

METHODS OF DIAGNOSTICS AND TECHNOLOGIES

Scientific Article

UDC 616-006.66

DOI: 10.17816/pmj423130-143

POSSIBILITY OF PREDICTING THE PROBABILITY OF THYROID CANCER RECURRENCE BY MACHINE LEARNING METHODS

M.A. Barulina¹, I.Yu. Bendik¹, I.I. Kovalenko¹, M.A. Polidanov^{2}, R.P. Petrunkin²,
V.N. Kudashkin³, K.A. Volkov⁴, A.R. Kravchenya⁴, V.V. Maslyakov^{4,5}, S.V. Kapralov⁴,
G.E. Aslanov⁴, Ye.V. Losyakova³, I.S. Obukhov³, A.D. Osina⁴, A.K. Kurmaeva⁴*

¹Perm State National Research University,

²University "Reaviz", Saint Petersburg,

³Samara State Medical University

⁴Saratov State Medical University named after V.I. Razumovsky

⁵Medical University "Reaviz", Saratov, Russian Federation

ВОЗМОЖНОСТЬ ПРЕДСКАЗАНИЯ ВЕРОЯТНОСТИ РЕЦИДИВА РАКА ЩИТОВИДНОЙ ЖЕЛЕЗЫ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

© Barulina M.A., Bendik I.Yu., Kovalenko I.I., Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Volkov K.A., Kravchenya A.R., Maslyakov V.V., Kapralov S.V., Aslanov G.E., Losyakova Ye.V., Obukhov I.S., Osina A.D., Kurmaeva A.K., 2025

e-mail: maksim.polidanoff@yandex.ru

[Barulina M.A. – DSc (Physics and Mathematics), Director of the Institute of Physics and Mathematics, ORCID: 0000-0003-3867-648X; Bendik I.Yu. – 1st-year Master's Student of the Institute of Physics and Mathematics, ORCID: 0009-0000-7851-9492; Kovalenko I.I. – Head of the Center for Artificial Intelligence of the Institute of Physics and Mathematics, ORCID: 0000-0003-4450-1184; Polidanov M.A. (*contact person) – Advisor to the Russian Academy of Natural Sciences (RANS), Research Department Specialist, Assistant of the Department of Biomedical Disciplines, ORCID: 0000-0001-7538-7412; Petrunkin R.P. – 3rd-year Student of the Faculty of Medicine, ORCID: 0009-0003-3206-7920; Kudashkin V.N. – Resident of the Department of Surgery with a Course in Cardiovascular Surgery of the Institute of Professional Education, ORCID: 0000-0001-9099-3517; Volkov K.A. – 3rd-year Student of the Institute of Clinical Medicine, ORCID: 0000-0002-3803-2644; Kravchenya A.R. – PhD (Medicine), Associate Professor, Associate Professor of the Department of Pediatric Diseases of the Faculty of Medicine, ORCID: 0000-0003-2738-4510; Maslyakov V.V. – DSc (Medicine), Professor, Professor of the Department of Mobilization Preparation of Healthcare and Disaster Medicine, Professor of the Department of Surgical Diseases, ORCID: 0000-0001-6652-9140; Kapralov S.V. – DSc (Medicine), Associate Professor, Head of the Department of Faculty Surgery and Oncology, ORCID: 0000-0001-5859-7928; Aslanov G.E. – 6th-year Student of the Institute of Clinical Medicine, ORCID: 0009-0009-9497-5725; Losyakova Ye.V. – 6th-year Student of the Institute of Pediatrics, ORCID: 0009-0003-8286-4266; Obukhov I.S. – 6th-year Student of the Institute of Pediatrics, ORCID: 0009-0007-5573-8431; Osina A.D. – 6th-year Student of the Institute of Clinical Medicine, ORCID: 0009-0001-5294-3436; Kurmaeva A.K. – 6th-year Student of the Institute of Clinical Medicine, ORCID: 0009-0002-0886-6290].

М.А. Барулина¹, И.Ю. Бендик¹, И.И. Коваленко¹, М.А. Полиданов^{2*}, Р.П. Петрунькин², В.Н. Кудашкин³, К.А. Волков⁴, А.Р. Кравченя⁴, В.В. Масляков^{4,5}, С.В. Капралов⁴, Г.Э. Асланов⁴, Е.В. Лосякова³, И.С. Обухов³, А.Д. Осина⁴, А.К. Курмаева⁴

¹Пермский государственный национальный исследовательский университет,

²Университет «Реавиз», г. Санкт-Петербург,

³Самарский государственный медицинский университет,

⁴Саратовский государственный медицинский университет имени В.И. Разумовского,

⁵Медицинский университет «Реавиз», г. Саратов, Российская Федерация

Objective. To develop a machine learning model for predicting the fact of recurrence in patients with thyroid cancer after surgical intervention.

