УДК 004.89

EDN KRDEXZ

Анализ алгоритмов искусственного интеллекта для прогнозирования гастроэнтерологических заболеваний

Яхшибоев Рустам Эркинбой угли, Муминов Баходир Болтаевич, Хусанов Уролбой Абдуманнон угли, Кудратиллаев Мейрбек Бахитбай угли*

Ташкентский университет информационных технологии имени Мухаммада Аль-Хорезми, пр. Амира Темура, 108, Ташкент, 100000, Узбекистан

*E-mail: m.qudratillayev@tuit.uz

Аннотация. В данной статье рассматривается анализ алгоритмов искусственного интеллекта для прогнозирования гастроэнтерологических заболеваний. Данные и параметры каждого пациента были получены гастроэнтерологами в клинике Ташкентской медицинской академии на кафедре гастроэнтерологии. С помощью этих алгоритмов и программного обеспечения можно сократить время на диагностику пациента, при этом повышается точность диагностики, а экономичное приложение будет удобным и может быть реализовано в различных клиниках Республики Узбекистан. В связи с Указом Президента Республики Узбекистан Ш. Мирзиеева “О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта” сферы медицины и фармацевтики направлены на использование технологий искусственного интеллекта для анализа и прогнозирования потребностей рынка. Приложение № 6 к Указу Президента Республики Узбекистан от № ПП - 4996 указывает на высшие учебные заведения Ташкентской медицинской академии, в связи с этим постановлением совместно были проведены анализ и разработка программного обеспечения с использованием алгоритмов искусственного интеллекта. В статье представлен анализ нескольких алгоритмов искусственного интеллекта. С помощью этих алгоритмов планируется в дальнейшем разработать программно-аппаратный комплекс для прогнозирования гастроэнтерологических заболеваний и внедрить его в медицинских учреждениях.

Ключевые слова: искусственный интеллект, алгоритм, прогнозирование, гастроэнтерологические заболевания, аппаратно-программный комплекс.

Analysis of artificial intelligence algorithms for predicting gastroenterological diseases

Yakhshiboyev Rustam Erkinboy o‘g‘li, Muminov Bahodir Boltayevich, Husanov O‘rolboy Abdumannon o‘g‘li, Kudratillaev Meirbek Baxitbay o‘g‘li*

Tashkent University of Information Technology named after Muhammad al-Khwarizmi, Tashkent, 100000, Uzbekistan

*E-mail: m.qudratillayev@tuit.uz

Abstract. This article discusses the analysis of artificial intelligence algorithms for predicting gastroenterological diseases. The data and parameters of each patient were obtained respectively with the help of gastroenterologists at the clinic of the Tashkent Medical Academy in the Department of Gastroenterology. With the help of these algorithms and software, it is possible to reduce the time for diagnosing a patient, the accuracy of diagnostics increases and the economic acquisition will be convenient and can be implemented in different clinics of the Republic of Uzbekistan. In connection with the Decree of the President of the Republic of Uzbekistan Sh. Mirziyoyev “On measures to create conditions for the accelerated introduction of artificial intelligence technologies”, the spheres of medicine and pharmaceuticals indicate the use of artificial intelligence technologies for analyzing and forecasting market needs. Appendix No. 6 to the Decree of the President of the Republic of Uzbekistan dated No. PP - 4996 indicates the higher educational institutions of the Tashkent Medical Academy, in connection with this decree, the analysis and development of software using artificial intelligence algorithms were jointly carried out. An analysis of several artificial intelligence algorithms has been made. With the help of these algorithms, it is planned to further develop a hardware-software complex for predicting gastroenterological diseases and implement it in medical institutions.

Keywords: artificial intelligence, algorithm, forecasting, gastroenterological diseases, hardware and software complex.

1. Introduction

At present, the development of artificial intelligence in all countries of the world is developing carefully and rapidly. Links in the development of artificial intelligence The President of the Republic of Uzbekistan Sh. Mirziyoyev issued a resolution “On measures to create conditions for the accelerated introduction of artificial intelligence technologies”, this resolution is in accordance with the strategy “Digital Uzbekistan - 2030”. [1,2,3]

In the field of medicine, digital technologies can be widely used in the diagnosis, treatment of various diseases and different degrees. With the help of digital technologies, the work of doctors can be facilitated, the human factor is reduced, research time is reduced and efficiency is increased.

