weighting factors for ranking the criteria. A weighting factor is a multiplier that determines the importance of how much a particular criterion can affect the final choice [4].

It remains to calculate the convolution value for all alternatives, and then compare: determine for each alternative the sum of the products of all the values of the criteria and their weighting factors. We will conduct an experiment to select the most optimal model for the task at hand, but first we will describe each of the possible options.

Let's consider the practical part of building forecasting and classification models. We will describe experiments for selected algorithms:

– multilayer perceptron;
– classification tree;
– *k*-nearest neighbors method;
– logistic regression.

Let's download two datasets from the website https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. Each of them is a response to a telephone survey on health topics conducted annually by the U.S. Center for Disease Control and Prevention (CDC). This Behavioral Risk Factor Surveillance System (BRFSS) collects responses from more than 400.000 Americans about unhealthy behaviors, chronic diseases, and use of preventive services. The surveys have been conducted since 1984, and the results for 2015 are available on *Kaggle* in a csv file format. The dataset contains 253.680 records [5–7].

The next step is to train the model. To do this, we divide the data for each forecasting method in the following ratio: 80% – training sample, and 20% – test sample.

The perceptron will have three layers. The first one contains the same number of neurons and independent attribute variables. The hidden layer will have two neurons for further processing. The final layer will have one neuron and will produce the final value of the object's probable belonging to one of the possible classes: if it is greater than a certain threshold, the prediction will produce 1, if less than 0.

Suppose we have the following initial conditions for the first data set:

- threshold of values – 0.5;
- number of epochs – 100;
- training sample – 200.000;
- testing sample – 50.000.

The algorithm proceeds as follows:

1) the model receives another object as input;

2) the first layer is filled with the values of its attributes;

3) calculate weighted sums for each of the two layers of the hidden layer;

4) calculate sigmoidal activation values for each of the following neurons;

5) repeat the same operations for the neurons of the hidden layer that are further included in the resulting neuron;

6) adjust the weights.

Repeat the algorithm and calculate the global error using the gradient descent method until it becomes minimal or until we reach the maximum number of training epochs.

Let’s build a classification tree. Consider the algorithm for the first data set. Initially, we have the following conditions:

- training sample – 200.000;
- testing sample – 50.000.

The basis of the algorithm is the *Gini Impurity* of the tree, an indicator calculated as the minimum value of all possible *Gini Impurity* attributes of an object. It reflects the extent to which an attribute can distinguish data by the value of the main dependent feature, and the lower its value, the fewer errors it can cause. If the values of the attribute are not binary, all its possible values are taken into account, sorted, and for each two values, their average number and *Gini Impurity* are calculated.

After selecting the appropriate attribute, the root is built from this key, the left branch containing information with a positive value for a particular attribute is added, and the right branch with a negative value. Then the branch with the larger amount of information is selected as the root.

The algorithm terminates when the tree contains all possible attributes or the branch becomes a leaf, i.e. contains data from one of the possible classes.

Initial conditions:

- training sample – 25.000;
- test sample – 30.000;
- number of nearest neighbors – 3.

This classifier works in the following way: the more neighbors of a certain class, the more likely it is that the object belongs to this class as well. The number *k* here is the number of neighbors that are closer to the next element than others. The distance in this algorithm is calculated using the Euclidean criterion. The method does not require training, so it is immediately applied to the test set items among the training set records. Therefore, it loads the system by calculating the distances from each element of the test set to all its neighbors in the training set on a large scale.

Let’s conduct an experiment for the first dataset using logistic regression, the main purpose of which is to divide objects into two classes.

At the beginning we have:

- number of stages – 100;
- training sample – 200.000;
- testing sample – 50.000.

The algorithm is as follows:

- initialization of synaptic connections and offsets;
- calculation of a linear combination of input features;
- calculating the probable value using a sigmoid function;
- determination of the loss function (log loss);
- adjusting weight relationships and offsets using gradient descent.

The algorithm terminates when the loss function has reached the desired value or the maximum number of training stages has been reached.

