

**Информация для цитирования:**

Колесникова Т.Г., Черноусов М.Е., Минина В.И., Вылегжанина А.В., Минин А.В., Баканова М.Л. АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В ОЦЕНКЕ ГЕНЕТИЧЕСКОЙ ПРЕДРАСПОЛОЖЕННОСТИ К РАКУ ЛЕГКИХ У РАБОТНИКОВ УГОЛЬНЫХ ШАХТ // Медицина в Кузбассе. 2025. №4. С. 82-88.

**Колесникова Т.Г., Черноусов М.Е., Минина В.И., Вылегжанина А.В., Минин А.В., Баканова М.Л.**

Кемеровский Государственный Университет,  
Федеральный исследовательский центр угля и углехимии СО РАН,  
Кемеровский Государственный Медицинский Университет,  
г. Кемерово, Россия



## АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В ОЦЕНКЕ ГЕНЕТИЧЕСКОЙ ПРЕДРАСПОЛОЖЕННОСТИ К РАКУ ЛЕГКИХ У РАБОТНИКОВ УГОЛЬНЫХ ШАХТ

**Цель исследования** – в работе представлены результаты исследования, направленного на определение генетической предрасположенности к раку легких у работников угольных предприятий Кемеровской области.

**Материалы и методы.** Были использованы данные, собранные во время медицинских профосмотров. Данные обрабатывались с использованием алгоритмов машинного обучения.

**Результаты.** Рассмотрены различные методы моделирования, включая деревья решений, случайный лес, логистическую регрессию и другие. Особое внимание уделено оценке точности предсказания риска развития рака легких и определению значимых признаков, таких как курение и определенные генетические варианты.

**Заключение.** Полученные результаты свидетельствуют о высокой эффективности алгоритмов случайного леса и логистической регрессии, а также о возможности применения методов машинного обучения для анализа сложных биомедицинских данных.

**Ключевые слова:** рак легких; генетическая предрасположенность; векторизация текста; алгоритмы машинного обучения; дерево решений; случайный лес; логистическая регрессия; метод опорных векторов; метод k ближайших соседей

**Kolesnikova T.G., Chernousov M.E., Minina V.I., Vylegzhanina A.V., Minin A.V., Bakanova M.L.**

Kemerovo State University,  
The Federal Research Center of Coal and Coal Chemistry of Siberian Branch of the Russian Academy of Sciences,  
Kemerovo State Medical University, Kemerovo, Russia

### MACHINE LEARNING ALGORITHMS IN ASSESSING GENETIC PREDISPOSITION TO LUNG CANCER AMONG COAL MINES WORKERS IN THE KEMEROVO REGION

**The aim of the research** – the paper presents the results of a study aimed at determining the genetic predisposition to lung cancer in workers at coal enterprises in the Kemerovo Region.

**Materials and methods.** Data collected during medical examinations were used. The data was processed using machine learning algorithms.

**Results.** Various modeling methods are considered, including decision trees, random forest, logistic regression, and others. Particular attention is paid to assessing the accuracy of predicting the risk of developing lung cancer and identifying significant features, such as smoking and certain genetic variants.

**Conclusion.** The results obtained indicate the high efficiency of random forest and logistic regression algorithms, as well as the possibility of using machine learning methods to analyze complex biomedical data.

**Key words:** lung cancer; genetic predisposition; text vectorization; machine learning algorithms; decision tree; random forest; logistic regression; support vector machine; method KNN

По данным ВОЗ, одним из самых распространенных видов онкологических заболеваний у мужчин является рак легких. Также он занимает одно из первых мест среди причин смерти от злокачественных новообразований, как у мужчин, так и у женщин [1].

Анализ научной литературы по данному вопросу показал, что в Российской Федерации статистические данные по заболеваемости и смертности по-

казывают общую тенденцию к снижению, начиная с 1997 года. Наибольшее количество зарегистрированных случаев заболевания было зафиксировано в 1997 году – 65660 человек, после чего наблюдалось постепенное снижение до минимального значения в 2020 году – 54375 человек. С 2021 года отмечается небольшой рост числа заболевших, достигший 58905 случаев в 2023 году. Количество умерших от рака легких также демонстрирует общую тенденцию

к снижению, начиная с 1997 года, когда было зарегистрировано 61364 смерти. Минимум был достигнут в 2021 году – 46798 смертей, после чего наблюдается некоторый рост показателей до 48104 случаев в 2023 году (табл. 1) [2].