Materials and methods. According to the aim of the study, the case histories of 300 patients who had undergone surgical intervention for thyroid cancer were analyzed. The average age was 43.54 years. All patients included in the study underwent a comprehensive examination according to the clinical recommendations on the diagnosis and treatment of patients with thyroid cancer. Selection of the most appropriate model in machine learning is critical as it directly affects the accuracy and efficiency of prediction. Selection of the best model was done through comparing the performance of different algorithms on the same training sample using cross-validation. Each model was evaluated on such metrics as average accuracy and standard deviation to determine which model demonstrates the best results. The random forest model performed best in terms of average accuracy and was used hereafter. The model was trained using a matrix of predefined features. Using param grid, we can efficiently adjust hyperparameters such as the number of trees, maximum depth and minimum number of samples for separation, which will help us to find the optimal settings for our task. RandomizedSearchCV method was used to select the hyperparameters. During the hyperparameter search process, the model was trained on training data selected as 70 % of the original dataset. The search resulted in the follow-

© Барулина М.А., Бендик И.Ю., Коваленко И.И., Полиданов М.А., Петрунькин Р.П., Кудашкин В.Н., Волков К.А., Кравченя А.Р., Масляков В.В., Капралов С.В., Асланов Г.Э., Лосякова Е.В., Обухов И.С., Осина А.Д., Курмаева А.К., 2025
e-mail: maksim.polidanoff@yandex.ru

[Барулина М.А. – доктор физико-математических наук, директор Физико-математического института; ORCID: 0000-0003-3867-648X; Бендик И.Ю. – магистр I курса Физико-математического института; ORCID: 0009-0000-7851-9492; Коваленко И.И. – заведующий Центром искусственного интеллекта Физико-математического института; ORCID: 0000-0003-4450-1184; Полиданов М.А. (*контактное лицо) – советник Российской академии естественных наук (РАЕ), специалист научно-исследовательского отдела, ассистент кафедры медико-биологических дисциплин, ORCID: 0000-0001-7538-7412; Петрунькин Р.П. – студент III курса лечебного факультета, ORCID: 0009-0003-3206-7920; Кудашкин В.Н. – врач-ординатор кафедры хирургии с курсом сердечно-сосудистой хирургии Института профессионального образования, ORCID: 0000-0001-9099-3517; Волков К.А. – студент III курса Института клинической медицины, ORCID: 0000-0002-3803-2644; Кравченя А.Р. – кандидат медицинских наук, доцент, доцент кафедры детских болезней лечебного факультета, ORCID: 0000-0003-2738-4510; Масляков В.В. – доктор медицинских наук, профессор, профессор кафедры мобилизационной подготовки здравоохранения и медицины катастроф, профессор кафедры хирургических болезней, ORCID: 0000-0001-6652-9140; Капралов С.В. – доктор медицинских наук, доцент, заведующий кафедрой факультетской хирургии и онкологии, ORCID: 0000-0001-5859-7928; Асланов Г.Э. – студент VI курса Института клинической медицины, ORCID: 0009-0009-9497-5725; Лосякова Е.В. – студент VI курса Института педиатрии, ORCID: 0009-0003-8286-4266; Обухов И.С. – студент VI курса Института педиатрии, ORCID: 0009-0007-5573-8431; Осина А.Д. – студентка VI курса Института клинической медицины, ORCID: 0009-0001-5294-3436; Курмаева А.К. – студентка VI курса Института клинической медицины, ORCID: 0009-0002-0886-6290].

ing best hyperparameters for the random forest model for our data specifically: `n_estimators = 161`; `min_samples_split = 5`; `max_leaf_nodes = 39`; `max_depth = 12`; `bootstrap = True`.

Results. A model that demonstrated high target feature accuracy was trained during the study. The proportion of patients with postoperative recurrence correctly identified by the model was 98 % of all patients with recurrence, and the proportion of patients without recurrence correctly classified by the model “as patients at no risk of recurrence” was 95 % of all patients without recurrence. This shows that the developed model effectively handles the task of classification based on medical parameters, which may be particularly important for decision making in clinical practice. The high accuracy indicates the reliability of the model and its ability to identify cases of recurrence correctly, this may contribute to the improvement of diagnostics and treatment.

Conclusions. A machine learning model to predict a high probability of thyroid cancer recurrence based on the analysis of medical parameters was developed while carrying out the study. The development process began with careful data preprocessing, which is a critical step in reliable models’ construction. During preprocessing, outliers and columns containing monotonic values were removed to improve the data quality and avoid distortions in the model training. Categorical variables were also coded to ensure that they could be used correctly in machine learning algorithms, and correlated features were excluded to minimize multicollinearity and increase the interpretability of the model.

Keywords. Thyroid cancer, thyroid, machine learning, recurrence prediction, random forest, python.

Цель. Разработка модели машинного обучения по предсказанию факта рецидива у пациентов с раком щитовидной железы после проведенного оперативного вмешательства.

Материалы и методы. В соответствии с целью исследования были проанализированы истории болезни 300 пациентов с выполненным оперативным вмешательством по поводу рака щитовидной железы. Средний возраст – 43,54 года. Всем включенным в исследование больным было проведено комплексное обследование согласно клиническим рекомендациям по диагностике и лечению больных РЩЖ. Выбор наиболее подходящей модели в машинном обучении критически важен, так как он напрямую влияет на точность и эффективность предсказания. Отбор лучшей модели был произведен через сравнение производительности различных алгоритмов на одной и той же обучающей выборке с использованием кросс-валидации. Каждая модель оценивалась по метрикам, таким как средняя точность и стандартное отклонение, что позволяет определить, какая из них демонстрирует наилучшие результаты. Лучше всего по показателю средней точности выявила себя модель случайного леса, она же в дальнейшем и использовалась. Обучение модели было произведено по матрице заранее определенных признаков. Используя параметрическую сетку (`param_grid`), можно эффективно настраивать гиперпараметры, такие как количество деревьев, максимальная глубина и минимальное количество образцов для разделения, что поможет найти оптимальные настройки для нашей задачи. Для подбора гиперпараметров использовался метод `RandomizedSearchCV`. В процессе поиска гиперпараметров модель обучалась на тренировочных данных, отобранных как 70 % от исходного датасета. Итогом поиска определились следующие лучшие гиперпараметры для модели случайного леса для конкретно наших данных: `n_estimators = 161`; `min_samples_split = 5`; `max_leaf_nodes = 39`; `max_depth = 12`; `bootstrap = True`.

