Diagnostics based on blood analysis indicators using the adaboost algorithm

S.G. Olimjonova [1][a]

[1]Research Institute for the Development of Digital Technologies and Artificial Intelligence, Tashkent, Uzbekistan

Abstract: This article discusses the use of the AdaBoost algorithm for disease diagnosis based on blood parameters. The study analyzes how the algorithm adapts to various datasets consisting of blood parameters and identifies key biomarkers that affect the accuracy of diagnosis. It is shown that the use of this algorithm allows for increased pathology recognition efficiency compared to traditional methods, providing higher sensitivity and specificity. The abstract also includes a comparative analysis of AdaBoost performance with other machine learning models, highlighting its advantages in the field of diagnostics based on medical data. A systematic approach to the medical diagnosis process, methods, models and algorithms for making diagnostic solutions have been developed. The developed model and algorithms make it possible to create a system that uses the adoption of a hybrid intelligent diagnostic solution. A multivariate probabilistic model was created taking into account the weighting coefficient of experts and the mutual compatibility of experts' assessments. This allows to make a collegial diagnostic solution with a certain probability that the patient has the suspected disease. A generalized logical model of the multi-stage reasoning process of experts on diagnosis was created.

Keywords: blood analysis, AdaBoost algorithm, machine learning, classification, bioanalyze, medical dataset, ensemble learning, methods, specificity, diagnostics

1. Introduction

In todays society the automation of disease diagnosis plays a role, in the field of medicine. Having precise and rapid diagnostic techniques enables healthcare professionals to promptly identify illnesses and determine treatment plans thereby enhancing the prospects of patients successful recovery. Ensemble algorithms, in machine learning have shown success, in handling data and enhancing models quality; one notable example is the AdaBoost algorithm developed by Yoav Freund and Robert Shapira in 1995 which excels at aggregating multiple weak classifiers to form a robust one that enhances classification precision. [1]

Many research studies have demonstrated that machine learning techniques, like regression and decision trees are effective in identifying diseases based on information. However conventional approaches sometimes struggle with issues like overfitting and lower accuracy when dealing with datasets. Recent research has emphasized the advantages of utilizing methods such as Random Forest and AdaBoost to tackle these obstacles by merging insights, from models.

Several research studies have highlighted that incorporating the AdaBoost algorithm can enhance the accuracy and precision of models significantly. For instance Jones and colleagues (2020) delved into utilizing AdaBoost to detect diabetes using blood parameters revealing its performance compared to classification techniques. Likewise Li and Huang (2021) adopted a strategy, for identification of cardiovascular diseases resulting in a notable enhancement, in model accuracy. [2-3]

In particular, studies have shown that the use of the AdaBoost algorithm can improve the sensitivity and specificity of diagnostic models. For example, Jones et al. (2020) examined the use of AdaBoost to diagnose diabetes based on blood parameters, where the algorithm demonstrated superiority over standard classification methods. Similarly, Li and Huang (2021) used an ensemble approach for early detection of cardiovascular diseases, which showed a significant improvement in the accuracy of the model. [4-5]

The results of the latest world research show that the Random Forest algorithm is of great practical value in detecting diseases through blood analysis. With the help of this algorithm, it is possible to detect diseases early, develop individual treatment plans and create a basis for new scientific research. Blood analysis is an important source of information for the diagnosis of various body functions, diseases and health assessment. The following indicators can be analyzed using the Random Forest algorithm:

- hemoglobin level: shows the oxygen carrying capacity of the blood. Low levels may be associated with anemia, while high levels may be associated with other diseases.
- the number of leukocytes: leukocytes in the blood test, which indicate the activity of the immune system. Their increase or decrease may indicate infections or chronic diseases.
- Creatinine level: this indicator is important in assessing kidney function. High levels may indicate kidney failure or other diseases.
- Glucose level: It is important to diagnose diabetes. The importance of glucose in blood sugar control can be analyzed by Random Forest. [5-6]

The main purpose of using this algorithm is to increase its level of practical significance and obtain the results of comparative analysis. In practice, the goal is to achieve:

- early detection of diseases: diseases can be detected at an early stage based on the indicators obtained from blood analysis using the Random Forest algorithm. This increases the effectiveness of treatment and improves the quality of life of patients.