**Таблица 1**  
**Динамика заболеваемости и смертности от рака легких в Российской Федерации с 1997 по 2023 год (чел./год)**  
**Table 1**  
**Dynamics in the Russian Federation from 1997 to 2023 (persons/year)**

Год	Абс. число новых случаев заболевания, чел.	Абс. число умерших, чел.
1997	65660	61364
1998	64727	60201
1999	63092	59301
2000	63125	58872
2001	62145	57018
2002	60337	56245
2003	58812	54181
2004	59057	53646
2005	57635	52695
2006	57219	51902
2007	57175	51912
2008	56767	51364
2009	57052	51433
2010	56985	51322
2011	56030	50440
2012	55475	49908
2013	56051	50068
2014	57685	49730
2015	60351	51280
2016	60467	51476
2017	62175	50186
2018	61686	50774
2019	60113	50046
2020	54375	49158
2021	56328	46798
2022	57421	46443
2023	58905	48104

Проблема заболеваемости раком легких чрезвычайно актуальна в регионах с развитой горнодобывающей промышленностью. Работники угледобывающих предприятий, находящиеся в условиях длительного воздействия угольной пыли, радона и других вредных веществ, подвергаются особенно высокому риску развития этого заболевания. К примеру, в Кемеровской области уровень заболеваемости и смертности от рака легких несколько выше обобщенных показателей по Российской Федерации, что можно увидеть, сравнивая стандартизированные коэффициенты заболеваемости и смертности от рака легких, рассчитанные для Кемеровской области и по России в целом (рис.) [2].

Представленные данные указывают на устойчивое снижение как заболеваемости, так и смертности от рака легких в Российской Федерации в целом, и

в Кемеровской области в частности. Это может быть связано со снижением распространенности курения, и другими профилактическими мерами, а также с успехами в лечении онкологических заболеваний. Тем не менее, высокий уровень заболеваемости и смертности сохраняется, особенно среди определенных групп населения, что требует дальнейшего изучения и разработки эффективных стратегий профилактики.

Для работников с вредными факторами проводятся периодические медосмотры по профессии согласно Приказу Минздрава России от 28.01.2021 № 29н. Медосмотры помогают выявить профессиональные заболевания на ранней стадии их развития. Если работа сотрудника будет связана с воздействием вредных и/или опасных факторов, предварительный медицинский осмотр позволит определить, подходит кандидат на должность по состоянию здоровья или нет. Предполагается, что 47% случаев рака легкого могут быть предотвращены при устранении вредной профессиональной экспозиции, связанной с риском развития злокачественных новообразований данной локализации.

Снизить уровень заболеваемости среди работников вредных производств возможно. Для этого необходимо избегать привлечения людей, имеющих генетическую предрасположенность к развитию рака легких, к работе в условиях, способных спровоцировать заболевание.

Таким образом, раннее выявление лиц с высоким риском развития рака легких в силу генетической предрасположенности становится критически важным.

Современные методы генетического анализа позволили накопить большой объем информации о вариациях генома, связанных с различными заболеваниями, в том числе с онкопатологией органов дыхания [3].

Различные коллективы исследователей разрабатывают новые модели на основе методов машинного обучения, которые могут применяться в узких областях медицинской практики. Например, большие языковые модели могут применяться для генерации отчетов и различного рода выписок из истории болезни, а также для составления прогнозов на основе полученных результатов анализов [4]. Также возможно применение больших языковых моделей для проведения удаленных консультаций с помощью умных чат-ботов [5, 6].

