

УДК 004.85,004.89  
DOI 10.52575/2687-0932-2025-52-3-665-680  
EDN UQKESM

## Использование методов машинного обучения для решения задачи предсказания развития отклонений в живых системах

<sup>1</sup> Маренков Я.А. , <sup>1</sup> Ломакин А.С. , <sup>1,2</sup> Донская А.Р. , <sup>2</sup> Панина А.А. 

<sup>1</sup> Волгоградский государственный технический университет,  
Россия, 400005, Волгоград, пр. им. Ленина, 28

<sup>2</sup> Волгоградский государственный медицинский университет Минздрава России,  
Россия, 400066, г. Волгоград, пл. Павших Борцов, 1  
yaroslavmarenkov5@gmail.com, arseny.lomakin@gmail.com, donsckaia.anastasiya@yandex.ru,  
panina0675@yandex.ru

**Аннотация.** Целью исследования является частичная автоматизация процесса ранней диагностики онкологических заболеваний за счет определения подозрения на наличие у пациента ракового заболевания на базе машинного обучения и предоставления интерпретирующей информации. В мире одними из самых смертоносных заболеваний человека являются онкологии различных органов, и ему подвергается большая группа населения. Согласно данным Минздрава России и расчётам Росстата в России в период с 2005 года по 2023 год в России ежегодно выявляли от 427 тыс. до 552 тыс. новых больных. Для решения задачи поиска эффективных и точных методов раннего диагностирования рака в медицине в последнее время все чаще используют различные компьютерные технологии, в том числе и методы машинного обучения. Главным недостатком данных методов является низкая интерпретируемость человеком получаемого результата. Поэтому возникает потребность в разработке интерпретируемого метода для определения подозрения на наличие ракового заболевания. В данной статье рассматривается проблема интерпретируемости результатов прогнозов, выполняемых моделями машинного обучения для решения задачи предсказания развития отклонений в живых системах на примере новообразований поджелудочной железы у человека. На основе результатов расширенного анализа крови, сопоставленных с фактом наличия или отсутствия у пациента рака поджелудочной железы, была разработана интерпретируемая модель машинного обучения. Использовалась локальная интерпретируемость модели с помощью метода SHAP (Shapley Additive exPlanations) с визуализацией в виде водопадного графика. Были обучены разнообразные модели классического машинного обучения, произведен сравнительный анализ данных моделей для выявления результата с большей точностью. На основе полученной модели было сделано формирование интерпретирующей информации. Было разработано пользовательское приложение для взаимодействия с моделью. Наилучший результат показала модель случайного леса (Random Forest) с f1-score, равным 0,859.

**Ключевые слова:** машинное обучение, интерпретация предсказаний моделей машинного обучения, модели машинного обучения, методы машинного обучения, ранняя диагностика онкологических заболеваний

**Для цитирования:** Маренков Я.А., Ломакин А.С., Донская А.Р., Панина А.А. 2025. Использование методов машинного обучения для решения задачи предсказания развития отклонений в живых системах. *Экономика. Информатика*, 52(3): 665–680. DOI 10.52575/2687-0932-2025-52-3-665-680. EDN UQKESM



# Applying Machine Learning Methods to Solve the Problem of Predicting the Progression of Abnormalities in Living Systems

<sup>1</sup> Yaroslav A. Marenkov , <sup>1</sup> Arseniy S. Lomakin ,

<sup>1,2</sup> Anastasia R. Donskaia , <sup>2</sup> Anna A. Panina 

<sup>1</sup> Volgograd State Technical University,  
28 Lenin Ave., Volgograd, 400005, Russia

<sup>2</sup> Volgograd State Medical University of Public Health Ministry of the Russian Federation,  
1 Pavshikh Bortsov Sq., Volgograd, 400131, Russia

yaroslavmarenkov5@gmail.com, arseny.lomakin@gmail.com, donsckaia.anastasiya@yandex.ru,  
panina0675@yandex.ru

**Abstract.** The purpose of the research is to partially automate the process of early cancer diagnosis by identifying suspected cancer in a patient based on machine learning and providing interpretive information. Cancer of various organs is among the deadliest human diseases in the world, and a large group of population is exposed to it. According to the data of the Ministry of Health of Russia and calculations of Rosstat for 2005–2023, from 427 thousand to 552 thousand new patients were detected in Russia annually. To solve the problem of searching for effective and accurate methods of early cancer diagnosis in medicine, various computer technologies, including machine learning methods, have recently become increasingly used. The main disadvantage of these methods is low human interpretability of the result obtained. Therefore, there is a need to develop an interpretable method to determine suspected cancer. This paper deals with the problem of interpretability of prediction results performed by machine learning models to solve the problem of predicting the development of abnormalities in living systems using human pancreatic cancer as an example. An interpretable machine learning model was developed based on the results of an extended blood test matched to the fact whether the patient has pancreatic cancer or not. We used local interpretability of the model using SHAP (Shapley Additive exPlanations) method with visualization in the form of a waterfall graph. A variety of classical machine learning models were trained; a comparative analysis of these models was made to identify the result with higher accuracy. Based on the obtained model, interpretive information was formed. A user application was developed to interact with the model. The Random Forest model performed best with an f1-score of 0.859.

**Keywords:** machine learning, interpretation of machine learning model predictions, machine learning models, machine learning methods, early diagnosis of cancer

**For citation:** Marenkov Ya.A., Lomakin A.S., Donskaia A.R., Panina A.A. 2025. Applying machine learning methods to solve the problem of predicting the progression of abnormalities in living systems. *Economics. Information technologies*, 52(3): 665–680 (in Russian). DOI 10.52575/2687-0932-2025-52-3-665-680. EDN UQKESM

## Введение

В России и в мире в целом одной из глобальных проблем в сфере здравоохранения остаются онкологические заболевания. Согласно сведениям Росстата, собранным на основании данных Министерства Здравоохранения России за период с 2005 по 2023 год, ежегодно регистрируют от 442 тыс. до 552 тыс. новых случаев злокачественных образований<sup>1</sup>.

Одним из ключевых вопросов данной проблемы является ранняя диагностика. На помощь с решением данного вопроса может прийти использование искусственного интеллекта и, в частности, машинного обучения. Его использование может ускорить работу врачей, так как оно позволяет быстро обрабатывать большие объемы данных. Построенные на этой базе программные средства могут упростить и ускорить процесс диагностики, при этом не исключая врача из него. Также немаловажным фактором в данном процессе является обеспечение врача возможностью интерпретации результатов, выдаваемых моделью, чтобы он

<sup>1</sup> Заболеваемость населения социально-значимыми болезнями: Здравоохранение // Росстат: официальный сайт. – URL: <https://rosstat.gov.ru/folder/13721> (дата обращения: 20.03.2025).