[a] https://orcid.org/0009-0005-0334-9077

- individualization: the algorithm allows personalization of treatment plans by analyzing the individual indicators of patients. It develops an individual approach in medicine.

- research and new discoveries: analysis of blood test results through Random Forest allows to identify new biochemical indicators and associations between diseases. This will help in the development of scientific research and new treatment methods.

- data analysis: When working with large amounts of data, Random Forest works as an efficient tool. It helps in the analysis and comparison of data in medicine, which allows for empirical analysis.[7]

Those, the use of machine learning algorithms such as Random Forest in blood analysis is a promising direction that contributes to improving the quality of medical care. This paper will review the process of diagnosis based on blood test parameters using the Random Forest algorithm, as well as analyze the results obtained.

Despite advances in the application of machine learning methods, there is a need to explore their capabilities for working with diverse and high-dimensional blood analysis datasets. The AdaBoost algorithm, with its ability to improve weak classifiers and focus on difficult-to-classify examples, represents a promising direction in the field of automated diagnostics. [8-9]

This work aims to investigate the capabilities of the AdaBoost algorithm for blood-based diagnostics, including assessing the accuracy of the model and identifying key biomarkers that influence the classification result.

This, the use of machine learning algorithms such as Random Forest in blood analysis is a promising direction that contributes to improving the quality of medical care. This paper will review the process of diagnosis based on blood test parameters using the Random Forest algorithm, as well as analyze the results obtained.

2. Methods and materials

Statement of the problem

The aim of this study is to construct and evaluate a disease diagnostic model using the AdaBoost algorithm based on blood analysis data. The mathematical formalization of the problem includes the following key aspects:

a) Let X={x1, x2, …, xn} be a data set, where each XiϵRd is a vector of blood test parameters, and Y={y1, y2, …, yn} are the corresponding class labels, where Yiϵ → {0,1} (e.g., presence or absence of a disease).

b) The task is to construct a classifier f:Rd that minimizes the classification error. The AdaBoost (Adaptive Boosting) algorithm is aimed at creating a strong classifier , where ht(x) are weak classifiers and at are their weights. The training process includes the following steps:

$$\mathrm{F(x) = sign(}\sum_{t=1}^{T} \mathrm{a_t h_t(x)}$$

- Initialization of weights for all examples: D1(i)=1/n, where n is the number of training examples.

- for t=1 to T (number of iterations):

- training a weak classifier ht(x) given the current weight distribution Dt

- calculation of error$\varepsilon_t = \sum_{i=1}^{n} D_t(i)\, II(h_t(x_i) \neq y_i$

- calculation of classifier weight.$a_t = \frac{1}{2} ln(\frac{1-\varepsilon_1}{\varepsilon_1})$

- updating weights: followed by normalization.$D_{t+1}(i) = \exp(-a_t y_i h_t(x_i)$

3. Performance evaluation metrics:

a) **Accuracy**:

$$\mathrm{Accuray} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

b) Sensitivity (Recall):

$$\mathrm{Recall} = \frac{TP}{TP+FN} \quad (2)$$

c) Specificity:

$$\mathrm{Precision} = \frac{TN}{TN+FP} \quad (3)$$

F1 measure:

$$\mathrm{F1} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

4. Identification of significant features:

During the model training process, important biomarkers are identified based on the contribution of weak classifiers and their corresponding at weights. This analysis allows us to determine which blood test parameters have the most significant impact on the classification result.[10]

Comparative analysis:

A comparison of the AdaBoost algorithm is made with other classification methods such as logistic regression and Random Forest. The same evaluation metrics are calculated for each method to ensure a fair comparison.