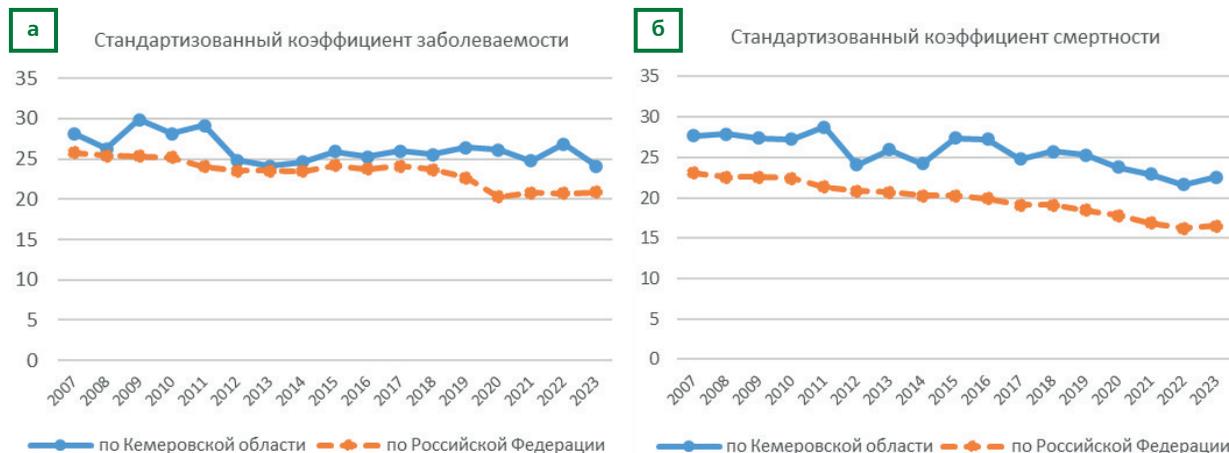
Кластеризация является одной из ключевых задач в медицинской диагностике, где целью является разделение большого массива данных на группы объектов, обладающих высокой степенью сходства [7, 8]. Методы машинного обучения, решающие задачу кластеризации, выявляют скрытые структуры в медицинских данных и позволяют разделить все множество пациентов на подгруппы в зависимости от их симптомов, анамнеза, лабораторных показателей и прочих характеристик [9]. Такие алгоритмы, как k-means, DBSCAN, метод главных компонент и пр., успешно применяются для поддержки

Рисунок

Стандартизированные коэффициенты заболеваемости/смертности по Кемеровской области и по Российской Федерации за период с 2007 по 2023 годы: а) коэффициент заболеваемости; б) коэффициент смертности

Figure

Standardized morbidity/mortality rates for the Kemerovo region and the Russian Federation for the period from 2007 to 2023: a) morbidity rate; b) mortality rate



принятия врачебных решений и выбора стратегии лечения [10-12].

В период пандемии COVID-19 мощное развитие получили методы распознавания и классификации изображений с помощью нейронных сетей. Разработка алгоритмов для автоматической диагностики COVID-19 на основе рентгеновских КТ-снимков грудной клетки стала одним из приоритетных направлений. Такие системы позволяли быстро и точно определять наличие инфекции, снижая нагрузку на медицинский персонал и ускоряя процесс постановки диагноза [13, 14].

Машинное обучение, как подмножество технологий искусственного интеллекта, способно внести большой вклад в анализ генетических данных благодаря своим уникальным возможностям обработки больших объемов информации и выявления сложных взаимосвязей. Оно может быть использовано для создания моделей, которые предсказывают вероятность развития онкологических заболеваний на основе генетической информации пациента [15].

В Кемеровской области были обнаружены свои особенности генома у местных жителей, которые могут сильно влиять на риск развития этого заболевания. Поэтому важно изучить уникальные генетические особенности работников угольных шахт и других вредных производств в регионе, чтобы иметь возможность предсказывать вероятность развития у них рака легких [16, 17].

В рамках современных исследований генетических основ рака легких особое внимание уделяется роли герминальных генетических вариантов, которые могут приводить к нарушениям клеточных процессов и способствовать злокачественному перерождению тканей легкого.