Результаты. В ходе исследования была обучена модель, которая продемонстрировала высокую точность целевого признака. Доля пациентов с послеоперационным рецидивом, правильно идентифицированных моделью, составила 98 % от общего числа пациентов с рецидивом, а доля пациентов без рецидива, верно классифицированных моделью «как пациенты, не имеющие риска рецидива», – 95 % от всех пациентов без рецидива. Это свидетельствует, что разработанная модель эффективно справляется с задачей классификации на основе медицинских параметров, что может быть особенно важно для принятия решений в клинической практике. Высокая точность указывает на надежность модели и ее способность правильно идентифицировать случаи рецидива, что может способствовать улучшению диагностики и лечения.

Выводы. В рамках исследования была разработана модель машинного обучения для предсказания высокой вероятности рецидива рака щитовидной железы на основе анализа медицинских параметров. Процесс разработки начался с тщательной предобработки данных, что является критически важным этапом в построении надежных моделей. В ходе предобработки были удалены выбросы и столбцы, содержащие однообразные значения, что позволило улучшить качество данных и избежать искажений в обучении модели. Также была проведена кодировка категориальных переменных, что обеспечило возможность их корректного использования в алгоритмах машинного обучения, и исключены коррелирующие признаки, чтобы минимизировать мультиколлинеарность и повысить интерпретируемость модели.

Ключевые слова. Рак щитовидной железы, щитовидная железа, машинное обучение, прогнозирование рецидивов, случайный лес, python.

INTRODUCTION

Thyroid cancer (TC) is one of the most common types of cancer among endocrine diseases [1–4]. Despite high survival rates with early detection and adequate treatment, the problem of recurrence remains relevant and requires special attention [5–7]. Recurrence of the disease can occur even after successful treatment, which makes it necessary to do patients' monitoring regularly [8–11]. However, predicting recurrence based on clinical parameters is a challenging task for medicine in general.

Oncology specialists face several challenges when assessing the risk of disease recurrence. First, patients' clinical data can be varied and complex, including age, gender, tumor grade, presence of metastases, and previous test results. These factors can interact with each other in complex ways, making them difficult to interpret. Secondly, traditional risk assessment methods are often based on subjective assessments by physicians, which can lead to variability in diagnoses and treatment recommendations.

Furthermore, the time spent analyzing data and making decisions can be significant. With limited resources and increasing workloads on healthcare professionals, it's important to optimize the process of diagnosing and monitoring patients. Incorrect assessment of the risk of recurrence can lead not only to a deterioration in the

patient's condition, but also to unnecessary costs for additional examinations and treatment.

The above-described challenges necessitate the development of an automated model for predicting thyroid cancer recurrence. Machine learning methods enable the analysis of large volumes of data and the identification of hidden patterns between various clinical indicators and the likelihood of recurrence. The model can be trained on historical patient data, allowing it to make more accurate predictions based on new input data.

Developing such a model will not only improve prediction accuracy but also significantly reduce the time required for data analysis, allowing physicians to focus on more important aspects of patient care and improving the quality of care. Furthermore, automating the recurrence risk assessment process could lead to cost savings for both healthcare providers and patients.

Thus, the development of a model for predicting thyroid cancer recurrence represents an important step towards improving the diagnosis and treatment of patients.

The aim of the study is to develop a machine learning model to predict the incidence of recurrence in patients with thyroid cancer after surgery.

MATERIALS AND METHODS

In accordance with the aim of the study, the medical records of 300 patients who had

undergone surgery for thyroid cancer were analyzed. The average age was 43.54 years. All patients included in the study underwent a comprehensive examination in accordance with clinical guidelines [12] for the diagnosis and treatment of patients with thyroid cancer. Patients meeting the following inclusion criteria based on the complex of examination results were selected: patients with thyroid cancer without confirmed metastases with the disease stage from T1N0M0 to T3N0M0; absence of previous and concomitant special treatment (immunotherapy or targeted therapy); availability of informed consent for the surgical intervention and participation in the study.

Selection of the most appropriate model in machine learning is critical as it directly impacts the accuracy and efficiency of predictions¹. The right model allows for better pattern detection in data and adaptability to the specifics of the task. An inappropriate model can lead to poor performance, overfitting, or underfitting, making it difficult to interpret results and make decisions.

Several models of the most common ones were considered:

- Logistic Regression (LR);
- Linear Discriminant Analysis (LDA);
- K-Nearest Neighbors (KNN);
- Classification and Regression Trees (CART);
- Gaussian Naive Bayes (NB);
- Support Vector Machines (SVM);
- Random Forest Classifier (RF).