Within a short time, the doctor can make a decision about the diagnosis. With the help of digital technologies, controversial points can be overcome. Digital technologies use artificial intelligence, neural networks, machine learning and modern Python programming languages.

Artificial intelligence- the science and technology of creating intelligent machines, especially intelligent computer programs. AI is related to the similar task of using computers to understand human intelligence, but is not necessarily limited to biologically plausible methods.

Preliminary diagnostics helps to find out the problems, determine the bottlenecks of the enterprise, draw up a program for future changes and should answer the questions: is it possible to solve the identified problems, in what sequence they need to be solved. [4,5,6]

2. Main part

The K-Nearest Neighbors (KNN) algorithm is a type of supervised ML algorithm that can be used for both classification and regression prediction problems.

The K-nearest neighbors algorithm works like "feature similarities", especially used to predict the values of new data points that a new data point will be assigned a value by how closely it matches the points in the training set.

KNN falls into supervised learning algorithms. This means that there is a dataset labeled with training dimensions (x, y), and find the relationship between x and y. The goal is to discover the function h:X→Y so that given an unknown observation x, h(x) can positively predict the identical output y (1,2).

For distance metrics, the Euclidean metric will be used:

$$d(x, x') = \sqrt{(x_1 + x_1')}^2 + \cdots + \sqrt{(x_n + x_n')}^2 \text{(one)}$$

input x is assigned to the class with the highest probability.

$$P_{(y=j|X=x)} = \frac{1}{K}\sum_{i\in A} I(y^i = j)(2)$$

For regression, the method will be the same, instead of neighbor classes, we will take the target value and find the target value for the invisible data point by taking the mean, mean, or any suitable function. [7,8,9]

Random forest algorithm("random forest") is a machine learning algorithm proposed by Leo Breiman and Adele Cutler, which consists in using a committee (ensemble) of decision trees.

Random forest fetches rows and columns with a decision tree as the basis. Models h1, h2, h3, h4 differ from each other more than when using only bags due to the selection of columns.

As the number of base learners (k) increases, the variance will decrease. When you decrease k, the variance increases. But the offset remains constant for the entire process. k can be found using cross validation. [10,11,12]

2.1. Implementation in Scikit-learn

For each decision tree, Scikit-learn calculates the node importance using the Gini importance, assuming only two child nodes (binary tree):

$$n_j^i = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

- ni sub(j)= importance of node j
- w sub (j) = weighted number of samples reaching node j
- C sub(j) = impurity value of node j
- left(j) = child node on left, split at node j
- right(j) = child node from right split at node j

Then the importance of each object in the decision tree is calculated as:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k\in all\ nodes} ni_k}$$

- fi sub(i)= importance of feature i

- ni sub(j)= importance of node j

They can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

The final importance of a feature at the random forest level is the average of all trees. The sum of the object importance values for each tree is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

- RFfi sub(i) = the importance of the feature I calculated from all the trees in the random forest model
- normfi sub(ij) = normalized feature importance for i in tree j
- T = total number of trees

2.2. Implementation in Spark

For each decision tree, Spark calculates feature importance by summing the gain scaled by the number of samples passing through the node:

$$fi_i = \sum_{j;nodes\ j\ splits\ on\ features\ i} s_j C_j$$

- fi sub (i) = importance of feature i
- s sub(j) = number of samples reaching node j
- C sub(j) = impurity value of node j

To calculate the final feature importance at the random forest level, first the feature importance for each tree is normalized with respect to the tree:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

- normfi sub(i) = normalized importance of object i
- fi sub (i) = importance of feature i

Then the importance values of objects from each tree are summarized and normalized:

$$RFfi_i = \frac{\sum_j norm\ fi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}}$$

- RFfi sub(i) = the importance of the feature I calculated from all the trees in the random forest model
- normfi sub(ij) = normalized feature importance for i in tree j