Традиционные методы оценки генетического риска (такие как полногеномный анализ ассоциаций,

ген-кандидатные исследования, сегрегационный анализ) имеют ряд ограничений:

- могут не учитывать всех возможных факторов влияния (образ жизни и вредные привычки, место работы, экология мест проживания, иммунитет конкретного человека);

- могут быть субъективны и зависят от качества собранной информации;

- не учитывают сложный характер взаимодействия генов, способный приводить как к усилению, так и подавлению возможных эффектов.

Целью данного исследования была оценка способности алгоритмов машинного обучения (МО) идентифицировать наиболее значимые генетические варианты, ассоциированные с предрасположенностью к заболеванию раком легких работников угольных шахт Кемеровской области.

## МАТЕРИАЛ И МЕТОДЫ

Набор данных, использованный для обучения моделей, содержит информацию о 514 работниках угольных предприятий Кемеровской области. Для каждого человека указаны следующие данные:

- информация о генетических вариантах: *IL1b rs16944*, *TNFA rs1800629*, *APEX1 rs1130409*, *XPD (ERCC2) rs13181*, *TGFB1 rs1800471*, *EGFR rs2227984*, *EPHX1 rs1051740*, *CHEK2 rs555607708*, которые по результатам проведенных ранее исследований могут быть связаны с развитием рака легких у шахтеров;

- курит ли человек (да/нет);

- наличие заболевания раком легких: болен или здоров;

- если человек болен, то указан диагноз и стадия развития болезни.

Датасет был разделен на обучающую и тестовую выборки, в тестовую выборку вошли 153 наблюдение

ния, остальные данные использовались для обучения моделей.

Данные являются подмножеством набора данных, собранных в г. Кемерово. Исследование соответствовало этическим стандартам биоэтического комитета (Протокол исследования был одобрен Этическим комитетом ИЭЧ ФИЦ УУХ СО РАН г. Кемерово, протокол № 1 от 01.07.2025), разработанным в соответствии с Хельсинкской декларацией Всемирной ассоциации «Этические принципы проведения научных медицинских исследований с участием человека» с поправками 2000 г. и «Правилами клинической практики в Российской Федерации», утвержденными Приказом Минздрава РФ от 19.06.2003 г. № 266. До включения в исследование у всех участников было получено письменное информированное согласие.

Подробное описание исследуемой группы пациентов и геномного исследования образцов крови описано в [18].

Были использованы следующие методы МО [19]:

- дерево решений,
- случайный лес,
- логистическая регрессия,
- метод опорных векторов,
- метод k ближайших соседей.

В первой серии экспериментов для перекодирования значений аллелей был использован метод «one-hot encoding», который позволяет представить аллели в виде бинарных признаков, где каждому уникальному значению соответствует отдельный столбец. Таким образом, перешли от представления данных в виде строк («GC», «TT», «AT» и др.) к числовым значениям 0 и 1. Обучение каждого алгоритма МО проводилось дважды: первый раз на полном наборе данных, второй раз после удаления дубликатов (21 строка). Это позволило оценить влияние избыточности данных на результаты моделирования.

Во второй серии экспериментов данные рассматривались как текстовые последовательности, где каждая строка представляла собой отдельное предложение, а каждое слово внутри строки соответствовало генотипу в формате {имя\_гена}:{генотип}. Такой подход позволил применить методы обработки естественного языка (векторизация текста способом TF-IDF) для анализа генетических данных.

Итоговые показатели качества работы алгоритмов на наборе данных, преобразованном с применением метода TF-IDF, показаны в таблице 2. Обучение алгоритмов проводилось на тренировочном наборе данных (дублирующиеся строки не удалялись).

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ОБСУЖДЕНИЕ

Исходя из полученных результатов, можно сделать следующие выводы:

1). Наибольшую точность показали два алгоритма: случайный лес и логистическая регрессия. На наборе данных с учетом дубликатов они достигли уровня Accurasy 0,80.

2). Алгоритмы «Дерево решений» и «Случайный лес» позволяют выделить наиболее значимые признаки для предсказания. Эти признаки связаны с более высоким риском развития рака легких. Также оба метода указывают на высокую значимость курения, известного социо-культурного фактора риска развития рака легких.