The best model was selected by comparing the performance of different algorithms on the same training set using cross-validation. Each model was evaluated using metrics such as average accuracy and standard deviation, which made it possible to determine which one demonstrated the best results (Table 1).

The random forest model showed the best average accuracy, and was used subsequently.

The model was trained using a matrix of predefined features, as this allows for a systematic study of the impact of various parameters on model performance. Using a parametric grid (param_grid), hyperparameters² can be adjusted efficiently², such as the number of trees, maximum depth, and minimum number of samples to split, which will help us find the optimal settings for our task.

The RandomizedSearchCV method was used to select hyperparameters. Its unique feature is that instead of testing all possible combinations of these hyperparameters RandomizedSearchCV randomly selects a given number of combinations, which allows for faster finding of optimal settings and is especially useful when there are a large number of hyperparameters or their values, as it helps to avoid excessive costs of resources and time for training models. In the course of the search for these hyperparameters, the model is trained on training data selected as 70% of the original dataset, which is subsequently specified when calling the training function. A visualization of the search and training is shown in Fig. 1.

¹Preliminary data preprocessing in machine learning: Instructions, Tools, and Useful Resources for Beginners, available at: <https://habr.com/ru/companies/skillfactory/articles/848858/>; A Simple Guide to Data Preprocessing in Machine Learning, available at <https://www.v7labs.com/blog/data-preprocessing-guide>; An overview of classification methods in machine learning using Scikit-Learn, available at: <https://tproger.ru/translations/scikit-learn-in-python>

²Hyperparameter selection, available at: <https://education.yandex.ru/handbook/ml/article/podbor-giperparametrov>; Hyperparameter search and model optimization, available at: <https://habr.com/ru/companies/otus/articles/754402/>; Quality metrics for binary classification models, available at: <https://loginom.ru/blog/classification-quality>; Quality assessment in classification and regression problems, available at: <https://neerc.ifmo.ru/wiki/index.php?title>

Table 1

Results of the models

Model name	Accuracy	Loss
Logistic Regression (LR)	0.894737	0.074432
Linear Discriminant Analysis (LDA)	0.884211	0.077352
K-Nearest Neighbors (KNN)	0.563158	0.094297
Classification and Regression Trees (CART)	0.921053	0.053931
Gaussian Naive Bayes (NB)	0.836842	0.086322
Support Vector Machines (SVM)	0.552632	0.026316
Random Forest Classifier (RF)	0.942105	0.043719

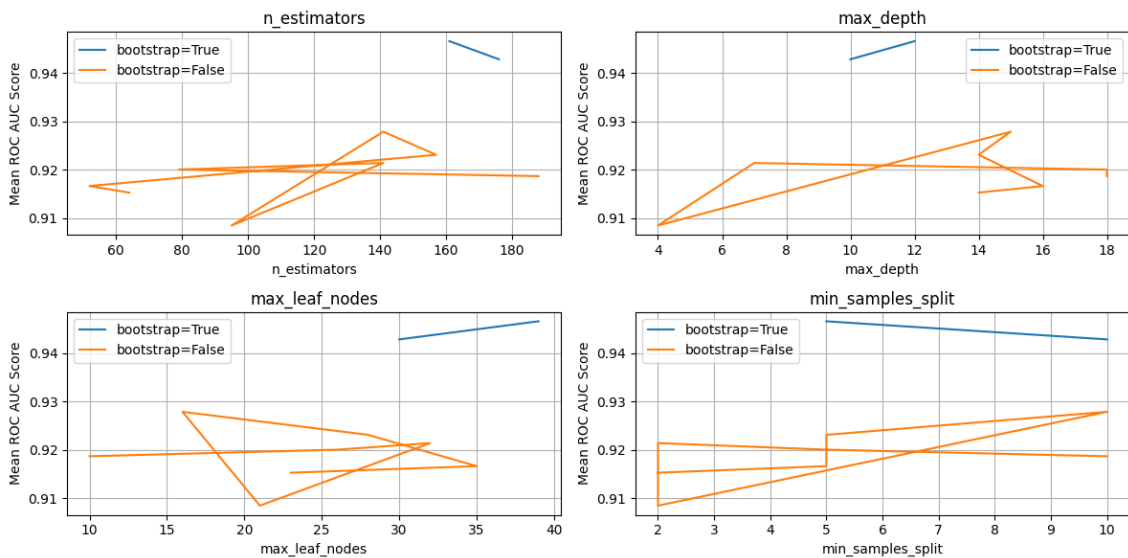


Fig. 1. Hyperparameter search and model training

The search identified the following best hyperparameters for the random forest model for our specific data:

- `n_estimators = 161`. This is the number of trees in a random forest. A larger number of trees typically leads to more accurate predictions, as the model becomes more robust to noise in the data;
- `min_samples_split = 5`. The minimum number of samples required to split a tree node. This parameter controls how “deep” the tree can grow; higher values prevent overfitting, reducing model complexity;

- `max_leaf_nodes = 39`. The maximum number of leaf nodes in a tree. Limiting the number of leaves helps control the complexity of the model and can improve generalization;
- `max_depth = 12`. Maximum tree depth. This setting limits how deep a tree can grow, which also helps prevent overtraining;
- `bootstrap = True`. Specifies whether bootstrapping is used to create subsamples of the data when training each tree. Setting this parameter to True means that each tree is trained on a random subsample of the data, which promotes tree diversity and improves overall model performance.