(врач) при необходимости мог перепроверить полученный результат или обратить внимание на те или иные показатели.

Авторами была поставлена цель – частичная автоматизация процесса ранней диагностики онкологических заболеваний за счет определения подозрения на наличие у пациента новообразований на базе машинного обучения и предоставления интерпретирующей информации.

В ходе данной статьи будут рассмотрены моменты по созданию интерпретируемого метода классического машинного обучения для решения задачи предсказания развития отклонений в живых системах на примере новообразований поджелудочной железы у человека, выполнена оценка и произведен сравнительный анализ полученных результатов, а также представлено разработанное для данного метода программное средство.

## **Объекты и методы исследования**

### ***Описание предметной области***

Для диагностирования различных видов онкологических заболеваний Министерством Здравоохранения России разработаны соответствующие клинические рекомендации. Каждая из них содержит описание симптомов, виды исследований, такие как первичный физикальный осмотр пациента, проведение лабораторно-диагностических исследований, проведение инструментально-диагностических исследований, а также иные виды исследований. Одним из недостатков является сложность и/или продолжительность процесса анализа полученных результатов.

Постоянное увеличение набора новых решений, направленных на выявление заболеваний на стадии лабораторных исследований, является важным вектором развития контура практического здравоохранения Российской Федерации, обобщая весь опыт, накопленный практикующими врачами, и сокращая временные издержки на постановку диагноза. Для решения данных задач могут применяться разнообразные подходы: классические (алгоритмические), онтологический инжиниринг или машинное обучение.

В первом случае алгоритм является сложным в построении, так как требует глубокого изучения предметной области, изучения детального поведения и различных состояний исследуемого объекта в отношении к результирующей характеристике.

Подход в онтологическом инжиниринге предлагает разработать формализованную математическую модель целевой области знаний. Для этого необходимо также детально изучить целевой объект и описать его в виде набора знаний взаимосвязи между ними. Это является на данный момент достаточно трудоёмким процессом для решения локальной задачи.

Третий подход основан на построении статистической модели, который обобщает набор накопленных знаний, строит собственные плохо интерпретируемые человеком связи, но имеет высокую точность на схожих задачах.

В ходе исследования был изучен ряд других публикаций по диагностированию онкологических заболеваний с использованием машинного обучения.

Например, исследователи [Гундырев И.А. и др., 2018] в своей работе рассматривали различные классификаторы для диагностирования рака легких по 8 биохимическим параметрам, полученным из анализа слюны (полученная точность – 72–78 %). В исследовании [Барчука А.А. и др., 2017] приводится сравнение 13 моделей в задачах классификации образцов патологической ткани при злокачественных процессах грудной клетки на основе уровней экспрессии генов (полученная точность – 82 %). В своем исследовании [Reddy, Khanaa, 2023] представляют алгоритм обнаружения рака легких с помощью сегментации изображений компьютерной томографии, используя компьютерное зрение, и последующей классификации рака с помощью моделей глубокого и неглубокого обучения (полученная точность – 92 %, классическое машинное обучение – 88 %). Исследователи [Qiu et al., 2018] в своей работе представили метод диагностирования рака поджелудочной железы на основе текстурного анализа компьютерной томографии (полученная точность – 79 %).



Для повышения качества моделей некоторые исследователи приходят к генерации новых дополнительных характеристик. Так, Демидова Л.А. в своем исследовании [Demidova, 2023] генерирует новые характеристики с помощью алгоритма нелинейного уменьшения размерности (UMAP), аппроксимации энтропии, а также фрактальных размерностей Каца и Хигучи на основе маркеров белков крови (полученное значение по метрике ассурасу – 0,96).

В связи с разнообразием различных наборов данных командой исследования будет исследована задача сравнения типовых моделей машинного обучения для классификации табличных данных на примере диагностирования онкологических заболеваний на данных лабораторных исследований пациентов, который описан ниже.

### **Формализация и постановка задачи**

Главной задачей данного исследования является разработка интерпретируемой модели машинного обучения для задачи диагностирования онкологических заболеваний (на примере рака поджелудочной железы по расширенным анализам крови пациента). Далее, основываясь на полученной модели, необходимо разработать приложение с графическим интерфейсом для пользователей, которое позволит им просмотреть результаты работы модели (предсказания) и интерпретирующую информацию. В следующих разделах будут подробно описаны шаги, выполненные для решения поставленной задачи.

Формально задача определения факта наличия онкологического заболевания представляется как задача классификации следующим образом. Имеется множество случаев  $M = \{m_1, \dots, m_n\}$ , где каждый элемент данного множества характеризуется набором характеристик расширенного анализа крови  $A = \{a_1, \dots, a_k\}$ , и ему в соответствие установлен факт наличия или отсутствия у пациента рака поджелудочной железы.

Алгоритм идеального бинарного классификатора представляется следующей формулой:

$$F(m_i, C_i) = \begin{cases} 0, & \text{если } m_i \notin C_i \\ 1, & \text{если } m_i \in C_i \end{cases}$$

Таким образом, задача сводится к поиску такого классификатора  $F'$ , который по точности будет максимально близок к  $F$ .

Классификация с точки зрения типа предсказания в данном исследовании рассматривалась с нескольких сторон:

- Бинарный ответ (принадлежность к классу наличия или отсутствия заболевания). Может принимать одно из двух значений  $\{0, 1\}$ . Таким образом,

$$F' = M \times C \rightarrow \{0, 1\};$$

- Степень подобия (степень, на которую параметры предсказания удовлетворяют для принадлежности к тому или иному классу). Может принимать значение в диапазоне  $[0; 1]$ . Таким образом,

$$F' = M \times C \rightarrow [0, 1];$$

- Использование вспомогательных интерпретирующих средств.

Исследование основано на базе размеченного набора данных. Данный набор был разделен на обучающую выборку –  $L$  и тестовую выборку –  $T$ . При этом данные подвыборки не пересекаются, то есть  $L \cap T = \emptyset$ .

