Фундаментальные проблемы современного материаловедения. 2024. Т. 21. № 3. С. 325–330
Fundamental'nye problemy sovremennogo materialovedenia (Basic Problems of Material Science (BPMS)). 2024; 3(21): 325–330

Научная статья
1.3.8. Физика конденсированного состояния (физико-математические науки)
УДК 004
doi: 10.25712/ASTU.1811-1416.2024.03.006

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ РАЗЛИЧНЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ПРЕДСКАЗАНИЯ СВОЙСТВ ЭКЗОГЕННЫХ ФЛУОРОФОРОВ

Михаил Андреевич Зарудских[1], Сергей Александрович Безносюк[2†]

[1, 2] Алтайский государственный университет, пр. Ленина, 61, 656049, Барнаул, Россия
[1] zarudskih@yandex.ru, https://orcid.org/0009-0007-5224-1665
[2] bsa1953@mail.ru†, https://orcid.org/0000-0002-4945-7197

Аннотация. В статье проанализирована эффективность использования различных моделей машинного обучения для предсказания спектральных свойств экзогенных флуорофоров, ключевых в диагностике онкозаболеваний. Исследуется применение алгоритмов ИИ для быстрого и экономически эффективного поиска новых флуорофоров, способствующих раннему выявлению рака. В статье оценивается эффективность различных моделей машинного обучения в предсказании свойств экзогенных флуорофоров, используемых в диагностике онкологических заболеваний. В работе исследуется применение алгоритмов искусственного интеллекта для быстрого поиска новых флуорофоров, способствующих раннему обнаружению рака. Особое внимание уделено оптической биопсии как неинвазивному методу исследования тканей для ранней диагностики патологий. В статье обобщаются данные из базы данных PubChem и GeoMcNamara и анализируются молекулярные свойства флуорофоров и их спектральные характеристики. Используя модели машинного обучения, такие как линейная регрессия, метод опорных векторов, случайный лес и XGBoost, получены результаты предсказания длины волны излучения для образцов флуорофоров. Результаты обучения и тестирования моделей свидетельствуют о высокой точности работы XGBoost и Random Forest. Исследование подчеркивает важность разработки эффективных флуорофоров для ранней диагностики рака и представляет модели машинного обучения в качестве инструментов для обработки и анализа данных в этой области, что позволяет акцентировать внимание на перспективности и применимости прогрессивных методов исследования в онкологии и медицинской химии.

Ключевые слова: компьютерное моделирование, машинное обучение, искусственный интеллект, экзогенные флуорофоры, медицинская химия, экзогенные флуорофоры.

Для цитирования: Зарудских М.А., Безносюк С.А. Исследование эффективности различных моделей машинного обучения в задаче предсказания свойств экзогенных флуорофоров // Фундаментальные проблемы современного материаловедения. 2024. Т. 21, № 3. С. 325–330. doi: 10.25712/ASTU.1811-1416.2024.03.006.

Original article

INVESTIGATION OF THE EFFECTIVENESS OF VARIOUS MACHINE LEARNING MODELS IN PREDICTING THE PROPERTIES OF EXOGENOUS FLUOROPHORES

Mikhail A. Zarudskikh[1], Sergey A. Beznosyuk[2†]

[1, 2] Altai State University, Lenina Pr., 61, Barnaul, 656049, Russia
[1] zarudskih@yandex.ru, https://orcid.org/0009-0007-5224-1665
[2] bsa1953@mail.ru†, https://orcid.org/0000-0002-4945-7197

Abstract. The article analyzes the effectiveness of using various machine learning models to predict the spectral properties of exogenous fluorophores, which are key in the diagnosis of cancer. The application of AI algorithms for the rapid and cost-effective search for new fluorophores contributing to the early detection of cancer is being investigated. The article evaluates the effectiveness of various machine learning models in predicting the properties of exogenous fluorophores used in the diagnosis of cancer. The paper explores the use of artificial intelligence algo-

rithms for the rapid search for new fluorophores that contribute to the early detection of cancer. Special attention is paid to optical biopsy as a non-invasive method of tissue examination for early diagnosis of pathologies. The article summarizes data from the PubChem and GeoMcNamara databases and analyzes the molecular properties of fluorophores and their spectral characteristics. Using machine learning models such as linear regression, the support vector machine method, random forest and XGBoost, the results of radiation wavelength prediction for fluorophore samples were obtained. The results of training and testing of models indicate the high accuracy of the work of XGBoost and Random Forest. The study highlights the importance of developing effective fluorophores for early cancer diagnosis and presents machine learning models as tools for processing and analyzing data in this area, which allows us to focus on the prospects and applicability of advanced research methods in oncology and medical chemistry.