RESULTS AND DISCUSSION

For analysis in the development environment, the column names in the provided data set were changed to shorter ones, using Latin letters, according to the following legend (Table 2).

The target feature for which a predictive model needed to be developed was the “Postoperative Recurrence” feature (represented as “pr” in Table 2). The class distribution for this feature was tested and is presented in the form of a pie chart (Fig. 2).

Table 2

Legend of columns renaming

New name	Old name
id	Patient ID
age	Age
dotdm	Duration of disease, months
dbtt	TNM. diagnosis
dbtn	TNM. N diagnosis
dbtm	TNM. M diagnosis
ap	Alkaline phosphatase
tc	Total calcium
ttg	TSH
tfpl	Free T4, pmol/L
cpm	Calcitonin, pg/mL
phpl	Parathyroid hormone, pmol/L
at	Thyroglobulin antibodies, IU/mL
rea	CEA, ng/mL
cc	Cytological classification after FNA according to the Bethesda system (diagnostic category from 1 to 5)
cd	Associated Diseases
cds	Associated Cardiovascular Diseases
cdg	Associated Gastrointestinal Diseases
cdd	Associated Diseases of Connective Tissue Dysplasia
td	Type of surgery
apas	Alkaline phosphatase after surgery
tcas	Total calcium after surgery
tao	TSH after surgery
t4ao	T ₄ after surgery
cas	Calcitonin after surgery
phas	Parathyroid hormone after surgery
atas	Thyroglobulin antibodies after surgery

End of Table 2

New name	Old name
ras	CEA after surgery
dtahT	TNM diagnosis after histology. T
dtahn	TNM diagnosis after histology. N
dtahm	TNM diagnosis after histology. M
hsas	Postoperative hospital stay, days
ic	Intraoperative complications
pr	Postoperative recurrence

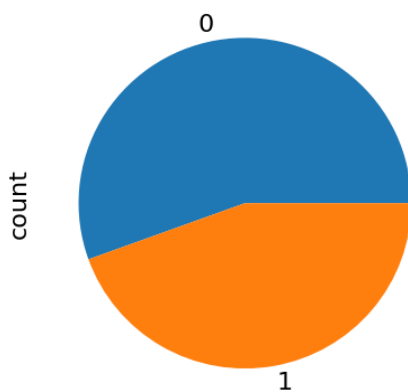


Fig. 2. Distributions of classes in the target feature

No class imbalance was observed. Therefore, the correlation of all features (excluding categorical ones) was calculated using the `corr` method, which calculates the Pearson correlation coefficient, a measure of the linear relationship between two variables, from the Pandas library for Python³ in the Visual Studio Code development environment. The feature correlation matrix is shown in Fig. 3.

The matrix shows that two features (“cas” and “cpm”) are highly correlated. To avoid information redundancy and, con-

sequently, a deterioration in the generalization ability of the predictive model, these features were removed. A correlation coefficient calculation method was used to identify highly correlated features. First, a correlation matrix is created, after which iteration is performed over its elements to identify pairs of features with an absolute correlation value above a given threshold (in this case, 0.75). All such features are added to a set, which is then used to remove these features from the original dataset. In our case, there is only one feature, and a strong correlation between the “cc” feature and the target feature is noticeable. A scatterplot of features was then constructed with respect to this feature, clearly dividing it into the classes of interest.

Using the `sns.boxplot(df)` function from the Seaborn library, we visualized the data distribution in the form of a boxplot. It allows us to compare the distributions of different data groups, identify outliers, and assess variability. The median (second quartile), first (Q_1), and third (Q_3) quartiles, which form the interquartile range (IQR), are displayed. The whiskers on the chart show the range of values within 1.5 IQR of the quartiles. Values outside these boundaries are designated as outliers and are displayed as individual dots.

³ AI with Python – Supervised Learning: Classification, available at: https://www.tutorialspoint.com/artificial_intelligence_with_python/artificial_intelligence_with_python_supervised_learning_classification.htm