### **Дизайн исследования**

#### **Описание набора данных**

Для исследования использовался набор данных  $M$ , который изначально (до предварительной обработки данных) содержал 62 показателя расширенного анализа крови  $A$  для 1254 записей деперсонализированных пациентов, проходивших обследование. Также каждому элементу данного набора установлен в соответствие факт наличия или отсутствия рака поджелудочной железы  $C$ . Изначально множество  $M$  содержало 1115 случаев без заболевания и 139 – с заболеванием. Таким образом, набор данных является несбалансированным.

### *Анализ и предварительная обработка набора данных*

Анализ набора данных начался с изучения типов характеристик, предоставляемых набором  $A$ . Далее производился поиск и удаление характеристик с большим количеством пропусков (более 10 % в общем или более 40 % для положительного класса). Затем производился количественный анализ и удаление характеристик с малым количеством вариаций (1-2). После чего с целью проверки данных на их принадлежность к нормальному распределению выполнялся одновыборочный тест Колмогорова – Смирнова [Berger, Zhou, 2014] (с уровнем статистической значимости, равным 0,05). По итогам данного теста данные не имеют нормального распределения. Исходя из того, что данные не имеют нормального распределения, и стоит задача бинарной классификации, то для поиска статистически незначимых характеристик для целевой переменной  $C$  использовался тест Мана – Уитни [McKnight, Najab, 2010] (с уровнем статистической значимости, равным 0,05). По итогам данного теста статистически незначимые характеристики удалялись. В результате данных действий в наборе характеристик осталось 13 параметров.

Далее перед началом обучения производилась предварительная обработка данных. Сначала для каждой характеристики из набора  $A$  по каждой записи  $M$  был произведен поиск и замена пропущенных значений на модальное значение данной характеристики. Затем производился поиск с последующим удалением выбросов при помощи метода межквартильного диапазона [Yang et al., 2019]. Далее данные разбивались на две непересекающиеся выборки: тренировочную (70 %) и тестовую (30 %). После чего для лучшего качества работы моделей и уменьшения риска переобучения данные масштабировались, используя метод стандартизации [Raju et al., 2020]. По итогу для обучения моделей машинного обучения получился набор данных  $M'$ , который содержит 808 записей (669 записей для пациентов без заболевания и 139 – с заболеванием). Тренировочная выборка  $L$  содержит 565 записей (468 – без заболевания, 97 – с заболеванием). Тестовая выборка  $T$  содержит 243 записи (201 – без заболевания, 42 – с заболеванием).

### *Задача классификации*

Классификация выполняет задачу установки одной метки из конечного набора меток (классов) для ранее не помеченного объекта. В машинном обучении алгоритм принимает на вход набор размеченных данных, а на выходе он формирует классификационную функцию (модель), которая в свою очередь принимает неразмеченные данные и возвращает либо саму метку, либо степень подобия каждого из возможных классов [Soofi, Awan, 2017].

Задача классификации в машинном обучении может решаться с помощью алгоритмов классического машинного обучения и глубокого машинного обучения. В данном исследовании было решено использовать алгоритмы классического машинного обучения, поскольку, во-первых, для качественного обучения моделей нейронных сетей имеющийся набор данных довольно-таки мал, и с задачей обучения на малых наборах данных как раз лучше справляются именно алгоритмы классического машинного обучения. Во-вторых, в ходе анализа предметной области было выявлено, что чаще используются и показывают лучший результат на табличных данных именно алгоритмы классического машинного обучения, а глубокое обучение используется в основном для анализа изображений.

В данном исследовании использовались следующие алгоритмы классического машинного обучения: дерево решений (Decision Tree) [Priyanka, Kumar, 2020], случайный лес (Random Forest) [Schonlau, Zou, 2020], гистограммный градиентный бустинг (Hist Gradient Boosting) [Guryanov, 2019]. Также использовались следующие статистические алгоритмы:  $k$  ближайших соседей [Cunningham, Delany, 2021] и машина опорных векторов (Support Vector Machine) [Mustafa, Abdulazeez, 2021]. Использовались реализации данных алгоритмов, предоставляемые библиотекой scikit-learn.

### *Настройка и валидация моделей классификации*

Валидация моделей происходила с помощью следующих метрик: accuracy, precision, recall, f1-мера, матрица ошибок (confusion matrix), roc-кривые, log loss и MCC. Также для



объективной оценки качества работы моделей для метрик, для которых это возможно, использовалась кросс-валидация [Bates et al., 2023]. Использовались реализации данных метрик из библиотеки `scikit-learn`.

Модели, реализации которых представлены в библиотеках по машинному обучению, имеют базовые гиперпараметры, на которых данный алгоритм выдает хороший результат для большинства задач. Но для получения лучшего результата в конкретной задаче их необходимо настраивать. В данном исследовании все рассматриваемые модели изначально обучались на базовых гиперпараметрах (за исключением одного, поскольку в данных имеется дисбаланс классов, то для борьбы с данной проблемой было выполнено перевзвешивание классов в моделях), затем для модели, показавшей лучший результат, был выполнен подбор гиперпараметров с помощью метода `Grid Search`.

### *Проблема интерпретируемости*

При работе с моделями машинного обучения одной из проблем является их интерпретируемость. Это означает, что в результатах предсказаний моделей машинного обучения не всегда понятно, из-за чего модель сделала тот, а не иной прогноз (работа как с «черным» ящиком). Причем между интерпретируемостью и сложностью моделей (в том числе и их точностью) существует обратно-пропорциональная зависимость – чем выше сложность модели, тем обычно выше ее точность, но меньше интерпретируемость [Stiglic et al., 2020].

Под интерпретируемостью в данной работе понимается смысл прогнозов. Интерпретируемость бывает локальной и глобальной.

Локальная интерпретируемость позволяет понять, почему модель сделала такой вывод для одного конкретного прогноза, показывает можно ли доверять данному прогнозу [Stiglic et al., 2020].

Глобальная же интерпретируемость, соответственно, демонстрирует, как модель работает в целом на множестве данных. Данный вид интерпретируемости позволяет найти взаимосвязь характеристик, определить наиболее важные характеристики для модели и просмотреть общую логику прогнозирования [Stiglic et al., 2020; Hakkoum et al., 2024].