Исследование сосредоточено на анализе небольшого количества генов, которые по результатам выполненных ранее исследований считаются связанными с риском рака легких. В данном случае рассматриваются только 8 генов из нескольких тысяч, и совпадение значений в имеющихся столбцах не обязательно отражает полную картину состояния здоровья человека. Состояние здоровья человека может зависеть от множества других генов и факторов окружающей среды, которые не учтены в данном исследовании.

Кроме того, даже если последовательности генов идентичны, уровень экспрессии (активности) генов может различаться, что также влияет на риск развития заболевания. Не стоит также исключать возможность взаимодействия генов: гены взаимодействуют друг с другом, и комбинация различных генов может влиять на риск развития заболевания сильнее, чем каждый ген в отдельности.

Во второй серии экспериментов наибольшую точность показали метод k ближайших соседей (Accurasy 0,81) и метод опорных векторов (Accurasy 0,78). Это указывает на их высокую способность

**Таблица 2**  
Результаты оценки точности предсказания риска развития рака легких различными алгоритмами МО на тестовом наборе данных, обработанном с помощью метода TF-IDF

**Table 2**  
Results of assessing the accuracy of predicting the risk of developing lung cancer using various ML algorithms on a test data set processed using the TF-IDF method

Алгоритм МО	Accurasy	AUC&ROC кривая
Дерево решений	0,73	0,7104
Случайный лес	0,78	0,7651
Логистическая регрессия	0,77	0,7595
Метод опорных векторов	0,78	0,7751
Метод k ближайших соседей	0,81	0,7954
Количество строк в датасете	153	

правильно классифицировать случаи рака легких и здоровых людей.

Сравнительный анализ результатов исследования, представленных в первой таблице и во второй таблице, позволяет сделать ряд наблюдений и выводов:

1) в случае использования метода TF-IDF наблюдается повышение точности предсказания у метода к ближайших соседей (с 0,78 до 0,81).

2) Также во второй таблице алгоритмы «Дерево решений», «Метод опорных векторов» и «Метод к ближайших соседей» показывают улучшенные значения AUC по сравнению с первой таблицей. Наибольшее увеличение AUC наблюдается у метода опорных векторов (с 0,7719 до 0,7751) и метода к ближайших соседей (с 0,7606 до 0,7954). Это указывает на более высокую точность предсказаний моделей как следствие улучшения способности разделять классы здоровых и больных людей.

Метод TF-IDF, использованный во второй таблице, вероятно, помог глубже понять взаимосвязи между отдельными словами (генотипами) и их сочетаниями, учесть их вклад в общее содержание, что привело к улучшению точности и AUC.

Применение метода TF-IDF показало возможность его использования при работе с сложными биомедицинскими данными, включая обработку генетических вариантов. Он позволил получить хорошее качество работы модели, выделив значимые признаки в имеющемся массиве генетической информации.

## ЗАКЛЮЧЕНИЕ

В процессе проведения научного исследования были протестированы пять методов машинного обучения для оценки риска возникновения рака лег-

ких: дерево решений, случайный лес, логистическая регрессия, метод опорных векторов и метод к ближайших соседей, в двух вариантах предварительной обработки набора данных (с преобразованием по методу TF-IDF и без преобразования).

В результате метод к ближайших соседей в сочетании с обработкой данных по методу TF-IDF показал наиболее хорошую точность прогнозов на имеющемся наборе данных. Данный подход может быть предложен к использованию для профилактики заболевания раком у работников горнодобывающих предприятий в рамках регулярных профилактических медосмотров, а также при анализе рисков возникновения заболевания при приеме на работу.

Алгоритмы машинного обучения (такие как дерево решений, случайный лес и др.) представляют собой мощный инструмент для анализа медицинских данных. Они способны выявлять скрытые закономерности и определять ключевые признаки, значительно влияющие на вероятность возникновения заболевания.

Тем не менее, важно отметить, что рассмотренные методы использовались исключительно в контексте конкретной задачи и не являются универсальным решением для всех видов анализа медицинских данных. В зависимости от специфики задачи и предъявляемых требований к результатам могут потребоваться другие подходы или комбинации методов для достижения оптимальных результатов.