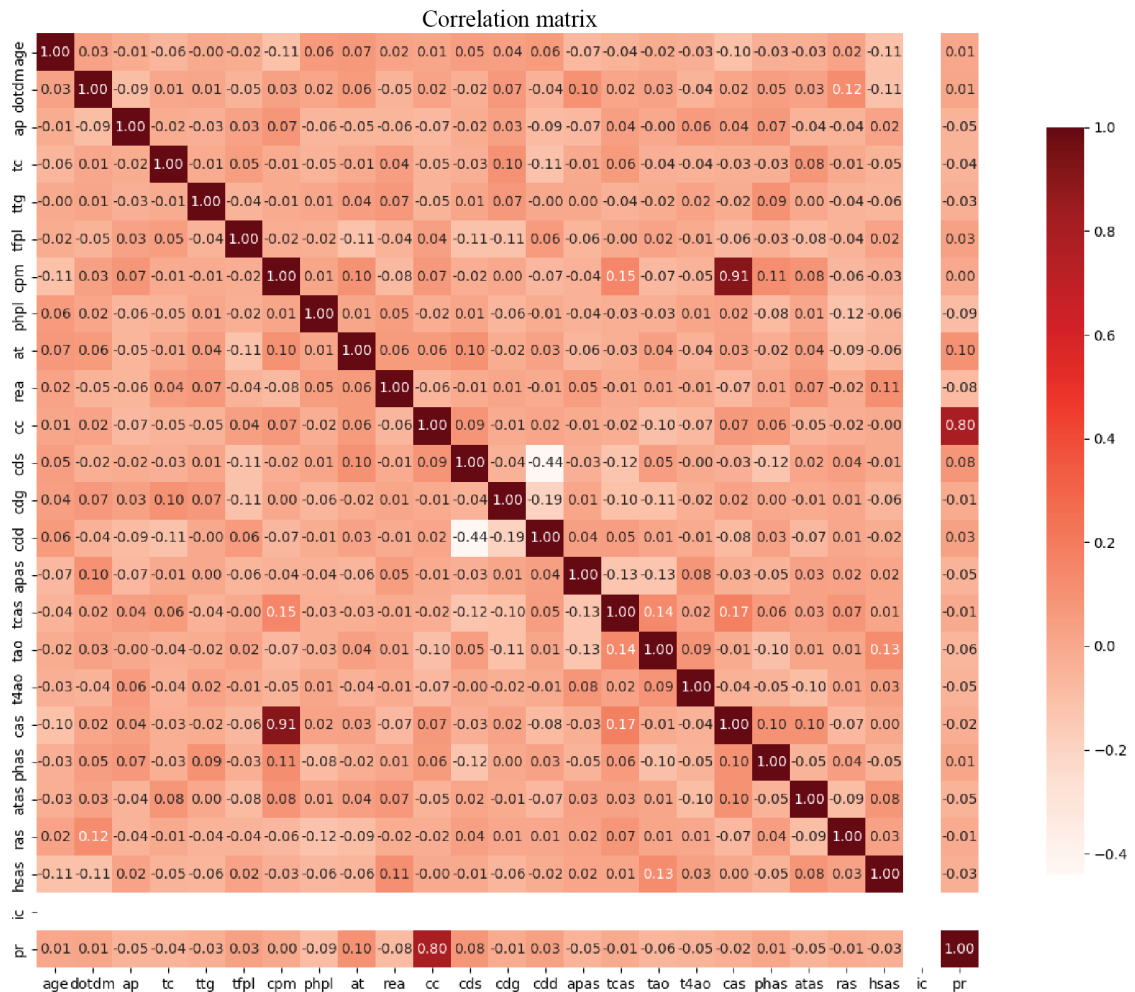


Fig. 3. Correlation matrix of features

The corresponding diagram is shown in Fig. 4. Before plotting the diagram, the data were normalized. The diagram in Fig. 4 shows that outliers in our data are present in the following parameters: TSH (indicated as “ttg” in the diagram), cytological classification after FNAB according to the Bethesda system (diagnostic category from 1 to 5) (“cc”), and total calcium after surgery (“tcas”).

Several simple methods can be used to remove outliers. One is the interquartile range (IQR). First, boundaries are calculated based on the first and third quartiles, and then values

outside these boundaries are removed. Another approach is the standard deviation: if a value is too far from the mean, it can be excluded.

The IQR method was chosen over standard deviation because it is more robust to the influence of outliers. Standard deviation value can be distorted by extreme values, leading to incorrectly defined boundaries for outlier removal. In contrast, IQR focuses on the central portion of the data and allows for the detection of anomalies without relying on extreme values. This makes IQR a more robust tool for data cleaning and model improvement.

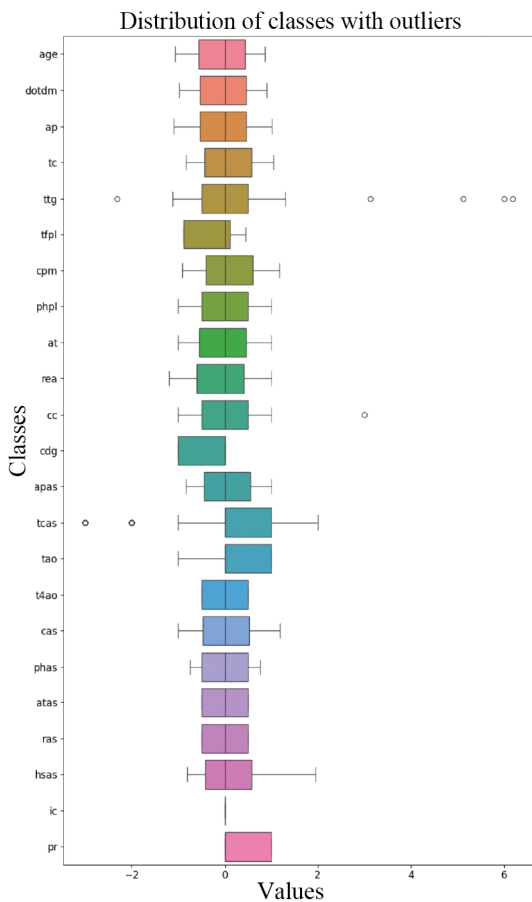


Fig. 4. Correlation matrix of features

Description of the interquartile range method (IQR)

1. Quartile determination: First, the first (Q_1) and third (Q_3) quartiles are calculated. The first quartile is the value below which 25 % of the data falls, and the third quartile is the value below which 75 % of the data falls.