2.3. The purpose of the algorithm involved in SVM

In other words, "The goal is to maximize the minimum distance." for the distance is given:

$$d_{H(\varphi(x_0))} = \frac{|w^T(\varphi(x_0)) + b|}{||w||_2}$$

$$w^* = arg_w \max\ [min_n d_H(\varphi(x_n))]$$

So, now that the goal is clear. By making predictions for the training data, which was binary, classified into positive and negative groups, if a point is replaced from the positive group in the hyperplane equation, we will get a value greater than 0 (zero), mathematically,

$$w\ T\ (\varphi\ (x)) + b > 0$$

And predictions from the negative group in the hyperplane equation would give a negative value as

$$w\ T(\Phi(x)) + b < 0.$$

But here the signs were about the training data, that is, how we train our model. This is for a positive class, give a positive sign, and for a negative class, give a negative sign.

But when testing this model on the test data, if we correctly predict a positive class (positive sign or sign greater than zero) as positive, then two positive results yield a positive and therefore greater than zero result. The same applies if we correctly predict the negative group, since two negatives will again result in a positive result.

But if the model error classifies the positive group as negative, then one plus and one minus constitute a minus, hence less than zero overall [19,20,21].

2.4. To summarize the above concept

The product of the predicted and the actual label will be greater than 0 (zero) if the prediction is correct, otherwise less than zero.

$$y_n[w^T \varphi(x) + b] = \begin{cases} \geq 0 \; if \; correct \\ < 0 \; if \; incorrect \end{cases}$$

For ideally separable datasets, the optimal hyper-plane classifies all points correctly, additionally substituting the optimal values into the weight equation.

Argmax is an abbreviation for maxima arguments, which are basically points in the area of the function in which the values of the function maximize.

Besides, taking the independent weight term outward gives:

$$w^* arg_w max \frac{1}{||w||_2} [min_n y_n |w^T (\varphi(x) + b)|]$$

The inner term (min n_{yn} | w T Φ (x) + b |) basically represents the minimum distance from the point to the decision boundary and the closest point to the decision boundary H.

Rescaling the distance to the nearest point as 1, i.e. (min n_{yn} | w T Φ (x) + b |) = 1. Here the vectors remain in the same direction, and the hyperplane equation does not change. It's like zooming in on an image; objects expand or contract, but the directions stay the same and the image stays the same [16,17,18].

Distance rescaling is done by replacing:

$$w \rightarrow cw, \qquad b \rightarrow cb$$

$$(cw)^T \varphi(x_n) + (cb) = c(w^T \varphi(x_n) + b) = 0$$

Now the equation becomes (describing that each point is at least 1/| | w || 2 away from the hyperplane) as

$$w^* = arg_w max \frac{1}{||w||_2}, s.t. min_n y_n [w^T \varphi(x_n) + b] = 1$$

This maximization problem is equivalent to the following minimization problem, which is multiplied by a constant, since they do not affect the results [13,14,15].

In a scientific study, human saliva was taken. With the help of saliva, you can predict about gastroenterological diseases. With illness, the composition of saliva changes

dramatically. The composition of saliva is parameters. By changing the composition of saliva, you can create a data set for training artificial intelligence algorithms. Table No. 1 shows the composition of a healthy person.

Table 1. The composition of the saliva of a healthy person.

№	The composition of saliva	Qty. (% and g/l)
1	Water	99.4-99.5%
2	Organic and inorganic components	0.5-0.6%
3	Squirrels	1.4-6.4 g/l
4	Mucin	0.8-6.0 g/l
5	cholesterol	0.02-0.5 g/l
6	Glucose	0.1-0.3 g/l
7	Ammonium	0.01-0.12 g/l
8	Uric acid	0.005-0.03 g/l

In the article “Analysis of algorithms for predicting and preliminary diagnosis of gastroenterological diseases” [21], the initial results were taken, which contain 200 patients. As a continuation of this study, the number of patients was increased to 1000 per data set.

Patient parameters were obtained. Based on the obtained parameters, the data set was trained using the KNN, ANN, SVM and Random Forest algorithms.

3. Results

An example of the date of the set is indicated in table No. 2. The parameters and the name of the composition of human saliva are indicated.

An analysis of the KNN, ANN, SVM and Random Forest algorithms was made, the number of patients was 100 and 1000. The corresponding results were obtained.

Table 2. Data set parameters.