Keywords: computer modeling, machine learning, artificial intelligence, exogenous fluorophores, medical chemistry, exogenous fluorophores.

For citation: Zarudskikh, M. A. & Beznosyuk, S. A. (2024). Investigation of the effectiveness of various machine learning models in predicting the properties of exogenous fluorophores. *Fundamental'nye problemy sovremennogo materialovedenia (Basic Problems of Material Science (BPMS))*, 21(3), 325–330. (In Russ.). doi: 10.25712/ASTU.1811-1416.2024.03.006.

Введение

В современном мире за последние десятилетия изменение экологической обстановки и образа жизни человека привели к распространению различных заболеваний, среди которых онкологические заболевания представляют основную угрозу для жизни и здоровья населения по нескольким причинам: агрессивность и непредсказуемый характер протекания болезни, трудности в диагностике и лечений, что делает актуальным поиск новых методов диагностики и терапии этой группы заболеваний.

Среди таких методов наиболее широкое применение в биомедицинских исследованиях живых тканей при контроле и эффективности терапии получают оптические методы, объединённые под общим названием «оптическая биопсия» [1-3]. Одним из ключевых преимуществ оптической биопсии является ее неинвазивный характер, позволяющий проводить исследования без причинения ущерба организму. Это открывает возможность быстро и в режиме реального времени изучать метаболические процессы внутри живых тканей, выявлять микроморфологические и биохимические изменения. Такой подход позволяет не только диагностировать патологии, но и отслеживать их динамику на ранних стадиях развития, что существенно повышает эффективность лечения. Во всех патологиях имеются определенные изменения в тканях, которые отличаются от здоровых образцов.

Используя методы люминесцентного анализа в оптической биопсии, специалисты могут выявить эти различия и провести диагностику заболеваний еще на ранних стадиях, когда изменения еще не видны невооруженным глазом. Благодаря возможности неразрушающего контроля состояния биологических объектов, люминесцентный анализ становится неоценимым инструментом как для научных исследований, так и для медицинской практики [4].

В области визуализации патологических очагов с использованием люминесцентного анализа ключевую роль играют методы, основанные на выявлении различий в флуоресцентном излучении между патологическими очагами и окружающими нормальными тканями при освещении их светом определенной длины в УФ и видимом диапазонах спектра. Эти различия описываются термином "флуоресцентный контраст", который может иметь эндогенное, которое возникает при использовании эндогенных флуорофоров, или экзогенное происхождение, которое возникает при использовании экзогенных флуорофоров соответственно. Эндогенные флуорофоры – это биологические вещества тканей, способные к флуоресценции [5]. К экзогенным флуорофорам относятся экзогенные красители или флуоресцентные маркеры, которые обладают лучшими спектральными свойствами, по сравнению с природными макромолекулами.

Основными спектральными характеристиками экзогенных флуорофоров являются параметры флуоресценции: интенсивность флуоресценции, спектр излучения, спектр возбуждения, квантовый выход флуоресценции. Изменение этих параметров несет определенную взаимно дополнительную информацию об флуорофоре и его микроокружении [6].

Поиск эффективных экзогенных флуорофоров с заданными свойствами становится

важным направлением в современной онкологии. Обнаружение раковых заболеваний на ранних стадиях с помощью высокоспецифичных флуорофоров способствует своевременной диагностике и началу лечения, что повышает эффективность и успешный исход терапии.

Цель настоящего исследования сравнение эффективности различных моделей машинного обучения в предсказании спектральных свойств экзогенных флуорофоров на основании их молекулярных свойств.

Методы машинного обучения

Быстрый поиск новых экзогенных флуорофоров с заданными свойствами, и минимальными финансовыми затратами представляется возможным благодаря современным возможностям цифровой индустрии, а именно развитию возможностей искусственного интеллекта (ИИ), и одного из его направлений – машинного обучения (МО) [7]. В настоящий момент методы машинного обучения находят широкое применение в экономике, производстве, медицине, и других областях жизнедеятельности человека, где необходима автоматизация процесса, и требуется работать с большими объёмами данных [8-10]. Главная идея машинного обучения заключается в том, чтобы алгоритмы и модели МО могли обнаруживать закономерности и особенности в данных, и использовать эту информацию для принятия решения или предсказания будущих результатов. Основы и принципы работы МО включают следующие компоненты [11]:

1. Данные. МО – это область, где ключевую роль играют данные, на которых обучается модель. Данные включают в себя разнообразную информацию, будь то числовые значения, текст, изображения или звуковые данные. Качество и разнообразие данных имеют огромное значение для успешного обучения модели.