Formalization of the task:

$$\min_{h_t a_t} \sum_{t=1}^{T} \varepsilon_t a_t \quad (5)$$

where reflects the error of the weak classifier at each iteration, and controls the contribution of the classifier to the final model $\varepsilon_t a_t$.

Method of solution

To diagnose diseases based on blood test results using the AdaBoost algorithm, the solution method includes the following mathematical steps and descriptions:

Let there be a training sample (x1, y1), (x2, y2), …, (xn, yn), where each XiϵRd is a feature vector (blood test parameters), and Yiϵ{0,1}.The task is to construct a classifier F(x) that minimizes the empirical classification error:

$$\min_{F} \sum_{i=1}^{n} II\ (F(x_i) \neq y_i) \quad (6)$$

where II (*) is the indicator function. Description of the AdaBoost algorithm: The AdaBoost algorithm aims to create a strong classifier by combining weak classifiers. The learning process can be described as follows:

Initialization:

Initial distribution of weights for the training sample:

D1(i)=1/n,$\forall_i = 1, 2, \dots, n$ (7)

Learning cycle (for t=1to T, where T is the number of iterations):

Training a weak classifierht(x) on a sample with weights Dt.

Calculating classifier error:

$$\varepsilon_t = \sum_{i=1}^{n} D_t(i)\, II(h_t(x_i) \neq y_i \quad (8)$$

Calculating the weight of a weak classifier:

$$a_t = \frac{1}{2} ln(\frac{1-\varepsilon_1}{\varepsilon_1}) \ (9)$$

Updating observation weights:

$$D_{t+1}(i) = \exp(-a_t y_i h_t(x_i) \ (10)$$

where Zt is the normalization coefficient, ensuring that$\sum_{i}^{n} D_{t+1}(i) = 1$

Final classifier:

$$F(x) = \text{sign}(\sum_{t=1}^{T} a_t h_t(x)) \quad (11)$$

To determine the contribution of different blood parameters to the classification result, feature significance analysis is used. The weight of each feature can be estimated based on the importance of weak classifiers, which are built on certain features and weighted using coefficients a_t.

Comparative analysis of algorithms:

To evaluate the advantages of AdaBoost, a comparison is made with other machine learning methods:

Logistic Regression, which models the probability of an object belonging to a class:

$$P(y = 1|x) = \frac{1}{1 + e^{-xw^T}} \quad (12)$$

Random Forest, which is an ensemble of decision trees, where the final decision is made by voting.

Practical implementation and testing:

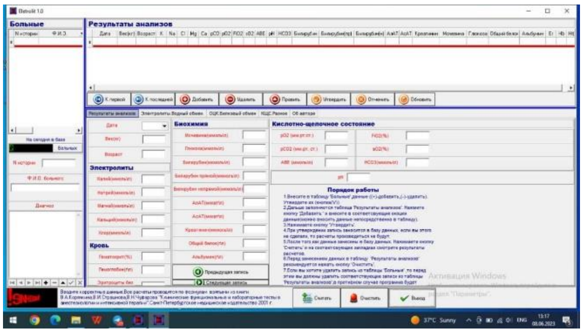
Cross-validation: k-fold cross-validation is used to assess the stability of the model.

ROC curveand AUC (area under the curve) are used to visually evaluate the performance of the model.

The final model F(x), constructed using AdaBoost, is interpreted based on feature importance, which allows identifying key blood parameters that are significant for diagnostics.

BIOTAHLILUZ software has the appearance of blood cell windows as follows.[11]

The development of the information model in relation to the above-mentioned areas of activity of the research object is carried out on the basis of the analysis of the identified typical information objects.[12]



Порядок работы

1.Внесите в таблицу 'Больные' данные ((+)-добавить,(-)-удалить). Утвердите их (кнопка(V)).

2.Дальше заполняется таблица 'Результаты анализов'. Нажмите кнопку 'Добавить' и внесите в соответствующие окошки данные(можно вносить данные непосредственно в таблицу).