Data set parameters	The name of the composition of saliva
Parameter_1	Squirrels
Parameter_2	Mucin

Parameter_3	cholesterol
Parameter_4	Glucose
Parameter_5	Ammonium
Parameter_6	Uric acid

The first time the training process was carried out on the number of 100 patients in the data set (figures 1,2).

	Patient	Parameter_1	Parameter_2	Parameter_3	Parameter_4	Parameter_5	Parameter_6
0	1	1.4	0.8	0.02	0.10	0.01	0.005
1	1	1.5	0.9	0.03	0.11	0.02	0.006
2	1	1.6	1.0	0.04	0.12	0.03	0.007
3	1	1.7	1.1	0.05	0.13	0.04	0.008
4	1	1.8	1.2	0.06	0.14	0.05	0.009
...	...	...	...	...	...	...	...
94	3	10.9	10.2	0.97	1.04	0.95	0.099
95	3	11.0	10.3	0.98	1.05	0.96	0.100
96	3	11.1	10.4	0.99	1.06	0.97	0.101
97	3	11.2	10.5	1.00	1.07	0.98	0.102
98	3	11.3	10.6	1.01	1.08	0.99	0.103

99 rows × 7 columns

Figure 1. Date set of 100 patients.

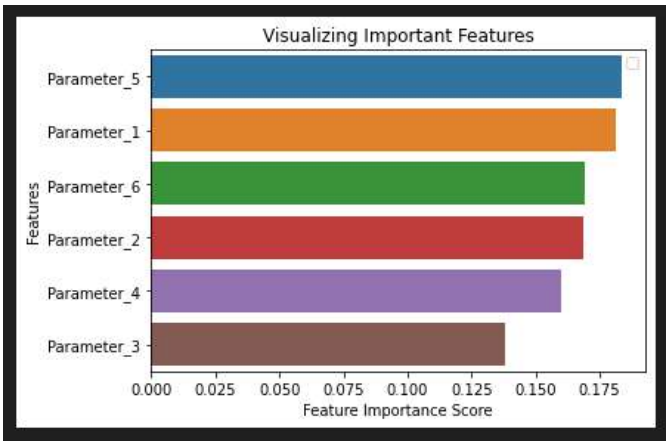

Figure 2. Importance of parameters from the set date (100).

The selected KNN, ANN, SVM and Random Forest algorithm determined the importance of the parameters using the correlation coefficient and predicted the probability of illness from the set date. (figures 3,4,5,6)

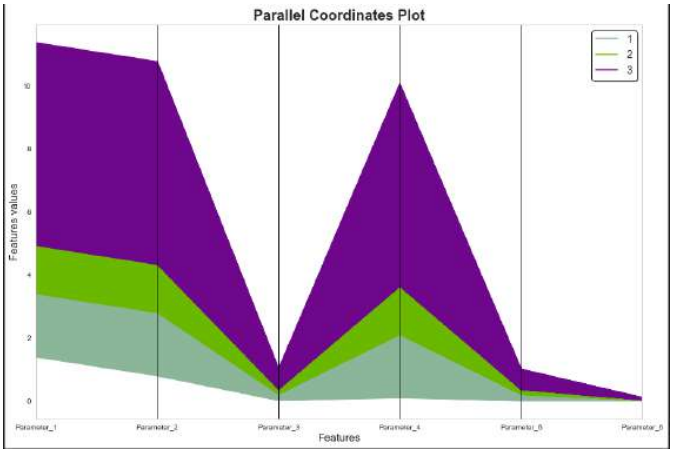

Figure 3. Determining the importance of parameters and predicting the probability of a patient's disease (KNN).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio – healthy