### *Обучение моделей машинного обучения*

#### *Результаты обучения на базовых гиперпараметрах*

Изначально все рассматриваемые модели (указанные в дизайне исследования) обучались на их базовых гиперпараметрах, предоставляемых реализацией данных алгоритмов в библиотеке `scikit-learn`<sup>1,2,3,4,5</sup>, исключая один параметр, который был изменен – `class_weight`. Для него было установлено значение «`balanced`», которое автоматически выполняет перевзвешивание классов в зависимости от его частоты появления в данных. Также для фиксации случайной составляющей и корректного сравнения метрик моделей было установлено значение «12» для гиперпараметра `random_state`.

Валидация моделей происходила с помощью способов, указанных в соответствующем разделе дизайна исследования. Сравнительная табл. 1 содержит значения числовых метрик для рассматриваемых алгоритмов.

---

<sup>1</sup> `DecisionTreeClassifier`: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> (дата обращения: 08.05.2025).

<sup>2</sup> `RandomForestClassifier`: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> (дата обращения: 08.05.2025).

<sup>3</sup> `HistGradientBoostingClassifier`: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html> (дата обращения: 08.05.2025).

<sup>4</sup> `KNeighborsClassifier`: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html> (дата обращения: 08.05.2025).

<sup>5</sup> `SVC`: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> (дата обращения: 08.05.2025).

Таблица 1  
Table 1

Сравнение числовых метрик  
Comparison of numerical metrics

| Метрика / Metric | Дерево решений / Decision Tree | Случайный лес / Random Forest | Гистограммный градиентный бустинг / Hist Gradient Boosting | К ближайших соседей / K Nearest Neighbors | Машина опорных векторов / Support Vector Machine |
|------------------|--------------------------------|-------------------------------|------------------------------------------------------------|-------------------------------------------|--------------------------------------------------|
| Accuracy         | 0,893                          | 0,926                         | 0,909                                                      | 0,914                                     | 0,876                                            |
| Precision        | 0,813                          | 0,958                         | 0,861                                                      | 0,902                                     | 0,785                                            |
| Recall           | 0,813                          | 0,786                         | 0,804                                                      | 0,778                                     | 0,775                                            |
| F1-мера          | 0,813                          | 0,842                         | 0,828                                                      | 0,822                                     | 0,779                                            |
| Log loss         | 3,856                          | 0,216                         | 0,352                                                      | 1,454                                     | 0,269                                            |
| MCC              | 0,625                          | 0,724                         | 0,663                                                      | 0,669                                     | 0,560                                            |

Далее в табл. 2 представлены изображения графических метрик для каждого из рассматриваемых алгоритмов.

Таблица 2  
Table 2

Сравнение графических метрик  
Comparison of graphical metrics

| Алгоритм / Algorithm           | Матрица ошибок / Confusion matrix | ROC-кривые / ROC-curve |
|--------------------------------|-----------------------------------|------------------------|
| 1                              | 2                                 | 3                      |
| Дерево решений / Decision Tree |                                   |                        |
| Случайный лес / Random Forest  |                                   |                        |

Окончание табл. 2  
End of the table 2

| Алгоритм /<br>Algorithm                                             | Матрица ошибок / Confusion matrix        | ROC-кривые / ROC-curve                   |
|---------------------------------------------------------------------|------------------------------------------|------------------------------------------|
| 1                                                                   | 2                                        | 3                                        |
| Гистограммный<br>градиентный<br>бустинг / Hist<br>Gradient Boosting | <p>Hist Gradient Boosting Classifier</p> | <p>Hist Gradient Boosting Classifier</p> |
| К ближайших<br>соседей /<br>K Nearest<br>Neighbors                  | <p>K Neighbors Classifier</p>            | <p>K Neighbors Classifier</p>            |
| Машина опорных<br>векторов / Support<br>Vector Machine              | <p>SVM Classifier</p>                    | <p>SVM Classifier</p>                    |

Так как в данных присутствует дисбаланс классов, то наиболее показательными метриками является F1-мера, MCC и ROC-кривые. Также немаловажной является и матрица ошибок. Таким образом, лучшей моделью на базовых гиперпараметрах является алгоритм случайного леса. Именно на данном алгоритме и выполнялся дальнейший подбор гиперпараметров с помощью метода Grid Search.



### *Интерпретируемость моделей машинного обучения, использованная в исследовании*

В данном исследовании большее значение имеет именно локальная интерпретируемость, так как важна интерпретация для прогноза модели по данным пациента, которые только что ввели в программу.

Для интерпретации моделей в данной работе использовался метод SHAP (Shapley Additive exPlanations), который является инструментом как для локальной, так и для глобальной интерпретируемости. Он основан на кооперативной теории игр, позволяет интерпретировать многие алгоритмы машинного обучения, описан в статьях [Li, 2022; Pelegina, 2023; Lundberg, Lee, 2017]. Данный инструмент предоставляет информацию в удобном графическом виде. Для данного исследования был выбран график водопадного вида, который иллюстрирует влияние каждой из характеристик на принятое решение моделью в сторону одного из классов, и представляет собой интерпретацию степени подобия для одного конкретного объекта.

Рассмотрим следующий пример такого интерпретирующего графика, который представлен на рис. 1.