2. Алгоритмы и модели. МО использует различные алгоритмы и модели для обучения на данных для решения разнообразных задач, среди которых можно выделить задачи классификации и регрессии. Задачи классификации – это процесс поиска или выбор модели, которая помогает разделить данные на категориальные классы, то есть дискретные значения. Задачи регрессии – это поиск модели для разделения данных на непрерывные реальные значения вместо использования классов и дискретных значений. Для решения задачи регрессии применяются различные модели машинного обучения, например, линейная регрессия (Linear Regression) – это один из самых основных и понятных методов в статистике и машинном обучении. Её преимущество заключается в простоте реализации и интерпретируемости результатов. Эта модель хорошо подходит для задач, где ожидается линейная зависимость между независимыми и зависимыми переменными. Модель линейной регрессии также используется как базовый стандарт при оценке производительности более сложных моделей; метод опорных векторов (Support Vector Machines) – это мощный алгоритм обучения, который можно использовать как для классификации, так и для регрессии. Основное преимущество SVM заключается в его способности обрабатывать даже небольшие и сложные наборы данных с большим разнообразием признаков за счёт использования разных ядер. SVM идеально подходит для задач, где необходима высокая точность и где данные являются разделимыми в многомерном пространстве; случайный лес (Random Forest) – это ансамблевый метод обучения, который строит множество деревьев решений и объединяет их, чтобы получить более стабильное и точное предсказание. Этот метод отличается высокой точностью и устойчивостью к переобучению, благодаря случайной выборке обучающих подмножеств и признаков. Случайный лес хорошо подходит для обработки больших данных с большим количеством признаков, он также позволяет оценить важность каждого признака для модели. XGBoost (eXtreme Gradient Boosting) – это один из видов алгоритма градиентного бустинга, который использован для работы с большими объемами данных и достижения высокой производительности предсказаний. XGBoost обеспечивает автоматическую обработку пропущенных значений, регуляризацию для избегания переобучения и эффективную масштабируемость, которая делает его одним из предпочтительных инструментов среди исследователей и практиков.

3. Обучение и обучающий набор данных. Для обеспечения высокого качества и надежности данных, используемых в процессе обучения моделей машинного обучения, требуется осуществить процесс сбора и подготовки соответствующих наборов данных. Этот процесс включает в себя несколько ключевых этапов. Во-первых, необходимо провести сбор данных

из различных источников, учитывая специфику проблемы, которую требуется решить моделью. Важно уделить внимание качеству и полноте данных, чтобы исключить искажения в процессе обучения. Во-вторых, подготовка данных является не менее важным этапом. Очистка данных от ошибок, пропусков и выбросов является необходимым условием для получения корректных результатов. Дополнительно, необходимо преобразовать данные в удобный формат, который может быть использован алгоритмами машинного обучения. Такое преобразование позволит эффективно работать с данными и извлечь из них максимальную информацию. Третий этап включает разделение данных на обучающую и тестовую выборки. Это позволяет оценить качество обученной модели на новых данных и избежать переобучения. Обучающая выборка используется для обучения модели, а тестовая – для оценки ее эффективности и обобщающей способности. Следует отметить, что качество данных имеет прямое влияние на результаты моделей машинного обучения. Поэтому весь процесс от сбора до подготовки данных является критически важным для успешного построения моделей и получения надежных результатов на практике.

4. Оценка и проверка модели. После завершения процесса обучения модели машинного обучения следует провести анализ ее производительности и убедиться в ее способности обобщать знания на новых данных. Для этого применяются разнообразные метрики оценки, которые позволяют количественно оценить эффективность работы модели. В случае решения задач регрессии применяются такие метрики, как среднеквадратичная ошибка (Mean Squared Error (MSE)), коэффициент детерминации (R2-Score), среднее абсолютное отклонение в процентах (Mean Absolute Percentage Error (MAPE)) и др. Также одной из широко используемых метрик является точность (accuracy), показывающая долю правильных предсказаний, сделанных моделью относительно общего числа предсказаний. Точность полезна в случае сбалансированных классов, но может оказаться недостаточной при несбалансированности классов. Выбор наиболее подходящих метрик оценки и методов проверки модели зависит от конкретной задачи и характеристик данных. Метрики оценки помогают понять, насколько эффективно работает модель, а проверка модели обеспечивает уверенность в ее способности обобщать знания на новые данные.

5. Применение и развертывание модели: Завершив обучение и оценку модели, она может быть интегрирована в рабочую среду или систему, чтобы модель могла генерировать прогнозы или принимать решения на новых данных. Этот этап является крайне важным, поскольку именно здесь модель начинает использоваться на практике для решения конкретных задач.