2. IQR calculation: The interquartile range (IQR) is defined as the difference between Q_3 and Q_1 :

$$IQR = Q_3 - Q_1 .$$

3. Outlier detection: Values below $Q_1 - 1.5$ IQR or above $Q_3 + 1.5$ IQR are considered out-

liers and removed from the dataset. This leaves only 272 rows of the original 300 in the table.

To improve the accuracy and reliability of the predictive model, a duplicate sample search was performed as follows:

1. Counting duplicates: first, it is necessary to determine how many times each unique row occurs in the data set.

2. Filtering results: only those lines that occur more than once, i.e. only duplicates, are retained.

3. Output generation: if duplicates are found, a text description is generated for each group of duplicates, indicating the number of occurrences and characteristics of these lines. If there are no duplicates, a message is displayed stating that none were found.

No duplicates were found for the dataset in question.

Next, columns populated with the same value were found and removed.

Columns in the dataset were found and removed (using the method below) whose values were identical for all objects. Such columns do not contain variability and therefore do not provide useful information for analysis or model training. Their presence can create redundancy in the data and does not affect the quality of predictions, so they are excluded for model optimization.

The search was carried out in the following way:

1. Search of columns: columns in the data set are identified in which all values are the same, i.e. the number of unique values is 1. These are listed.

2. Output of results: then iterates through the found columns and displays in-

formation about each of them, indicating that the column is filled with one value and what that value is.

These were the following: TNM diagnosis. M (after renaming it “dbtm”), associated diseases (“cd”), TNM diagnosis after histology. N (“dtahn”), TNM diagnosis after histology. M (“dtahm”), and intraoperative complications (“ic”).

The found columns were removed by overwriting the original dataset with the same dataset, with the columns excluded from the list compiled during the search. In our case, there were five of them, leaving 28 of the 33 columns.

After training the model, its performance was evaluated using quality metrics such as accuracy, recall, and precision. These are calculated based on the classification results, presented as a confusion matrix, which includes four categories:

- true positive (TP): the number of positive cases correctly predicted;
- false positives (FP): the number of incorrectly predicted positive cases;
- true negatives (TN): the number of negative cases correctly predicted;
- false negatives (FN): the number of incorrectly predicted negative cases.

Accuracy measures how often a model makes correct predictions. It is calculated using the formula

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

This value shows the proportion of all correct predictions relative to the total number of predictions.

Recall reflects the model's ability to find all positive cases. It is calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

This metric shows what proportion of all actual positive cases the model was able to identify correctly.

Precision measures the proportion of correctly predicted positive cases among all cases classified as positive by the model. It is calculated using the formula

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

These metrics provide insight into how well the model performs on the task and identify potential issues, such as overfitting or insufficient ability to identify positive classes. The resulting calculations are presented in Table 3. Additionally, a classification report was generated, which provides more detailed information about the model's performance for each class. It is presented in Table 4.

Next, a matrix of errors was constructed. This allows one to evaluate how the model classifies the data by showing the distribution of true and predicted values. It includes correct and incorrect predictions for each class. This matrix is shown in Fig. 5.

The results show that the trained model demonstrates high accuracy of the target feature. The proportion of patients with postoperative recurrence correctly identified by the model was 98 % of the total number of patients with recurrence, while the proportion of patients without recurrence, correctly classified by the model as “patients at no risk

Table 3

Model quality metrics

Metrics	Value
Accuracy	0.963
Recall	0.963
Precision	0.964

Table 4

Classification report

Precision	Recall	F1-score
0.957	0.978	0.967
0.972	0.946	0.959
0.963	0.963	0.963
0.964	0.962	0.963
0.964	0.963	0.963

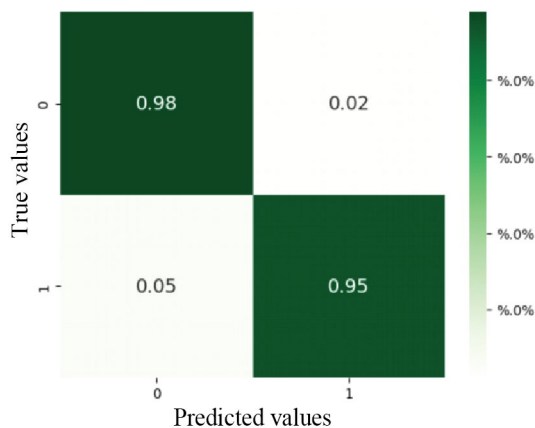


Fig. 5. Error matrix

of recurrence”, was 95% of all patients without recurrence. This demonstrates that the model effectively performs the classification task based on medical parameters, which may be particularly important for decision-making in clinical practice. High accuracy indicates the model's robustness and its ability to correctly identify recurrence cases, which may contribute to improved diagnosis and treatment.

CONCLUSIONS

The study developed a machine learning model to predict a high probability of thyroid cancer recurrence based on an analysis of medical parameters⁴. The development process began with careful data preprocessing, a critical step in building robust models. Outliers and columns containing uniform values were deleted during preprocessing, improving data quality and avoiding biases in model training. Categorical variables were also coded to ensure their correct use in machine learning algorithms, and correlated features were excluded to minimize multicollinearity and improve model interpretability.