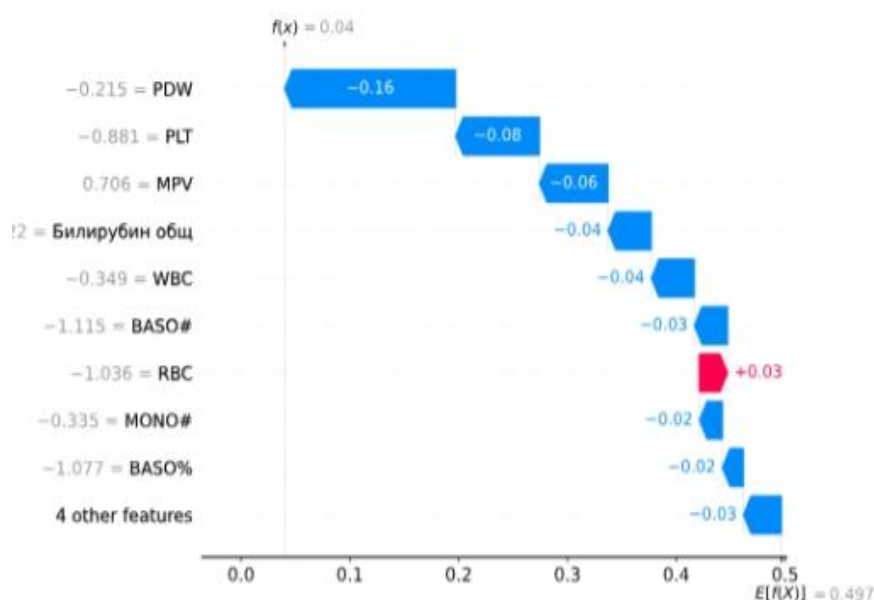


Рис. 1. Пример водопадного интерпретирующего графика SHAP

Fig. 1. An example of a waterfall interpretive chart SHAP

На данном примере представлен интерпретирующий график SHAP для одного из предсказаний. По данному графику видно, что большинство характеристик внесли свой вклад в сторону предсказания отрицательного класса. При этом также видно, за счет водопадной стилистики, какая характеристика внесла наибольший вклад в предсказание (в данном случае PDW), поскольку отображение данных отсортировано по вносимому вкладу в предсказание. Данный факт может позволить эксперту в данной предметной области определить те факторы, которым модель отвела большее значение, и, сопоставив их со своими знаниями и опытом, определить, насколько можно довериться результатам предсказания для каждого отдельного случая.

### **Результаты и их обсуждение**

#### *Результирующая модель машинного обучения*

Результирующая модель была получена в результате подбора гиперпараметров для модели случайного леса, которая дала лучший результат на базовых гиперпараметрах. В табл. 3 указаны гиперпараметры алгоритма случайного леса, подбор которых был выполнен.

Таблица 3  
Table 3

Подбираемые значения гиперпараметров  
Selectable hyperparameter values

| Гиперпараметр / Hyperparameter | Набор значений / Value set    |
|--------------------------------|-------------------------------|
| n_estimators                   | 100, 200, 300, 50, 25, 75, 10 |
| criterion                      | 'gini', 'entropy', 'log_loss' |
| max_depth                      | None, 5, 7, 10, 15            |
| max_features                   | 'sqrt', 'log2'                |
| min_samples_split              | 2, 3, 4, 5, 7, 10             |
| min_samples_leaf               | 1, 2, 3, 4, 5                 |

В итоге лучший результат по Grid Search показала модель случайного леса со следующим набором гиперпараметров:

- n\_estimators: 25;
- criterion: «entropy»;
- max\_depth: 10;
- max\_features: «sqrt»;
- min\_samples\_split: 10;
- min\_samples\_leaf: 1.

Данная модель имеет следующие числовые метрики:

- accuracy: 0,926;
- precision: 0,895;
- recall: 0,833;
- f1-score: 0,859;
- log loss: 0,261;
- mss: 0,725.

На рис. 2 приведены матрица ошибок и ROC-кривая для полученной модели.

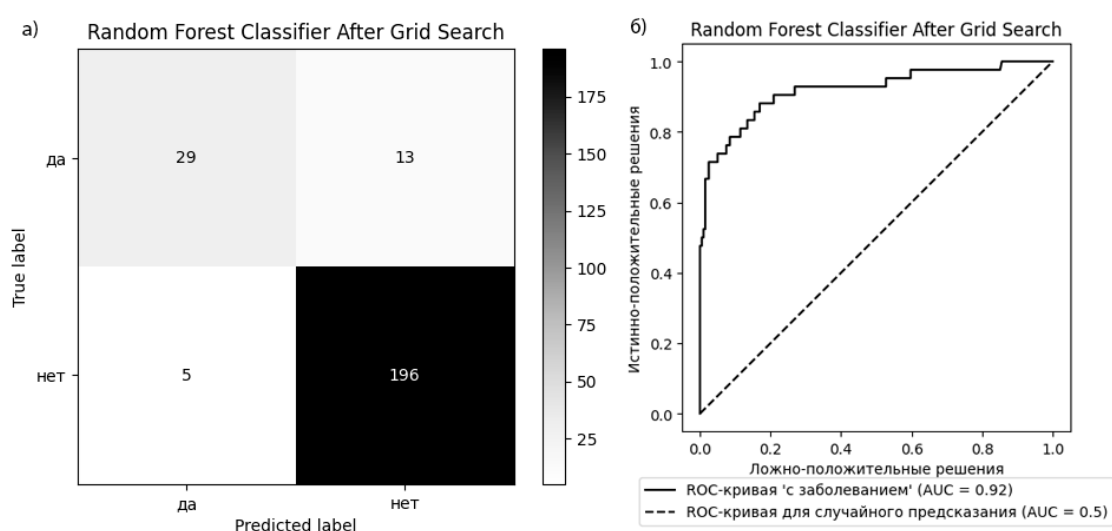


Рис. 2. Графические метрики для случайного леса после подбора гиперпараметров:

а) матрица ошибок; б) ROC-кривые

Fig. 2. Graphical metrics for Random Forest after hyperparameter selection:

a) error matrix; b) ROC curves

Полученная модель использовалась в дальнейшем для создания интерпретирующей информации и разработки программного средства, описанного в следующем разделе.

### Разработанное программное средство

Далее полученная модель использовалась при разработке клиент-серверного веб-приложения. Данное приложение разрабатывалась на языке программирования Python с использованием фреймворка Flask. По своей сути приложение представляет собой одностраничное приложение.

Клиент-серверное взаимодействие – это распределение системы на две части: клиент и сервер. В простом случае клиент выполняет функции логики представления и общения с сервером с помощью запросов. Сервер выполняет вычислительную часть приложения, то есть занимается обработкой полученных данных, выполнением необходимых вычислений и формированием результата<sup>1</sup>.

Далее на рис. 3, 4, 5 приведены примеры работы программы.

Рис. 3. Пример работы программы: запуск программы  
Fig. 3. Example of the program functioning: start the program

| Метрика   | Значение |
|-----------|----------|
| accuracy  | 0.9259   |
| recall    | 0.7857   |
| precision | 0.9589   |
| f1        | 0.8422   |
| log_loss  | 0.2162   |
| mcc       | 0.7242   |

Рис. 4. Пример работы программы: предсказание для здорового пациента  
Fig. 4. Example of the program functioning: prediction for a healthy patient

<sup>1</sup> Иванько, А.Ф. Информационные системы в издательском деле : учебное пособие / А.Ф. Иванько, М.А. Иванько. – Санкт-Петербург : Лань, 2022. – 148 с. – ISBN 978-5-8114-3843-3.