Результаты и их обсуждение

В настоящей работе использованы данные из базы данных химических соединений и смесей PubChem [12]. PubChem содержит в себе информацию о разнообразных свойствах химических соединений, такие как молекулярный вес, коэффициент липофильности, количество атомов и т.д. В описываемом исследовании для предсказания длинны волны излучения флуорофоров были отобраны следующие молекулярные свойства: заряд молекулы, количество атомов, количество доноров водородной связи, количество акцепторов водородной связи, молекулярный вес, топологическая плотность полярной поверхности.

Данные о длине волны излучения были получены из базы данных доктора философских наук Джорджа Макнамары – GeoMcNamara, которая содержит в себе данные оптических свойств различных светоактивных красителях (Alexa Fluor, ATTO, Brilliant, Chromeo, dye), группы аминокислот, которые обладают свойствами флуоресценции (триптофан, тирозин), а также бактериохлоринов. Используемые базы данных находятся в открытых источниках, и имеют свободны доступ.

Всего из 200 соединений, представленных в базе данных GeoMcNamara, после агрегации, и обработки, был получен набор данных из 119 соединений.

Для обучения выбраны следующие популярные и широко используемые модели машинного обучения: линейная регрессия (Linear Regression); метод опорных векторов (Support Vector Machines); случайный лес Random Forest и XGBoost.

Для оценки результатов использовалась метрика MSE. Результаты прогнозирования приведены на рисунке 1 и таблице 1.

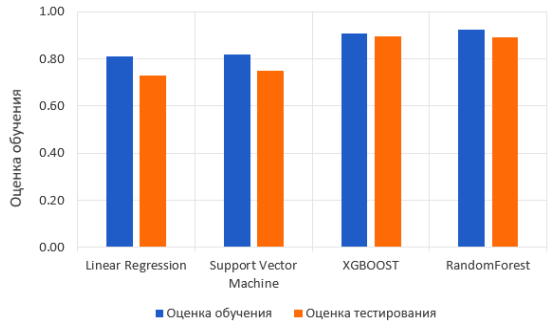

Рис.1. Результаты обучения и тестирования моделей машинного обучения

Fig.1. The results of training and testing of machine learning models

Таблица 1. Результат предсказания спектра излучения для разных моделей на основании метрики MSE

Table 1. The result of the prediction of the radiation spectrum for different models based on the MSE metric

Модель	Оценка обучения	Оценка тестирования
Linear Regressiom	0,81	0,73
Support Vector Machine	0,82	0,75
XGBoost	0,91	0,90
Random Forest	0,92	0,89

При анализе результатов машинного обучения на основе оценки обучения и тестирования видно, что XGBoost и Random Forest показывают достаточно хорошие результаты обучения и тестирования, что свидетельствует о том, что модели хорошо адаптируются к обучению, и способны эффективно извлекать закономерности из данных. Однако из-за малого размера набора данных результаты могут варьироваться при разных разбиениях на тренировочную и тестовую выборки.

Выводы

Поиск экзогенных флуорофоров с заданными свойствами является одной из наиболее перспективных направлений исследования в области оптической диопсии и медицинской химии в целом. Развитие искусственного интеллекта и цифровой индустрии в настоящее время предоставляют достаточно возможностей и вычислительных мощностей для теоретических расчётов химических соединений с различными заданными свойствами.

Рассмотренные в данной работе модели машинного обучения могут быть использованы для предсказания свойств химических соединений, что позволит снизить затраты времени и ресурсов.

Список литературы

1. Bigio I., Mourant J. Ultraviolet and visible spectroscopies for tissue diagnostics: fluorescence spectroscopy and elasticscattering spectroscopy // Phys. Med. Biol. 1997. N 42. P. 803–814.

2. Лощенов В.Б., Стратонников А.А. Физические основы флуоресцентной диагностики и фотодинамической терапии // В сб. трудов МИФИ. 2000. Т. 4. С. 53–54.

3. Лычковская Е.В., Вайс Е.Ф., Салмина А.Б, Салмин В.В. Оптическая биопсия с использованием экзогенных флуорофоров // Сибирское медицинское обозрение. 2015. Т. 92, № 2. С. 5–14.

4. Лисовский В.А., Щедрунов В.В., Барский И.Я. Люминесцентный анализ в гастроэнтерологии. Л.: Наука, 1984. 236 с.

5. Лакович Дж. Основы флуоресцентной спектроскопии. М.: Мир, 1986. 496 с.