## Информация о финансировании и конфликте интересов

Работа выполнена в рамках государственного задания. Проект № 124041100077-1

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

## ЛИТЕРАТУРА / REFERENCES:

1. World Health Organization (WHO): Lung cancer URL: <https://www.who.int/ru/news-room/fact-sheets/detail/lung-cancer> (дата обращения: 26.01.2025)
2. Zabolevaemost onkologii i smertnost ot nee s 1997 goda //MNIОI im. P.A. Gertsena; obrabotka: «Esli byt tochnym», 2024. Russian (Заболелаемость онкологией и смертность от нее с 1997 года» //МНИОИ им. П.А. Герцена; обработка: «Если быть точным», 2024.) URL: [https://tochno.st/datasets/cancer\\_incidence](https://tochno.st/datasets/cancer_incidence) (дата обращения: 16.02.2025).
3. Geneticheskii pasport – osnova individualnoi i prediktivnoi meditsiny /Pod red. V.S. Baranova. SPb: OOO «Izdatelstvo N-L», 2009. 528 s. Russian (Генетический паспорт – основа индивидуальной и предиктивной медицины /Под ред. В.С. Баранова. СПб: ООО «Издательство Н-Л», 2009. 528 с.)
4. Akinin AA. Primenenie bol'shikh yazykovykh modelej dlya diagnostiki. *Scientific and Technical Volga region Bulletin*. 2024; 11: 138-140. Russian (Акинин А.А. Применение больших языковых моделей для диагностики //Научно-технический вестник Поволжья. 2024. № 11. С. 138-140.)
5. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. *Health Care Sci*. 2023; 2(4): 255-263. doi: 10.1002/hcs2.61
6. Primenenie bolshikh iazykovykh modelei v meditsine [Elektronnyi resurs] Webiomed. Russian (Применение больших языковых моделей в медицине [Электронный ресурс] Webiomed.) URL: <https://webiomed.ru/blog/primenenie-bolshikh-izykovykh-modelei-v-meditsine> (дата обращения: 25.01.2025)
7. Vasyaeva AS. Review of clustering algorithms for the purpose of processing medical data. *Algorithms, methods and systems for data processing*. 2017; 2(36): 17-22. Russian (Васяева Д.О., Белякова А.С. Обзор алгоритмов кластеризации с целью обработки медицинских данных //Алгоритмы, методы и системы обработки данных. 2017. Т. 2, № 36. С. 17-22.)
8. Kasyuk ST, Didenko GA, Stepanova OA. Cluster analysis of medical research data using the R language. *Modern science: current problems of theory and practice. Series: Natural and Technical Sciences*. 2021; 04/2: 23-32. Russian (Касюк С.Т., Диденко Г.А., Степанова О.А. Кластерный анализ данных медицинских исследований с использованием языка R //