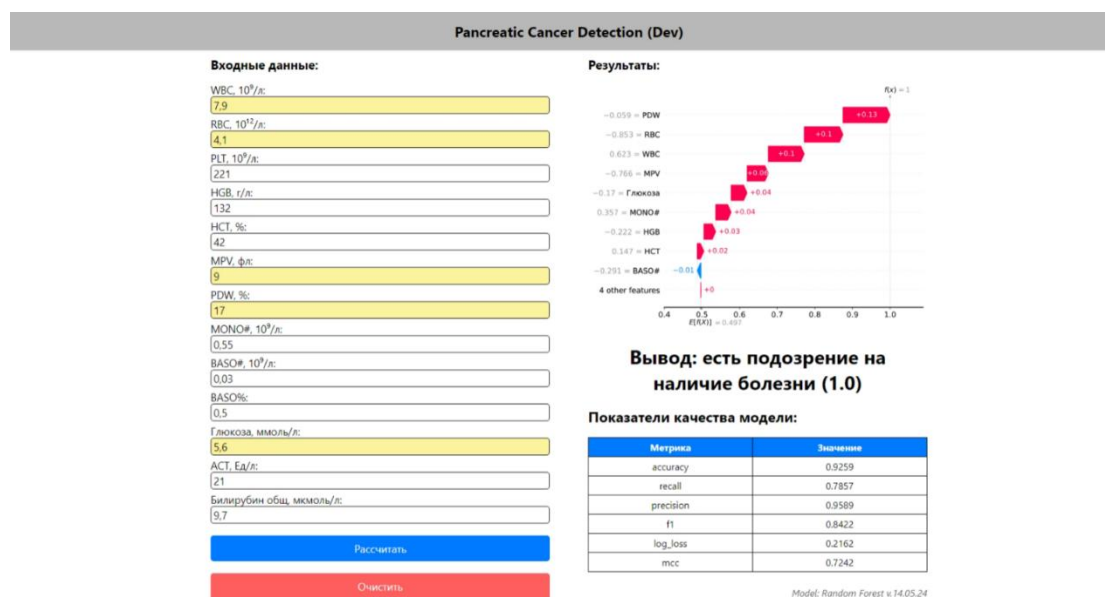


Рис. 5. Пример работы программы: предсказание для больного пациента  
Fig. 5. Example of the program functioning: prediction for a patient with disease

## Заключение

В ходе данной работы был произведен анализ предметной области, поставлена, формализована и исследована задача определения факта наличия или отсутствия у пациента онкологического заболевания на примере раннего диагностирования рака поджелудочной железы.

В ходе выполнения задачи был произведен анализ и предварительная обработка исходного набора данных для дальнейшего обучения. Далее был произведен выбор набора классификаторов и произведен их сравнительный анализ и настройка для исследуемой задачи. По результатам наилучший результат показал алгоритм случайного леса (Random Forest) после подбора гиперпараметров с f1-score 0,859. После чего полученная модель дополнительно использовалась для создания интерпретирующей информации с помощью метода SHAP. Затем для полученной модели было разработано программное средство с пользовательским графическим интерфейсом для взаимодействия с моделью и отображения результатов работы модели, включая интерпретирующую информацию в виде графика водопадного вида.

В дальнейшем командой исследователей планируется улучшение качества предсказания и добавление дополнительной интерпретирующей информации. Также планируется масштабирование полученного программного средства для применения в массовой лабораторной проверке анализов.

## Благодарности

Авторы выражают благодарность за помощь со сбором данных и в проведении исследований лаборатории ГУЗ «КДП №2» г. Волгограда, а также следующим людям:

– Панин Станислав Игоревич, доктор медицинских наук, профессор, заведующий кафедрой общей хирургии ФГБОУ ВО «Волгоградский государственный медицинский университет» Минздрава России (г. Волгоград, Россия);

– Суворов Владимир Александрович, кандидат медицинских наук, врач-онколог, заведующий торакоабдоминальным отделением;

– Зубков Александр Владимирович, кандидат технических наук, начальник управления информационного развития ФГБОУ ВО «Волгоградский государственный медицинский университет» Минздрава России (г. Волгоград, Россия), доцент кафедры программного обеспечения автоматизированных систем ФГБОУ ВО «Волгоградский государственный технический университет» (г. Волгоград, Россия);

– *Безбородов Сергей Александрович*, кандидат технических наук, доцент, заведующий кафедрой клинической инженерии и технологий искусственного интеллекта ФГБОУ ВО «Волгоградский государственный медицинский университет» Минздрава России (г. Волгоград, Россия)

– *Коваленко Надежда Витальевна*, кандидат медицинских наук, доцент, заведующий кафедрой онкологии, гематологии и трансплантологии Института НМФО (Непрерывного Медицинского и Фармацевтического Образования) ФГБОУ ВО «Волгоградский государственный медицинский университет» Минздрава России (г. Волгоград, Россия);

– *Шушкова Ирина Геннадьевна*, к.м.н., ассистент кафедры лучевой, функциональной, лабораторной диагностики ИНМФО ФГБОУ ВО «Волгоградский государственный медицинский университет» Минздрава России (г. Волгоград, Россия);

– *Быков Александр Викторович*, д.м.н., профессор, профессор кафедры хирургических болезней № 1 ИНМФО, ФГБОУ ВО «Волгоградский государственный медицинский университет» Минздрава России (г. Волгоград, Россия).

### Список источников

1. Заболеваемость населения социально-значимыми болезнями: Здравоохранение // Росстат: официальный сайт. – URL: <https://rosstat.gov.ru/folder/13721> (дата обращения: 20.03.2025).
2. DecisionTreeClassifier: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> (дата обращения: 08.05.2025).
3. RandomForestClassifier: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> (дата обращения: 08.05.2025).
4. HistGradientBoostingClassifier: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html> (дата обращения: 08.05.2025).
5. KNeighborsClassifier: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html> (дата обращения: 08.05.2025).
6. SVC: [Справочник API] // Scikit-learn: официальный сайт. – URL: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> (дата обращения: 08.05.2025).
7. Иванько, А.Ф. Информационные системы в издательском деле: учебное пособие / А.Ф. Иванько, М.А. Иванько. – Санкт-Петербург: Лань, 2022. – 148 с. – ISBN 978-5-8114-3843-3.