To select the most suitable model, a comparative analysis of several classification algorithms was conducted.

As a result, the random forest method was selected, which demonstrated high efficiency in solving the classification problem. Using the random hyperparameter search method, the model was optimized, which allowed us to determine the best parameters to improve its performance. A prediction accuracy of 96% was achieved in the obtained model, which indicated its high reliability and ability to correctly classify.

The results of the studies highlight the potential for machine learning in medicine, particularly in the context of early diagnosis and disease monitoring. The model's high accuracy can significantly improve clinical decision-making and enhance the quality of medical care.

⁴ Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Volkov K.A., Kravchenya A.R., Rafeeva P.D., Trukhina M.K., Kapralov S.V., Amirov E.V., Maslyakov V.V. A system for predicting the occurrence of recurrences after surgery for thyroid cancer: Certificate of Registration of Computer Program No. 2024689824 dated 11.12.2024. Application dated 28.11.2024.

REFERENCES

1. *Berstein L.M.* Thyroid cancer: epidemiology, endocrinology, factors and mechanisms of carcinogenesis. *Praktical Onkology* 2007; 8 (1): 1–8 (in Russian).
2. *Lusbnikov E.F., Tsyb A.F., Yamashita S.* Thyroid cancer in Russia after Chernobyl. Moscow: Medicine 2006; 128 (in Russian).
3. *Bentz B.G. et al.* B-RAF V600E mutational analysis of fine needle aspirates correlates with diagnosis of thyroid nodules. *Otolaryngol. Head Neck Surg.* 2009; 140 (5): 709–714.
4. *Barchuk A.S.* Recurrences of differentiated thyroid cancer. *Practical Oncology* 2007; 8 (1): 35 (in Russian).
5. *Amin M.B., Greene F.L., Edge S.B. et al.* The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA Cancer J Clin.* 2017; 67 (2): 93–99.
6. *Kane S.M., Mulhern M.S., Poursabidi L.K. et al.* Micronutrients, iodine status and concentrations of thyroid hormones: a systematic review. *Nutr Rev.* 2018; 76 (6): 418–431.
7. *Agretti P. et al.* MicroRNA expression profile helps to distinguish benign nodules from papillary thyroid carcinomas starting from cells of fine-needle aspiration. *J. Eur. Endocrinol.* 2012; 167 (3): 393–400.
8. *Rumyantsev P.O., Ilyin A.A., Rumyantseva U.V. et al.* Thyroid cancer: modern approaches to diagnosis and treatment. Moscow: GEOTAR-Media 2009; 448 (in Russian).
9. *Bellevicine C. et al.* Cytological and molecular features of papillary thyroid carcinoma with prominent hobnail features: a case report. *Acta Cytol.* 2012; 56 (5): 560–564.
10. *Elisei R. et al.* The BRAFV600E mutation is an independent, poor prognostic factor for the outcome of patients with low-risk intrathyroid papillary thyroid carcinoma: single-institution results from a large cohort study. *J. Clin. Endocrinol. Metab.* 2012; 97 (12): 4390–4398.
11. *Makarin V.A.* Thyroid cancer. A manual for patients. Moscow 2016; 168 (in Russian).
12. Clinical guidelines. Differentiated thyroid cancer. Coding according to the International Statistical Classification of Diseases and Related Health Problems: C 73. Age group: adults. Moscow 2020 (in Russian).

Funding. The study had no external funding.

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

Author contributions:

Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Kravchenya A.R., Maslyakov V.V., Kapralov S.V. – study concept and design.

Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Volkov K.A., Kravchenya A.R., Maslyakov V.V., Kapralov S.V., Aslanov G.E., Losyakova E.V., Obukhov I.S., Osina A.D., Kurmaeva A.K. – data collection and processing.

Barulina M.A., Bendik I.Yu., Kovalenko I.I., Polidanov M.A. – statistical processing.

Barulina M.A., Bendik I.Yu., Kovalenko I.I., Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Volkov K.A., Kravchenya A.R., Maslyakov V.V., Kapralov S.V., Aslanov G.E., Losyakova E.V., Obukhov I.S., Osina A.D., Kurmaeva A.K. – editing.

All authors approved the final version of the paper.

Limitation of the study. Permission to conduct the study was obtained from the Local Ethics Committee (LEC) of Reaviz Medical University (Protocol No. 9 dated September 10, 2024). The study was conducted with the patients' voluntary informed consent in accordance with the declaration of compliance with international and Russian ethical principles and standards (extract from Protocol No. 19 of the Bioethics Committee meeting dated October 26, 2018). The study was conducted in accordance with the requirements of the World Medical Association's Declaration of Helsinki (as amended in 2013).

Received: 03/05/2025

Revised version received: 05/08/2025

Accepted: 05/23/2025

Please cite this article in English as: Barulina M.A., Bendik I.Yu., Kovalenko I.I., Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Volkov K.A., Kravchenya A.R., Maslyakov V.V., Kapralov S.V., Aslanov G.E., Losyakova Ye.V., Obukhov I.S., Osina A.D., Kurmaeva A.K. Possibility of predicting the probability of thyroid cancer recurrence by machine learning methods. *Perm Medical Journal*, 2025, vol. 42, no. 3, pp. 130-143. DOI: 10.17816/pmj423130-143