The experiment was conducted for the first dataset with the initial conditions specified above. Now let’s calculate the quality of the implemented algorithms based on the calculated metric values.

Let’s simulate and solve the vector optimization problem. We need to choose the best possible model with a high percentage of reliable results. Let’s prepare the decision-making process for choosing the right algorithm [8–9].

1) Let us describe the set of alternatives:

- multilayer perceptron (MLP) [10];
- logistic regression;
- decision tree;
- *k*-nearest neighbors method.

2) Let’s describe the selection criteria:

- *precision* – the number of objects that, according to the algorithm, belong to a positive class, which is a true statement;
- *accuracy* – the number of correctly predicted results;

– *recall* – the number of positively identified items from the entire scope of this class;

– *f*-measure is a quality criterion that combines accuracy and recall.

3) Let's describe the scoring scales by criteria:

– all indicators have values between 0 and 1.

We present the model of the task in the form of a table with known indicators (Table 1).

Table 1. *Output indicators*

	Multilayer perceptron	Classification tree	*k*-nearest neighbors	Logistic regression
Accuracy	84.854	82.8	99.78	85.5575
Precision	1	0.845	0.9974	0.8683
Recall	0.84854	1	1	0.978
F-measure	0.91	0.9162	0.9987	0.92

For an accurate assessment, the indicators should be in the same range, so let's normalize the data: divide the accuracy value by the reference value of 100 and we will have normalized indicators (Table 2).

Table 2. *Table with normalized indicators*

	Multilayer perceptron	Classification tree	*k*-nearest neighbors	Logistic regression
Accuracy	0.84854	0.828	0.9978	0.855575
Precision	1	0.845	0.9974	0.8683
Recall	0.84854	1	1	0.978
F-measure	0.91	0.9162	0.9987	0.92

At this stage, the networks are not compared according to the Pareto principle, so we will perform linear additive convolution with normalizing factors for the entire problem model and consider the results of the study (Table 3).

As we can see, under the condition of linear additive convolution, the k-nearest neighbors algorithm is better at predicting whether objects belong to a certain class. However, we should note that the study was not conducted for the entire sample, but for a part of the information due to the long operation of the method. This alternative is suitable for a small amount of data, so logistic regression is a better option for processing a significant amount of information. If we repeat the calculations for the second dataset (Table 4), the result will be the same.

Table 3. *Linear additive convolution*

	Normalization multiplier	Classification tree	*k*-nearest neighbors	Logistic regression	Classification tree
Accuracy	0.2832929	0.84854	0.828	0.9978	0.855575
Precision	0.2694909	1	0.845	0.9974	0.8683
Recall	0.2613327	0.84854	1	1	0.978
F-measure	0.2670298	0.91	0.9162	0.9987	0.92
		0.97462471	0.96827181	1.079475332	0.977628

Table 4. *Indicators of the second dataset study*

	Multilayer perceptron	Classification tree	*k*-nearest neighbors	Logistic regression
Accuracy	86.482	82.8	99.68	86.925
Precision	0.86482	0.8419	0.9967	0.8796
Recall	1	1	0.9995	0.9833
F-measure	0.9275	0.9141	0.9833	0.9286

Thus, as the study confirmed, regression analysis is the best method for predicting the development of diabetes among people.

Conclusions

Machine learning models are the key to new opportunities in various fields. In particular, this applies to medicine, where machine learning is used to diagnose and predict the occurrence of possible diseases [11–14].

To date, many studies have been conducted on an important problem of humanity – identifying the cause of disease development. Various models are used for this purpose: neural networks, classifiers, decision trees, etc.

This paper investigates the most common prediction and classification models used in the medical field, namely the multilayer perceptron, decision tree, *k*-nearest neighbors method, and logistic regression. Each of them was analyzed on the basis of the *accuracy*, *precision*, *recall*, and *f*-measure criteria. Linear additive convolution was used to determine the best of these methods – *k*-nearest neighbors. However, given that this model functions slowly when working with a large amount of information, logistic regression was found to be the best model for predicting the development of diabetes.