- Современная наука: актуальные проблемы теории и практики. Серия: Естественные и Технические Науки. 2021. Т. 04/2. С. 23-32.) doi: 10.37882/2223-2966.2021.04-2.11
9. Bogatyr IN. Clustering algorithms in the analysis of medical data. *Scientific and technical bulletin of the Volga region*. 2023; 12: 586-588. Russian (Богатырь И.Н. Алгоритмы кластеризации в анализе медицинских данных //Научно-технический вестник Поволжья. 2023. № 12. С. 586-588.)
  10. Zimina EYu. Application of clustering methods in the development of a system for automated analysis of cardiac information. *New information technologies and systems (NITIS-2021): Collection of scientific articles based on the materials of the XVIII International Scientific and Technical Conference, Penza, November 24-26, 2021*. Penza, 2021. P. 63-66. Russian (Зими́на Е.Ю. Применение методов кластеризации при разработке системы автоматизированного анализа кардиологической информации //Новые информационные технологии и системы (НИТиС-2021): Сб. науч. статей по матер. XVIII Междунар. науч.-техн. конф., Пенза, 24-26 ноября 2021 года. Пенза, 2021. С. 63-66.)
  11. Serobabov AS, Denisova LA, Serobabova AL. Development of a system to support medical decision-making when prescribing treatment to a patient. *News of Tula State University. Technical Sciences*. 2023; 9: 321-325. Russian (Серобабов А.С., Денисова Л.А., Серобабова А.Л. Разработка системы поддержки принятия врачебных решений при назначении лечения пациенту //Известия Тульского государственного университета. Технические науки. 2023. № 9. С. 321-325.) doi: 10.24412/2071-6168-2023-9-321-322
  12. Al-Funjan A, Meziane F, Aspin R. Describing Pulmonary Nodules Using 3D Clustering. *Advanced Engineering Research (Rostov-on-Don)*. 2022; 22(3): 261-271. Russian (Аль-Фунжан А, Мезиан Ф, Аспин Р. Описание легочных узелков с помощью трехмерной кластеризации //Передовые инженерные исследования (Ростов-на-Дону). 2022. Т. 22, № 3. С. 261-271.) doi: 10.23947/2687-1653-2022-22-3-261-271
  13. Kotov SA, Starkov SO. Determination of the results of computed tomography studies of the chest organs with signs of COVID-19 using computer vision methods. *The future of nuclear energy – AtomFuture 2021: Abstracts of the XVII International Scientific and Practical Conference, Obninsk, November 22-23, 2021*. Obninsk, 2021. P. 180-182. Russian (Котов С.А., Старков С.О. Определение результатов исследований компьютерной томографии органов грудной клетки с признаками COVID-19 с помощью методов компьютерного зрения //Будущее атомной энергетики – AtomFuture 2021: Тезисы докладов XVII Междунар. науч.-практ. конф., Обнинск, 22-23 ноября 2021 года. Обнинск, 2021. С. 180-182.)
  14. Limanova NI, Morozov DA. Review of computer vision models for determining the presence of COVID-19 disease. *Eurasian Scientific Association*. 2021; 4-1(74): 57-59. Russian (Лиманова Н.И., Морозов Д.А. Обзор моделей компьютерного зрения для определения наличия заболевания COVID-19 //Евразийское Научное Объединение. 2021. Т. 4-1, № 74. С. 57-59.)
  15. Vinokurov MA, Mironov KO, Domonova EA, Romanuk TN, Popova AA, Akimkin VG. Comparison of machine learning methods for predicting cervical cancer risk based on genetic susceptibility. *Epidemiology and infectious diseases. Current issues*. 2024; 14(1): 77-82. Russian (Винокуров М.А., Миронов К.О., Домонова Э.А., Романюк Т.Н., Попова А.А., Акимкин В.Г. Сравнение методов машинного обучения для прогнозирования риска развития рака шейки матки на основе генетической предрасположенности //Эпидемиология и инфекционные болезни. Актуальные вопросы. 2024. Т. 14, № 1. С. 77-82.) doi: 10.18565/epidem.2024.14.1.77-82
  16. Minin AV, Katanakhova MV. Analysis of whole-exome sequencing data in Kuzbass patients with lung cancer: search for rare pathogenic variants. *Interdisciplinary approaches in biology, medicine and earth sciences: theoretical and applied aspects: Proceedings of the symposium of the XIX (LI) International scientific conference of students, graduate students and young scientists, Kemerovo, April 23, 2024*. Kemerovo, 2024. P. 116-119. Russian (Минин А.В., Катанахова М.В. Анализ данных полноэкзомного секвенирования у жителей Кузбасса больных раком легкого: поиск редких патогенных вариантов //Междисциплинарные подходы в биологии, медицине и науках о земле: теоретические и прикладные аспекты: Матер. симп. XIX (LI) Междунар. науч. конф. студентов, аспирантов и молодых ученых, Кемерово, 23 апреля 2024 года. Кемерово, 2024. С. 116-119.)
  17. Minin AV, Rkunov DS. Lung cancer in miners: genetic factors. *Development – 2021: Scientific electronic publication*. Kemerovo: FRC CCC SB RAS, 2021. P. 219-227. Russian (Минин А.В., Рекунов Д.С. Рак легкого у шахтеров: генетические факторы //Развитие – 2021: Научное электронное издание. Кемерово: ФИЦ УХ СО РАН, 2021. С. 219-227.) doi: 10.53650/9785902305620\_219
  18. Timofeeva AA, Minin AV, Bakanova ML, Zakharova YaA, Torgunakova AV, Prosekov AYU, et al. Polymorphism of genes encoding DNA repair enzymes in coal mine workers with lung cancer. *Journal of the Siberian Federal University. Series: Biology*. 2024; 17(1): 33-44. Russian (Тимофеева А.А., Минин А.В., Баканова М.Л., Захарова Я.А., Торгунакова А.В., Просеков А.Ю., и др.. Полиморфизм генов ферментов репарации ДНК у работников угольных шахт больных раком легкого //Журнал Сибирского федерального университета. Серия: Биология. 2024. Т. 17, № 1. С. 33-44.)
  19. Kulincha PV. Ispol'zovanie sluchajnogo lesa dlya klassifikatsii dannyh. *Molodoj uchenyj*. 2024; 24(523): 88-91. Russian (Кулинча П.В. Использование случайного леса для классификации данных //Молодой ученый. 2024. № 24(523). С. 88-91.)