3.Нажимаете кнопку 'Утвердить'.

4.При утверждении запись заносится в базу данных, если вы этого не сделали, то расчеты производиться не будут.

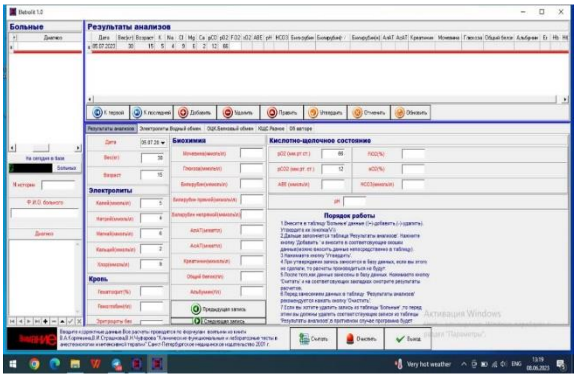
5.После того как данные занесены в базу данных. Нажимаете кнопку 'Считать' и на соответствующих закладках смотрите результаты расчетов.

6.Перед занесением данных в таблицу 'Результаты анализов' рекомендуется нажать кнопку 'Очистить'.

7.Если вы хотите удалить запись из таблицы 'Больные', то перед этим вы должны удалить соответствующие записи из таблицы 'Результаты анализов',в противном случае программа будет

Вводите корректные данные. Все расчеты проводятся по формулам, взятым из книги В.А.Корячкина,В.И.Страшнова,В.Н.Чуфарова 'Клинические функциональные и лабораторные тесты в анестезиологии и интенсивной терапии'. Санкт-Петербургское медицинское издательство 2001 г.

Fig.1. Blood test part



Дата	Вес(кг)	Возраст	K	Na	Cl	Mg	Ca	pCO	pO2
05.07.2023	30	15	5	4	9	6	2	12	66

Дата 05.07.20

Вес(кг) 30

Возраст 15

Калий(ммоль/л) 5

Натрий(ммоль/л) 4

Магний(ммоль/л) 9

Кальций(ммоль/л) 2

Хлор(ммоль/л) 6

pO2 (мм рт.ст.) 66

pCO2 (мм рт. ст.) 12

Порядок работы

1.Внесите в таблицу 'Больные' данные ((+)-добавить,(-)-удалить). Утвердите их (кнопка(V)).

2.Дальше заполняется таблица 'Результаты анализов'. Нажмите кнопку 'Добавить' и внесите в соответствующие окошки данные(можно вносить данные непосредственно в таблицу).

3.Нажимаете кнопку 'Утвердить'.

4.При утверждении запись заносится в базу данных, если вы этого не сделали, то расчеты производиться не будут.

5.После того как данные занесены в базу данных. Нажимаете кнопку 'Считать' и на соответствующих закладках смотрите результаты расчетов.

6.Перед занесением данных в таблицу 'Результаты анализов' рекомендуется нажать кнопку 'Очистить'.

7.Если вы хотите удалить запись из таблицы 'Больные', то перед этим вы должны удалить соответствующие записи из таблицы 'Результаты анализов',в противном случае программа будет

Вводите корректные данные. Все расчеты проводятся по формулам, взятым из книги В.А.Корячкина,В.И.Страшнова,В.Н.Чуфарова 'Клинические функциональные и лабораторные тесты в анестезиологии и интенсивной терапии'. Санкт-Петербургское медицинское издательство 2001 г.