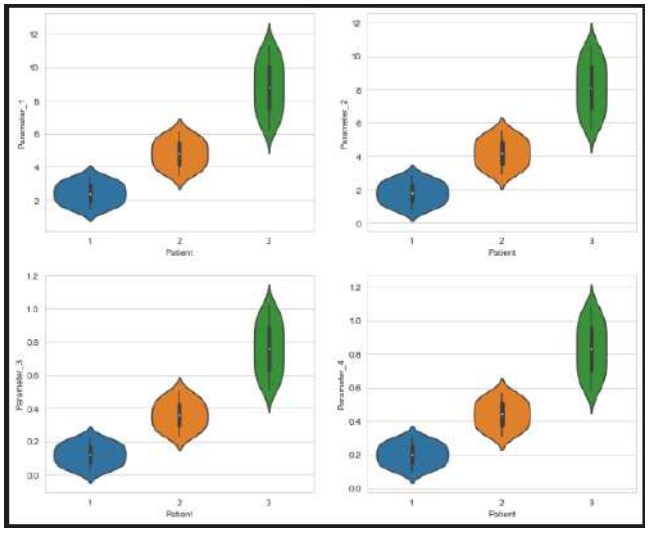

Figure 4. Determining the importance of parameters and predicting the likelihood of a patient's illness (ANN).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio – healthy

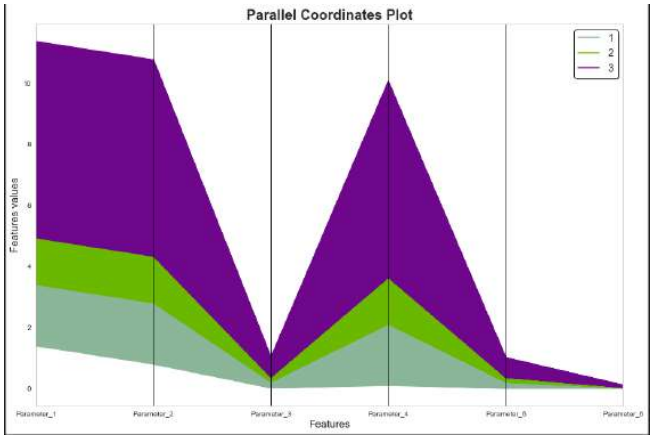

Figure 5. Determining the importance of parameters and predicting the likelihood of a patient's illness (SVM).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio – healthy

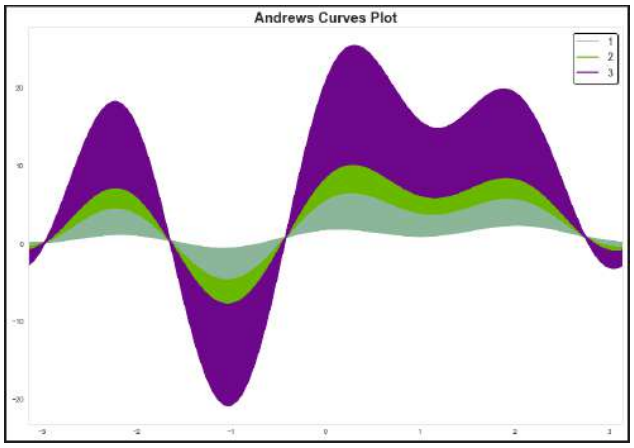

Figure 6. Determining the importance of parameters and predicting the likelihood of a patient's illness (Random Forest).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio - healthy

The second time the training process was carried out on the number of 1000 patients in the data set. The selected KNN, ANN, SVM and Random Forest algorithm determined the importance of the parameters and predicted the probability of illness from the set date (figure 7).

	Patient	Parameter_1	Parameter_2	Parameter_3	Parameter_4	Parameter_5	Parameter_6
0	1	1.4	0.8	0.02	0.10	0.01	0.005
1	1	1.5	0.9	0.03	0.11	0.02	0.006
2	1	1.6	1.0	0.04	0.12	0.03	0.007
3	1	1.7	1.1	0.05	0.13	0.04	0.008
4	1	1.8	1.2	0.06	0.14	0.05	0.009
...	...	...	...	...	...	...	...
94	3	10.9	10.2	0.97	1.04	0.95	0.099
95	3	11.0	10.3	0.98	1.05	0.96	0.100
96	3	11.1	10.4	0.99	1.06	0.97	0.101
97	3	11.2	10.5	1.00	1.07	0.98	0.102
98	3	11.3	10.6	1.01	1.08	0.99	0.103

99 rows × 7 columns

Figure 7. Date set of 1000 patients.