The main advantages of the chosen model in medicine are:

– the ability to search for relationships in very complex situations when they are difficult to notice when assessing the situation;

– due to the ability to learn, the model can find solutions to problems even in the absence of a priori knowledge of the initial information, the development of the phenomenon under study, the dependence between parameters, input indicators and expected results;

– the accuracy of forecasts does not depend on the availability of different types of less informative or missing data.

However, despite the effectiveness of machine learning models, they have several drawbacks:

– training takes some time, the neural network has to go through retraining stages during repeated use, and a large amount of input information requires more time;

– the reliability of the results could be better [15].

Logistic regression meets all the requirements set out in this paper, but its efficiency needs to be improved. Therefore, further research will be aimed at optimizing the chosen model. As of today, it has drawbacks that need to be eliminated.

Logistic regression uses first-order optimization methods, namely the gradient descent method. Its essence lies in the fact that synaptic weights change iteratively in the direct or opposite direction of the target gradient function. The update of values reflecting how the attributes of the elements affect the dependent feature continues until the most optimal results are achieved. The speed of this process, namely the number of iterations, depends on the value of the training parameter.

The gradient descent method is easy to implement when developing models in machine learning, but it has several drawbacks.

The first one is that in the case of a large amount of information, the algorithm is complicated by long redundant calculations. This is when stochastic gradient descent (SGD) comes in handy. This method uses a randomly selected sample to adjust the gradient during each iteration without calculating its exact value, which is what the method estimates. The number of data records does not affect the performance of the algorithm, and it is capable of achieving sublinear convergence speed. Therefore, stochastic gradient descent spends less time updating model parameters and does not accept large-scale calculations.

The second problem is to determine the optimal model training parameter, or more precisely, the hyperparameter. The process of model creation, speed, and accuracy of results depend on the hyperparameter. In choosing the best value that can be used in machine learning, various methods are used that return a tuple of hyperparameters and losses, namely:

– lattice search;

– random search;

– Bayesian optimization;

– optimization based on gradients;

– evolutionary optimization;

– population-based optimization.

The study used the method of tuning hyperparameters by random search. Therefore, in the future, it can be removed from the options for modifying and optimizing the implemented algorithm.

Thus, the purpose of further research is to optimize the model, eliminate shortcomings in its functioning in order to accurately determine the possibility of developing diseases.

References

1. Sharonova, N., Kyrychenko, I., Tereshchenko, G. (2021), "Application of big data methods in E-learning systems", *Computational Linguistics and Intelligent Systems (COLINS 2021): 5th International Conference, Lviv, 22–23 April 2021: CEUR workshop proceedings*, No. 2870, P. 1302–1311.

2. Smelyakov K., Hurova Y., Osiievskyi S. (2023), "Analysis of the Effectiveness of Using Machine Learning Algorithms to Make Hiring Decisions", *Computational Linguistics and Intelligent Systems (COLINS 2023): 7th International Conference, Kharkiv, 20–21 April 2023: CEUR workshop proceedings*, No. 3387, P. 77–92.

3. Kyrychenko I., Nazarov O., Huliiev N., Avdieiev O. (2023), "Selection of Artificial Neural Networks for Disease Prediction", *Computational Linguistics and Intelligent Systems (COLINS 2023): 7th International Conference, Kharkiv, 20–21 April 2023: CEUR workshop proceedings*, No. 3387, P. 236–248.

4. Haglin, J. M., Jimenez, G., Eltorai A. (2019), "Artificial neural networks in medicine", *Health and Technology*, No. 9, P. 1–6. DOI: 10.1007/s12553-018-0244-4.

5. Gaur, L., Bhatia, U., Jhanjhi, N. Z., Muhammad, G. (2023), "Medical image-based detection of COVID-19 using Deep Convolution Neural Networks", *Multimedia Systems*, No. 29, P. 1729–1738. DOI: 10.1007/s00530-021-00794-6

6. IHME (2022), *11 global health issues to watch in 2023, according to IHME experts*, available at: https://www.healthdata.org/acting-data/11-global-health-issues-watch-2023-according-ihme-experts (last accessed 18.05.2023).