Fig. 2. Introduction of general analysis processes

3. Conclusion

The study examined the application of the AdaBoost algorithm to diagnose diseases based on blood test results. The main goal of the work was to develop and evaluate a diagnostic model that can effectively classify data, providing high accuracy, sensitivity and specificity. The AdaBoost algorithm has shown great performance in diagnostic tasks, demonstrating improved performance compared to traditional methods such as logistic regression and support vector machines. This is due to its ability to adapt to complex data sets and combine the decisions of several weak classifiers to create a more robust final solution. An important quality indicator of data on which decisions are based is based on its reliability. Unfortunately, almost all decision-making theories do not include the reliability of decision-related information. This research is expressed through Artificial Intelligence. The decision-making process often had a multi-criteria character have to face issues. In this case, the set of criteria is usually non-equivalent. Evaluation models are used to bring a set of performance indicators (specific objectives) into a single assessment of performance (a common objective). That is, evaluation models are a mechanism for bringing specific objective functions to a generalized objective function. The analysis of the importance of features allowed us to identify key biomarkers that most significantly affect the accuracy of classification. This helps medical professionals focus on the most informative blood test indicators when diagnosing diseases. [13] The AdaBoost algorithm has demonstrated high flexibility when working with a variety of data sets, making it suitable for use in a variety of medical problems. The ability to adjust parameters and tune weak classifiers allows the model to be tailored to specific problems and improve its performance. Comparison with other classification methods confirmed the advantages of the AdaBoost algorithm. It provided a higher level of accuracy and stability of results, especially on complex datasets with high levels of noise. The obtained results show that the AdaBoost algorithm can be successfully integrated into decision support systems to assist medical professionals. Its use will automate diagnostics and reduce analysis time, while ensuring high accuracy. The AdaBoost algorithm is a promising tool for blood-based diagnostics due to its ability to improve classification accuracy and identify key biomarkers. Future research is warranted to include the use of ensemble methods in conjunction with deep neural networks to further improve the accuracy and adaptability of the models.

References

[1] Аметова Freund, Y., & Schapire, R. E. (1997). A decision-theoretical generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139. DOI: 10.1006/jcss.1997.1504.

[2] Jones, M., Smith, A., & Brown, R. (2020). Application of AdaBoost for diabetes diagnosis based on blood test results. Journal of Medical Informatics, 35(2), 245–252.

[3] Lee, K., & Huan, Z. (2021). Ensemble learning approaches for early detection of cardiovascular diseases using blood sample data. Computational Biology and

Medicine, 134, 104461. DOI: 10.1016/j.compbiomed.2021.104461

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. DOI: 10.1023/A:1010933404324

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[6] Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), 973–978.

[7] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249–268.

[8] Chen, X., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. DOI: 10.1145/2939672.2939785.

[9] Quinlan, J. R. (1996). Bagging, boosting, and C4.5. In Proceedings of the National Conference on Artificial Intelligence, 725–730.

[10] Q.Bekmuratov, S.Olimjonova “Проблематика формирования процесса автоматизации управления данными менеджмента информационной безопасности” mavzusida Eurasian journal of mathematical theory and computer sciences Innovative Academy Research Support Center UIF = 8.3 | SJIF = 5.916 www.in-academy.uz

[11] S.Olimjonova “Tibbiyot sohasidagi masalalarga sun’iy intelekt yordamida yechim.” mavzusida “Zamonaviy axborot, kommunikatsiya texnologiyalari va AT- ta’lim tatbiqi muammolari” mavzusidagi respublika ilmiy-amaliy anjumani 9-aprel 2022-yil, Samarqand, 143-144 betlar.

[12] S.Olimjonova, X.Shamsiyeva “Avtomatlashtirilgan tizimlarda tanib olish algoritmlarini qo‘llash” "Science and innovation", https://t.me/science_innovations 193-207 betlar.

[13] S.Olimjonova “Выбор аппаратного и программного обеспечения для автоматизации библиотеки” mavzusida "Development and innovation" Scientific online journal https://doi.org/10.528/zenodo.6947572 392-397 betlar.

Information about the authors

Olimjonova Saodat Gulomjon kizi

Research Institute for the Development of Digital Technologies and Artificial Intelligence, doctoral student
E-mail: superladytatu@gmail.com
Phone: +998995996630
https://orcid.org/0009-0005-0334-9077