**Сведения об авторах:**

КОЛЕСНИКОВА Татьяна Геннадьевна, канд. техн. наук, доцент кафедры цифровых технологий, КемГУ, г. Кемерово, Россия.  
E-mail: tgkolesnikova@yandex.ru

**Information about authors:**

KOLESNIKOVA Tatyana Gennadyevna, candidate of technical sciences, docent of the department of digital technologies, Kemerovo State University, Kemerovo, Russia. E-mail: tgkolesnikova@yandex.ru

ЧЕРНОУСОВ Михаил Евгеньевич, ассистент кафедры цифровых технологий, КемГУ, г. Кемерово, Россия. E-mail: chme@live.ru

МИНИНА Варвара Ивановна, доктор биол. наук, доцент, зав. кафедрой генетики и фундаментальной медицины, КемГУ; главный научный сотрудник лаборатории цитогенетики, ФИЦ УУХ СО РАН, г. Кемерово, Россия. E-mail: vminina@mail.ru

ВЫЛЕГЖАНИНА Алина Владимировна, канд. техн. наук, доцент кафедры цифровых технологий, КемГУ, г. Кемерово, Россия. E-mail: alina\_com@rambler.ru

МИНИН Артем Вячеславович, аспирант, КемГУ, г. Кемерово, Россия. E-mail: a.v.minin2000@mail.ru

БАКАНОВА Марина Леонидовна, канд. биол. наук, ст. науч. сотрудник лаборатории цитогенетики, ФИЦ УУХ СО РАН; доцент кафедры биологии с основами генетики и паразитологии, ФГБОУ ВО КемГМУ Минздрава России, г. Кемерово, Россия. E-mail: mari-bakano@yandex.ru

CHERNOUSOV Mikhail Evgenievich, assistant of the department of digital technologies, Kemerovo State University, Kemerovo, Russia. E-mail: chme@live.ru

MININA Varvara Ivanovna, doctor of biological sciences, docent, head of the department of genetics and fundamental medicine, Kemerovo State University; chief researcher of the laboratory of cytogenetics, FRC CCC SB RAS, Kemerovo, Russia. E-mail: vminina@mail.ru

VYLEGZHANINA Alina Vladimirovna, candidate of technical sciences, docent of the department of digital technologies, Kemerovo State University, Kemerovo, Russia. E-mail: alina\_com@rambler.ru

MININ Artem Vyacheslavovich, postgraduate student, Kemerovo State University, Kemerovo, Russia. E-mail: a.v.minin2000@mail.ru

BAKANOVA Marina Leonidovna, candidate of biological sciences, senior researcher of the laboratory of cytogenetics, FRC CCC SB RAS; docent of the department of biology with fundamentals of genetics and parasitology, Kemerovo State Medical University, Kemerovo, Russia. E-mail: mari-bakano@yandex.ru

**Корреспонденцию адресовать:** БАКАНОВА Марина Леонидовна, 650066, г. Кемерово, ФГБОУ ВО КемГМУ Минздрава России  
E-mail: mari-bakano@ya.ru