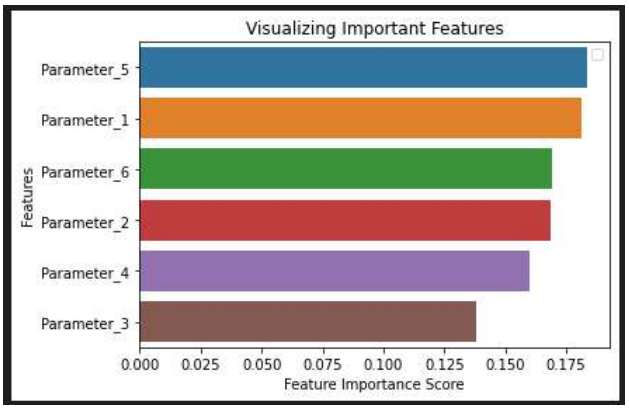

Figure 8. Importance of parameters from the set date (1000).

The selected KNN, ANN, SVM and Random Forest algorithm determined the importance of the parameters using the correlation coefficient and predicted the probability of illness from the set date (figures 9,10,11,12).

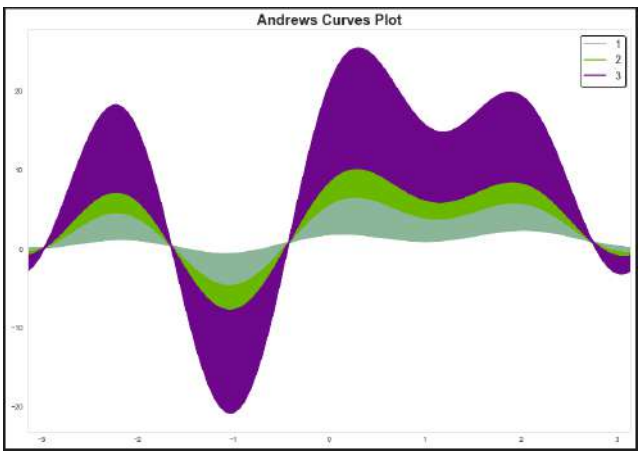

Figure 9. Determining the importance of parameters and predicting the likelihood of a patient's illness (Random Forest).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio – healthy

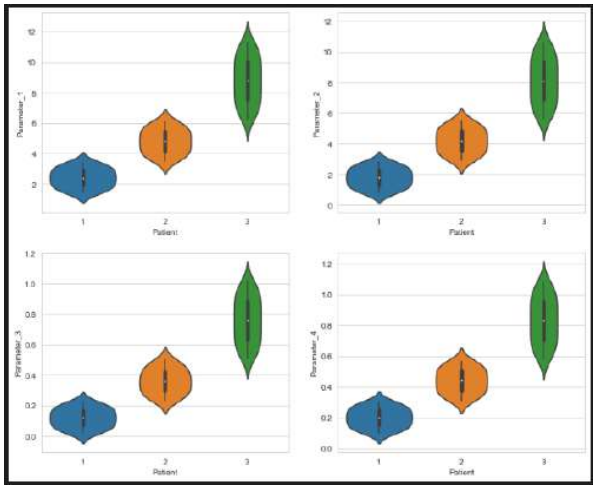

Figure 10. Determining the importance of parameters and predicting the patient's probability of illness (KNN).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio – healthy

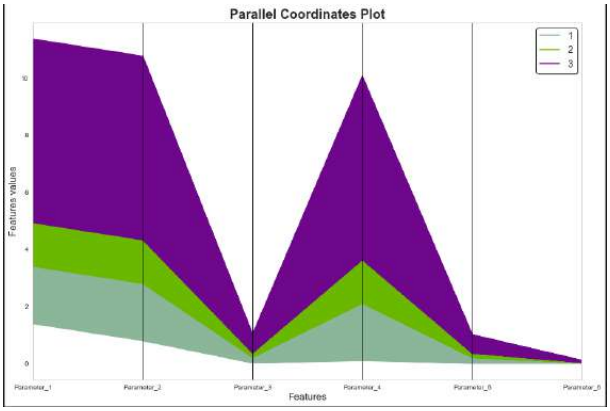

Figure 11. Determining the importance of parameters and predicting the likelihood of a patient's illness (ANN).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio – healthy