7. Nuha, A., et al. (2022), "Introduction and Methodology: Standards of Care in Diabetes–2023", *Diabets Care*, No. 46 (1), P. 1–4. DOI: 10.2337/dc23-Sint

8. Khan, G., Siddiqi, A., Ghani Khan, M. U., Qayyum Wahla, S., Samyan, S. (2019), "Geometric positions and optical flowbased emotion detection using MLP and reduced dimensions", *IET Image Processing*, No. 13 (4), P. 634–643. DOI: 10.1049/iet-ipr.2018.5728

9. Verma, S., Razzaque, M. A., Sangtongdee, U., Arpnikanondt, C., Tassaneetrithep, B., Hossain, A. (2021), "Digital Diagnosis of Hand, Foot, and Mouth Disease Using Hybrid Deep Neural Networks", *IEEE Access*, No. 9, P. 143481–143494. DOI: 10.1109/ACCESS.2021.3120199

10. Rimi, T. A., Sultana, N., Ahmed Foysal, M. F. (2020), "Derm-NN: Skin Diseases Detection Using Convolutional Neural Network", *4th International Conference on Intelligent Computing and Control Systems* (ICICCS), Madurai, India, P. 1205–1209. DOI: 10.1109/ICICCS48265.2020.9120925

11. Sarvamangala, D. R., Kulkarni, R. V. (2022), "Convolutional neural networks in medical image understanding: a survey", *Evolutionary Intelligence*, No. 15, P. 1–22. DOI: 10.1007/s12065-020-00540-3

12. Liu, Y., Jain, A., Eng, C. et al. (2020), "A deep learning system for differential diagnosis of skin diseases", *Nature Medicine*, No. 26 (6), P. 900–908. DOI: 10.1038/s41591-020-0842-3

13. NIDDKD (2022), *Diabetes Prevention Program (DPP)*, available at: https://www.niddk.nih.gov/about-niddk/research-areas/diabetes/diabetes-prevention-program-dpp (last accessed 18.05.2023).

14. Tison, G. H., Zhang, J., Delling, F. N., Deo, R. C. (2020), "Automated and Interpretable Patient ECG Profiles for Disease Detection, Tracking, and Discovery". *Circulation: Cardiovascular Quality and Outcomes*. No. 12 (9). DOI: 10.1161/circoutcomes.118.005289

15. Smelyakov, K., A., Chupryna, Bohomolov, O., Ruban, I. (2020), "The Neural Network Technologies Effectiveness for Face Detection", *3rs International International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, P. 201–205. DOI: 10.1109/DSMP47368.2020.9204049

Received 15.06.2023

Відомості про авторів / About the Authors

Гулієв Нурал Бахадур огли – Харківський національний університет радіоелектроніки, магістр зі спеціальності "Інженерія програмного забезпечення", Харків, Україна; e-mail: nural.huliiev@nure.ua; ORCID ID: https://orcid.org/0000-0003-2123-0377

Перетяга Максим Юрійович – Харківський національний університет радіоелектроніки, магістр зі спеціальності "Інженерія програмного забезпечення", Харків, Україна; e-mail: maksym.peretiaha@nure.ua; ORCID ID: https://orcid.org/0000-0002-9675-1305

Ховрат Артем Вячеславович – Харківський національний університет радіоелектроніки, магістр зі спеціальності "Інженерія програмного забезпечення", Харків, Україна; e-mail: artem.khovrat@gmail.com; ORCID ID: https://orcid.org/0000-0002-1753-8929

Тесленко Денис Максимович – Харківський національний університет радіоелектроніки, магістр зі спеціальності "Інженерія програмного забезпечення", Харків, Україна; e-mail: denys.teslenko@nure.ua; ORCID ID: https://orcid.org/0000-0002-6289-1633

Назаров Олексій Сергійович – кандидат технічних наук, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії, заступник декана факультету "Комп’ютерні науки", Харків, Україна; e-mail: oleksii.nazarov1@nure.ua; ORCID ID: https://orcid.org/0000-0001-8682-5000

Huliiev Nural – Kharkiv National University of Radio Electronics, M. Sc. in Software Engineering, Kharkiv, Ukraine.