6. Белков А.В., Скрипник А.В. Лазерные биомедицинские технологии. Спб.: СпбГУ Итмо, 2008. 116 с.

7. Бритвина П.В. Тенденции развития машинного обучения и его влияние // Вестник науки. 2023. Т. 65, № 8. С. 101–103.

8. Еслямгалиева, А. М. Современные методы диагностики онкологических заболеваний слизистой полости рта // Актуальные научные исследования в современном мире. 2020. Т. 68, № 12–9. С. 53–57.

9. Романова Е.А. Машинное обучение в экономике и производстве // В сб. материалов научной конференции «XLIX Огарёвские чтения», Саранск, 07–13 декабря 2020 года. Часть 3. Саранск: Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва, 2021. С. 691–697.

10. Свищев А.В., Морошкин Н. А., Ефремова С. Г. Исследование целесообразности применения технологий искусственного интеллекта в промышленности // Актуальные научные исследования в современном мире. 2021. Т. 73, № 5–8. С. 245–251.

11. Бабкина Е.А., Гаев Л.В. Искусственный интеллект и машинное обучение // Инновационные исследования: проблемы внедрения. 2023. С. 155.

12. Kim S., Chen J., Cheng T., Gindulyte A., He J. et al. PubChem in 2021: new data content and improved web interfaces // Nucleic acids research. 2021. V. 49, N D1. P. D1388–D1395.

Информация об авторах

М. А. Зарудских – аспирант кафедры физической и неорганической химии Алтайского государственного университета.

С. А. Безносюк – доктор физико-математических наук, профессор, заведующий кафедрой физической и неорганической химии Алтайского государственного университета.

References

1. Bigio, I. & Mourant, J. (1997). Ultraviolet and visible spectroscopies for tissue diagnostics: fluorescence spectroscopy and elasticscattering spectroscopy. *Phys. Med. Biol.*, (42), 803–814.

2. Loshenov, V. B. & Stratonnikov, A. A. (2000). Physical foundations of fluorescent diagnostics and photodynamic therapy. *Proceedings of MEPhI*, *4*, 53–54. (In Russ.).

3. Lychkovskaya, E. V., Weiss, E. F., Salmina, A. B. & Salmin, V. V. (2015). Optical biopsy using exogenous fluorophores. *Siberian Medical Review*, *92*(2), 5–14. (In Russ.).

4. Lisovsky, V. A., Shchedrunov, V. V. & Barsky, I. Ya. (1984). Luminescent analysis in gastroenterology. L.: Nauka. P. 236. (In Russ.).

5. Lakovich, J. (1986). Fundamentals of fluorescence spectroscopy. M.: Mir. P. 496. (In Russ.).

6. Belkov, A. V. & Skripnik, A. V. (2008). Laser biomedical technologies. St. Petersburg: St. Petersburg State University Itmo. P. 116. (In Russ.).

7. Britvina, P. V. 2023. Trends in the development of machine learning and its impact. *Bulletin of Science*, *65*(8), 101–103. (In Russ.).

8. Eslyamgalieva, A. M. (2020). Modern methods of diagnosis of oncological diseases of the oral mucosa. *Current scientific research in the modern world*, *68*(12–9), 53–57. (In Russ.).

9. Romanova, E. A. (2021). Machine learning in economics and production. Proceedings of the scientific conference «XLIX Ogarev Readings», Saransk, December 07–13, 2020. Part 3. Saransk: National Research Mordovian State University named after. N.P. Ogaryova, 691–697.

10. Svishchev, A. V., Moroshkin, N. A. & Efremova, S. G. (2021). Research into the feasibility of using artificial intelligence technologies in industry. *Current scientific research in the modern world*, *73*(5–8), 245–251. (In Russ.).

11. Babkina, E. A. & Gaev, L. V. (2023). Artificial intelligence and machine learning. *Innovative research: problems of implementation*, 155. (In Russ.).

12. Kim, S., Chen, J., Cheng, T., Gindulyte, A. & He, J. et al. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, *49*(1), D1388–D1395.

Information about the authors

M. A. Zarudskikh – Postgraduate Student at the Department of Physical and Inorganic Chemistry of the Altai State University.

S. A. Beznosyuk – Doctor of Physical and Mathematical Sciences, Professor, Head of the Department of Physical and Inorganic Chemistry, Altai State University.

Авторы заявляют об отсутствии конфликта интересов.
The authors declare that there is no conflict of interest.

Статья поступила в редакцию 08.06.2024; одобрена после рецензирования 17.07.2024; принята к публикации 01.08.2024.
The article was received by the editorial board on 08 June 24; approved after reviewing 17 July 24; accepted for publication 01 Aug. 24.