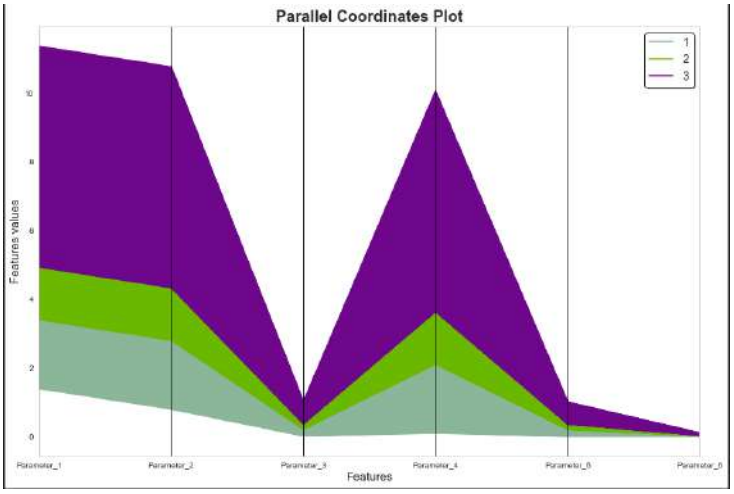

Figure 12. Determining the importance of parameters and predicting the likelihood of a patient's illness (SVM).

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio - healthy

In this algorithm, the result appeared on three colors. Accordingly, the result of training can be determined by colors:

- Purple - the probability of illness is higher
- Green - the probability of illness is lower
- pistachio – healthy

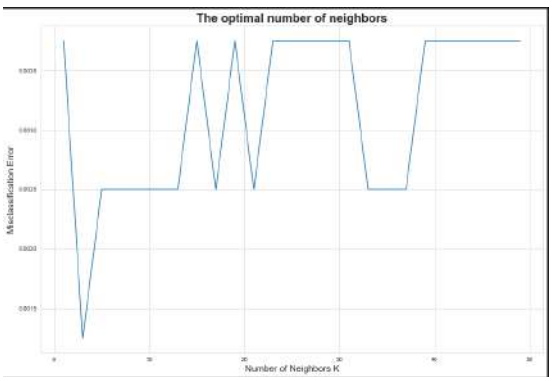

Figure 13. Choose the optimal number of algorithms.

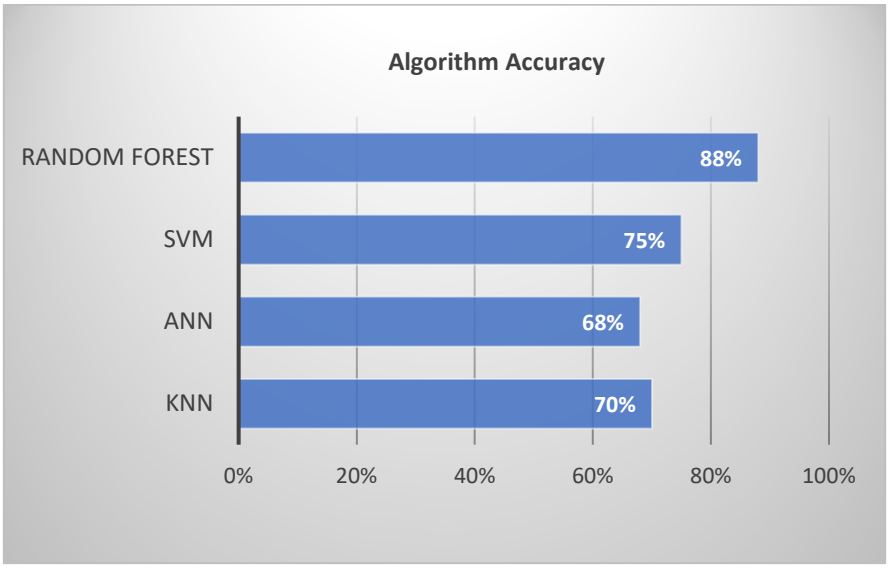

Figure 14. Accuracy of algorithms.

4. Conclusion

As a result of scientific research, the corresponding training accuracy of the KNN, ANN, SVM and Random Forest algorithms was obtained. For each algorithm, two different date sets were used, i.e. the number of patients is from 100 to 1000.

Acknowledgements

The work was supported by the clinic of the Tashkent Medical Academy and the Department of Biomedical Engineering, Informatics and Biophysics.