### Список литературы

- Барчук А., Атрощенко А., Гайдуков В., Виноградов П., Тараканов С., Канаев С., Арсеньев А., Комаров Ю., Харитонов М., Барчук, А., Мерабишвили В., Кузнецов В., Трофимов В., Гусарова Н., Коцюба И., Беляев А., Подольский М., Нефедова, А. 2017. Автоматизированная диагностика в популяционном скрининге рака легкого. *Вопросы онкологии*, 63(2): 215–220. <https://doi.org/10.37469/0507-3758-2017-63-2-215-220>.
- Гундырев И.А., Бельская Л.В., Косенок В.И., Сарф Е.А. 2018. Применение синтетических образов для решения задачи классификации на примере диагностики рака легкого. *Вестник Российской Академии медицинских наук*, 73(2): 96–104. <https://doi.org/10.15690/vramn946>.
- Bates S., Hastie T., Tibshirani R., 2023. Cross-Validation: What Does It Estimate and How Well Does It Do It. *Journal of the American Statistical Association*, 119(546): 1434–1445. <https://doi.org/10.1080/01621459.2023.2197686>.
- Berger V., Zhou Y., ред. Balakrishnan N., Colton T., Everitt B., Piegorsch W., Ruggeri F., Teugels J.L., 2014. Kolmogorov–Smirnov Test: Overview. Wiley StatsRef: Statistics Reference Online. <https://doi.org/10.1002/9781118445112.stat06558>.
- Cunningham P., Delany S.J., 2021. K-Nearest Neighbour Classifiers - A Tutorial. *ACM Comput. Surv*, 54(6): article number 128. <https://doi.org/10.1145/3459665>.
- Demidova L.A., 2023. A Novel Approach to Decision-Making on Diagnosing Oncological Diseases Using Machine Learning Classifiers Based on Datasets Combining Known and/or New Generated Features of a Different Nature. *Mathematics*, 11(4). Article number: 792. <https://doi.org/10.3390/math11040792>.



- Guryanov A., 2019. Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees. Analysis of Images, Social Networks and Texts: 8th International Conference, *AIST 2019*. Cham: Springer. 39–50. [https://doi.org/10.1007/978-3-030-37334-4\\_4](https://doi.org/10.1007/978-3-030-37334-4_4).
- Hakkoum H., Idri A., Abnane I., 2024. Global and local interpretability techniques of supervised machine learning black box models for numerical medical data. *Engineering Applications of Artificial Intelligence*, 131. <https://doi.org/10.1016/j.engappai.2023.107829>.
- Li Z., 2022. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, 96: article number 101845. <https://doi.org/10.1016/j.compenvurbsys.2022.101845>.
- Lundberg S.M., Lee S., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems. Red Hook: Curran Associates Inc. 4768–4777. <https://doi.org/10.48550/arXiv.1705.07874>.
- McKnight P.E., Najab J., ред. Weiner I.B., Craighead W.E., 2010. Mann-Whitney U Test. The Corsini Encyclopedia of Psychology. <https://doi.org/10.1002/9780470479216.corpsy0524>.
- Mustafa A.D., Abdulazeez A.M., 2021. Machine Learning Applications based on SVM Classification A Review. *Qubahan Academic Journal*, 1(2): 81–90. <https://doi.org/10.48161/qaj.v1n2a50>.
- Pelegrina, Duarte L.T., Grabisch M., 2023. A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artificial Intelligence*, 325: article number 104014. <https://doi.org/10.48550/arXiv.2211.02166>.
- Priyanka, D. Kumar., 2020. Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, 12(3): 246–269. <https://doi.org/10.1504/ijids.2020.10029122>.
- Qui J., Wu Y., Chen J., Hui B., Huang Z., Ji L., 2018. A texture analysis method based on statistical contourlet coefficient applied to the classification of pancreatic cancer and normal pancreas. Proceedings of the International Symposium on Big Data and Artificial Intelligence (ISBDAI '18). New York: ACM. 9–13. <https://doi.org/10.1145/3305275.3305278>.
- Raju V.N.G., Lakshmi K.P., Jain V.M., Kalidindi A., Padma V., 2020. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. Third International Conference on Smart Systems and Inventive Technology (ICSSIT). Tirunelveli: IEEE. 729–735. <http://dx.doi.org/10.1109/ICSSIT48917.2020.9214160>.
- Reddy N.S., Khanaa V., 2023. Intelligent deep learning algorithm for lung cancer detection and classification. *Bulletin of Electrical Engineering and Informatics*, 12(3): 1747–1754. <https://doi.org/10.11591/eei.v12i3.4579>.
- Schonlau M., Zou R. Y., 2020. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1): 3–29. <https://doi.org/10.1177/1536867X20909688>.
- Soofi A.A., Awan A., 2017. Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, 13: 459–465. <http://dx.doi.org/10.6000/1927-5129.2017.13.76>.
- Stiglic G., Kocbek P., Fijacko N., Zitnik M., Verbert K., Cilar L., 2020. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5): e1379. <https://doi.org/10.1002/widm.1379>.
- Yang J., Rahardja S., Fränti P., 2019. Outlier detection: how to threshold outlier scores. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing (AIIPCC '19). New York: ACM. 1–6. <https://doi.org/10.1145/3371425.3371427>.

## References

- Barchuk A., Atroshchenko A., Gaidukov V., Vinogradov P., Tarakanov S., Kanaev S., Arsen'ev A., Komarov YU., Kharitonov M., Barchuk, A., Merabishvili V., Kuznetsov V., Trofimov V., Gusarova N., Kotsyuba I., Belyaev A., Podol'skii M., Nefedova, A., 2017. Automated diagnosis in a population-based screening for lung cancer. *Problems in oncology*, 63(2): 215–220. <https://doi.org/10.37469/0507-3758-2017-63-2-215-220>. (in Russian)
- Gundyrev I.A., Bel'skaya L.V., Kosenok V.K., Sarf E.A., 2018. The use of synthetic images for solving the classification problem by the example of lung cancer diagnosis. *Annals of the Russian academy of medical sciences*, 73(2): 96–104. <https://doi.org/10.15690/vramn946>. (in Russian)
- Bates S., Hastie T., Tibshirani R., 2023. Cross-Validation: What Does It Estimate and How Well Does It Do It. *Journal of the American Statistical Association*, 119(546): 1434–1445. <https://doi.org/10.1080/01621459.2023.2197686>.