Peretiaha Maksym – Kharkiv National University of Radio Electronics, M. Sc. in Software Engineering, Kharkiv, Ukraine.

Khovrat Artem – Kharkiv National University of Radio Electronics, M. Sc. in Software Engineering, Kharkiv, Ukraine.

Teslenko Denys – Kharkiv National University of Radio Electronics, M. Sc. in Software Engineering, Kharkiv, Ukraine.

Nazarov Alexei – PhD (Engineering Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Software Engineering, Kharkiv, Ukraine.

ДОСЛІДЖЕННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ТА КЛАСИФІКАЦІЇ В ЗАДАЧАХ НАЯВНОСТІ ДІАБЕТУ СЕРЕД ПАЦІЄНТІВ З ІНСУЛЬТОМ У РІЗНИХ УМОВАХ ЖИТТЄДІЯЛЬНОСТІ

Предметом дослідження є методи прогнозування розвитку цукрового діабету. Цукровий діабет – неінфекційне захворювання, що вразило 425 млн людей, а до 2045 р. їхня кількість збільшиться в півтора раза. Доведено, що це захворювання є незалежним фактором, що спричиняє розвиток інсульту. Коли в крові занадто багато цукру, він негативно впливає на артерії та судини. Пацієнти з діабетом більш схильні до утворення атеросклеротичних бляшок і тромбів, що може призвести до блокади серця та ішемічного інсульту. Наявність діабету збільшує ризик інсульту й погіршує його перебіг. За результатами Фремінгемського дослідження, кількість повторних випадків захворювання серця подвоюється. **Мета дослідження** – вивчити методи прогнозування та класифікації розвитку діабету серед людей, зокрема пацієнтів з інсультом, для запобігання іншим захворюванням. Складність проблеми полягає в тому, що недіагностованих випадків стільки ж, скільки й діагностованих, тому близько половини людей страждають від хвороби й спричинених ускладнень через неналежне або запізніле діагностування. Тому важлива вчасна діагностика захворювання, яке важко виявити, з метою запобігання розвитку подальших ускладнень. **Завданням статті** є вибір найкращого алгоритму прогнозування виникнення захворювання. У роботі використано такі **методи**: багатошаровий персептрон, метод *k*-найближчих сусідів, дерево рішень і логістична регресія. На сьогодні для вирішення подібних проблем широко застосовується машинне навчання. Упродовж 1950–1960-х рр. були спроби об’єднати наявні на той час підходи до створення нейронних мереж, що дало змогу обчислювати кількісні описи людського інтелекту, а також запам’ятовувати, аналізувати та обробляти інформацію, що нагадувало роботу людського мозку. Медицина – одна з основних галузей, де різноманітні класифікатори та нейромережні алгоритми з кожним роком набувають все більшої популярності. Вони є пріоритетними, зокрема, і в діагностиці захворювань. **Результати**: з’ясовано, що початковим умовам вибору найкращої моделі відповідає логістична регресія. **Висновки**: унаслідок дослідження обрано оптимальну модель для прогнозування розвитку захворювання.

Ключові слова: багатошаровий персептрон; нейронна мережа; прогнозування; інсульт; цукровий діабет.

Бібліографічні описи / Bibliographic descriptions

Гулієв Н. Б. огли, Перетяга М. Ю., Ховрат А. В., Тесленко Д. М., Назаров О. С. Дослідження моделей прогнозування та класифікації в задачах наявності діабету серед пацієнтів з інсультом у різних умовах життєдіяльності. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 2 (24). С. 54–61. DOI: https://doi.org/10.30837/ITSSI.2023.24.054

Huliiev, N., Peretiaha, M., Khovrat, A., Teslenko, D., Nazarov, A. (2023), "Study of prediction and classification models in the problems of diabetes among patients with a stroke in different conditions of living conditions", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (24), P. 54–61. DOI: https://doi.org/10.30837/ITSSI.2023.24.054