References

1. On measures to create conditions for the accelerated introduction of artificial intelligence technologies http://lex.uz//docs/5297051
2. Balashova, A. Fakes and robots: what will be the main technological trends in 2019.
3. Balashova, A., Posypkina, A., Balenko, E. RBC. – 2018.
4. Christopher Bishop. Pattern Recognition and Machine Learning, 2006
5. Stenneth, Leon, Philip, S.Yu. Monitoring and mining GPS traces in transit space, SIAM International Conference on Data Mining.
6. Ganesh, J., Gupta, M., Varma, V. Interpretation of Semantic Tweet Representations. arXivpreprint arXiv:1704.00898. – 2017.

7. Zhang, A. Characterizing Online Discussion Using CoarseDiscourse Sequences / A. Zhang, B. Culbertson, P. Paritosh // Proceedings of the International AAAI Conference on Web andSocial Media. – 2017.
8. Hastie, T., Tibshirani R., Friedman J. Chapter 15. Random Forests / T. Hastie, R. Tibshirani, J. Friedman // The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – Springer-Verlag, 2009. – 746 p.
9. Stalkamp, J. man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition / J. Stalkamp et al. // Neural networks. – 2012. – T. 32. – P. 323-332.
10. Masci, J. Stacked convolutional auto-encoders for hierarchical feature extraction / J. Masci et al. // Artificial Neural Networks and Machine Learning-ICANN 2011. – Springer Berlin Heidelberg, 2011. – P. 52-59.
11. Krizhevsky, A. Imagenet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, Hinton G.E. //Advances in neural information processing systems. – 2012. – P. 1097-1105.
12. A nanoelectronics-blood-based diagnostic biomarker for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). https://www.pnas.org/doi/full/10.1073/pnas.1901274116
13. Яхшибоев, Р.Э. Разработка программного средства для идентификации номерных знаков транспортных средств на основе методов компьютерного зрения / Р.Э. Яхшибоев, Т.Д. Очилов, Б.Н. Сиддиков // Journal of new century innovations. – 2022. – Т. 15. – №. 1. – С. 81-93.
14. Yaxshiboyev, R.E. Forecasting groundwater evaporation using multiple linear regression / R. E. Yaxshiboyev et al. // Galaxy International Interdisciplinary Research Journal. – 2021. – T. 9. – №. 12. – P. 1101-1107.
15. Yakhshibaev, R. Development of a mathematical model for balancing the level and device for remote monitoring of groundwater parameters / R. Yakhshibaev et al. // 2021 International Conference on Information Science and Communications Technologies (ICISCT). – IEEE, 2021. – P. 1-4.
16. Djumanov, J. Mathematical model and software package for calculating the balance of information flow / J. Djumanov et al. // 2021 International Conference on Information Science and Communications Technologies (ICISCT). – IEEE, 2021. – P. 1-6.

17. Gafurjonovich, M.V. Technologies of Organization of Practical Lessons in the Modern Education System / M.V. Gafurjonovich, E.E. Yaxshivayevich, Y.D. Erkinbayevna // European Multidisciplinary Journal of Modern Science. – 2022. – T. 4. – P. 332-336.
18. Khamzaev, J. Driver sleepiness detection using convolution neural network / J. Khamzaev [et al.] // Central asian journal of education and computer sciences (CAJECS). – 2022. – T. 1. – №. 4. – P. 31-35.
19. Yaxshiboyev, R. Development of a software and hardware complex for primary diagnostics based on deep machine learning / R. Yaxshiboyev // Central asian journal of education and computer sciences (CAJECS). – 2022. – T. 1. – №. 4. – P. 20-24.
20. Yaxshiboyev, R. Development of a model of object recognition in images based on the «transfer learning» method / R. Yaxshiboyev // Central asian journal of education and computer sciences (CAJECS). – 2022. – T. 1. – №. 4. – С. 36-41.
21. Yaxshiboyev, Rustam. Analysis of algorithms for prediction and preliminary diagnostics of gastroenterological diseases / Rustam Yaxshiboyev, Dilbar Yaxshiboyeva // Central asian journal of education and computer sciences (CAJECS) 1.2. – (2022). – P. 49-56.