- Berger V., Zhou Y., ред. Balakrishnan N., Colton T., Everitt B., Piegorsch W., Ruggeri F., Teugels J.L., 2014. Kolmogorov–Smirnov Test: Overview. Wiley StatsRef: Statistics Reference Online. <https://doi.org/10.1002/9781118445112.stat06558>.
- Cunningham P., Delany S.J., 2021. K-Nearest Neighbour Classifiers - A Tutorial. *ACM Comput. Surv.*, 54(6): article number 128. <https://doi.org/10.1145/3459665>.
- Demidova L.A., 2023. A Novel Approach to Decision-Making on Diagnosing Oncological Diseases Using Machine Learning Classifiers Based on Datasets Combining Known and/or New Generated Features of a Different Nature. *Mathematics*, 11(4). Article number: 792. <https://doi.org/10.3390/math11040792>.
- Guryanov A., 2019. Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees. Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019. Cham: Springer. 39–50. [https://doi.org/10.1007/978-3-030-37334-4\\_4](https://doi.org/10.1007/978-3-030-37334-4_4).
- Hakkoum H., Idri A., Abnane I., 2024. Global and local interpretability techniques of supervised machine learning black box models for numerical medical data. *Engineering Applications of Artificial Intelligence*, 131. <https://doi.org/10.1016/j.engappai.2023.107829>.
- Li Z., 2022. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, 96: article number 101845. <https://doi.org/10.1016/j.compenvurbsys.2022.101845>.
- Lundberg S.M., Lee S., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems. Red Hook: Curran Associates Inc. 4768–4777. <https://doi.org/10.48550/arXiv.1705.07874>.
- McKnight P.E., Najab J., ред. Weiner I.B., Craighead W.E., 2010. Mann-Whitney U Test. The Corsini Encyclopedia of Psychology. <https://doi.org/10.1002/9780470479216.corpsy0524>.
- Mustafa A.D., Abdulazez A.M., 2021. Machine Learning Applications based on SVM Classification A Review. *Qubahan Academic Journal.*; 1(2): 81–90. <https://doi.org/10.48161/qaj.v1n2a50>.
- Pelegrina, Duarte L.T., Grabisch M., 2023. A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artificial Intelligence*, 325: article number 104014. <https://doi.org/10.48550/arXiv.2211.02166>.
- Priyanka, D. Kumar., 2020. Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, 12(3): 246–269. <https://doi.org/10.1504/ijids.2020.10029122>.
- Qui J., Wu Y., Chen J., Hui B., Huang Z., Ji L., 2018. A texture analysis method based on statistical contourlet coefficient applied to the classification of pancreatic cancer and normal pancreas. Proceedings of the International Symposium on Big Data and Artificial Intelligence (ISBD AI '18). New York: ACM. 9–13. <https://doi.org/10.1145/3305275.3305278>.
- Raju V.N.G., Lakshmi K.P., Jain V.M., Kalidindi A., Padma V., 2020. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. Third International Conference on Smart Systems and Inventive Technology (ICSSIT). Tirunelveli: IEEE. 729–735. <http://dx.doi.org/10.1109/ICSSIT48917.2020.9214160>.
- Reddy N.S., Khanaa V., 2023. Intelligent deep learning algorithm for lung cancer detection and classification. *Bulletin of Electrical Engineering and Informatics*, 12(3): 1747–1754. <https://doi.org/10.11591/eei.v12i3.4579>.
- Schonlau M., Zou R. Y., 2020. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1): 3–29. <https://doi.org/10.1177/1536867X20909688>.
- Soofi A.A., Awan A., 2017. Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, 13: 459–465. <http://dx.doi.org/10.6000/1927-5129.2017.13.76>.
- Stiglic G., Kocbek P., Fijacko N., Zitnik M., Verbert K., Cilar L., 2020. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5): e1379. <https://doi.org/10.1002/widm.1379>.
- Yang J., Rahardja S., Fränti P., 2019. Outlier detection: how to threshold outlier scores. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing (AIPCC '19). New York: ACM. 1–6. <https://doi.org/10.1145/3371425.3371427>.

**Конфликт интересов:** о потенциальном конфликте интересов не сообщалось.

**Conflict of interest:** no potential conflict of interest related to this article was reported.

Поступила в редакцию 03.06.2025

Поступила после рецензирования 31.08.2025

Принята к публикации 08.09.2025

Received June 03, 2025

Revised August 31, 2025

Accepted September 08, 2025



## ИНФОРМАЦИЯ ОБ АВТОРАХ

**Маренков Ярослав Анатольевич**, студент магистратуры 1 курса кафедры «Программное обеспечение автоматизированных систем», Волгоградский государственный технический университет, г. Волгоград, Россия.

ORCID: 0009-0005-7314-8924

**Ломакин Арсений Сергеевич**, студент магистратуры 1 курса кафедры «Программное обеспечение автоматизированных систем», Волгоградский государственный технический университет, г. Волгоград, Россия.

ORCID: 0009-0001-9340-1748

**Донская Анастасия Романовна**, старший преподаватель кафедры программного обеспечения автоматизированных систем, Волгоградский государственный технический университет, г. Волгоград, Россия; старший преподаватель кафедры клинической инженерии и технологий искусственного интеллекта, Волгоградский государственный медицинский университет Минздрава России, г. Волгоград, Россия.

ORCID: 0000-0003-3086-4929

**Панина Анна Александровна**, доктор медицинских наук, профессор кафедры лучевой, функциональной, лабораторной диагностики ИНМФО, Волгоградский государственный медицинский университет Минздрава России, г. Волгоград, Россия.

ORCID: 0000-0003-2750-8579

## INFORMATION ABOUT THE AUTHORS

**Yaroslav A. Marenkov**, 1st year Master's degree student of the Department of Software Engineering, Volgograd State Technical University, Volgograd, Russia.

ORCID: 0009-0005-7314-8924

**Arseniy S. Lomakin**, 1st year Master's degree student of the Department of Software Engineering, Volgograd State Technical University, Volgograd, Russia.

ORCID: 0009-0001-9340-1748

**Anastasia R. Donskaia**, Senior Lecturer at the Department of Software Engineering, Volgograd State Technical University, Volgograd, Russia; Senior lecturer at the Department of Clinical Engineering and Artificial Intelligence Technologies, Volgograd State Medical University of the Ministry of Health of Russia, Volgograd, Russia.

ORCID: 0000-0003-3086-4929

**Anna A. Panina**, MD, Associate Professor of the Department of Radiation, Functional, Laboratory Diagnostics, Volgograd State Medical University of the Ministry of Health of Russia, Volgograd, Russia.

ORCID: 0000-0003-2750-8579