Федеральное государственное бюджетное образовательное

учреждение высшего образования

Московский государственный университет имени М.В. Ломоносова

Факультет биоинженерии и биоинформатики

УТВЕРЖДАЮ

И.о.декана факультета

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / А.А. Замятнин /

«\_\_» \_\_\_\_\_\_\_\_\_\_2023 г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**Межфакультетский курс «Применение методов искусственного интеллекта в анализе биомедицинских данных»**

Уровень высшего образования:

**Специалитет**

Направление подготовки (специальность) высшего образования

**06.05.01 Биоинженерия и биоинформатики**

Направленность (профиль) программы

**Фундаментальная биоинженерия и биоинформатика**

Форма обучения:

**очная**

Рабочая программа рассмотрена и одобрена

Учебно-методической комиссией факультета

(протокол №1 от 30 апреля 2021 г)

Москва 2023

**Аннотация программы**

В последние десятилетия мы стали свидетелями бурного развития и значительных достижений в области методов искусственного интеллекта (ИИ). Технологии ИИ уже успешно применяются во многих сферах науки и технологии. Важной областью применения новых методов ИИ является биология и медицина, где анализ больших объемов данных является одной из серьезных пока не решенных проблем.

Методы ИИ, включая машинное обучение и глубокое обучение, способны обрабатывать и анализировать большие объемы данных с высокой точностью и скоростью. Это позволяет выявлять сложные закономерности и зависимости, недоступные для традиционных статистических методов. ИИ способен анализировать геномные и протеомные данные, выявляя связи между генетическими маркерами и определенными заболеваниями. Это открывает путь к разработке персонализированных лечебных стратегий и целенаправленной терапии.

Использование ИИ в процессе поиска и разработки новых лекарственных средств может значительно сократить время и затраты. Алгоритмы ИИ могут предсказывать взаимодействие молекул, способствуя открытию новых перспективных соединений-кандидатов для разработки терапевтических средств. ИИ-алгоритмы, обученные на больших наборах медицинских изображений, способны с высокой точностью диагностировать различные заболевания, включая рак, зачастую превосходя в этом специалистов. Анализ биомедицинских данных с помощью ИИ в будущем может позволить разрабатывать индивидуализированные подходы к лечению, учитывая генетические особенности и образ жизни пациента.

Курс предоставит студентам уникальную возможность погрузиться в мир биоинформатики, изучая передовые методы ИИ и их приложения в биомедицинском анализе. Студенты получат базу для дальнейшего применения методов ИИ в своих исследований, расширят кругозор, получат знания, необходимые для вклада в будущее биомедицины и биологии.

**1. Место дисциплины (модуля) в структуре ОПОП ВО:**

**Цель курса:** Курс направлен на формирование у студентов глубокого понимания того, как методы ИИ могут быть применены для решения сложных задач в области биомедицинских наук, и подготовку их к будущим исследованиям и карьере в этой быстрорастущей области.

**Задачи курса:**

- Ознакомить слушателей с основами методов искусственного интеллекта

- Показать сильные стороны и ограничения этих методов

- Дать информацию об основных важных прорывах в современной биологии и медицине, сделанных с помощью этих методов в последние годы.

**2. Входные требования для освоения дисциплины (модуля), предварительные условия (если есть):**

Линейная алгебра: понимание основных концепций, таких как векторы, матрицы, операции с матрицами, собственные значения и векторы.

Математический анализ: основы дифференциального и интегрального исчисления, включая производные, интегралы, пределы и ряды.

Теория вероятностей и математическая статистика: основные принципы вероятности, случайные переменные, распределения вероятностей, математическое ожидание и дисперсия, основы статистического вывода.

Основы информатики и программирования: навыки программирования на уровне, достаточном для работы с данными и реализации алгоритмов. Предпочтительно знание языков программирования, используемых в анализе данных, таких как Python или R.

Структуры данных и алгоритмы: понимание основных структур данных (списки, стеки, очереди, деревья, графы) и алгоритмов (сортировка, поиск, основные алгоритмы на графах).

Предварительные знания в области биологии и медицины: хотя это не строгое требование, базовое понимание биологических и медицинских терминов и принципов будет полезным для лучшего понимания контекста и применения ИИ в биомедицине. Для успешного прохождения курса достаточно базовых знаний биологии и химии в рамках школьной программы.

**3. Результаты обучения по дисциплине (модулю):**

|  |
| --- |
| **Планируемые результаты обучения по дисциплине (модулю)** |
| **Знать:**Основные методы искусственного интеллекта, включая машинное обучение и глубокое обучение.Ключевые подходы к обработке и анализу больших объемов биомедицинских данных, включая геномные и протеомные данные.Современные тенденции и прорывы в применении ИИ в биологии и медицине, включая разработку лекарств, диагностику заболеваний и персонализированную медицину.**Уметь:**Применять методы машинного обучения и глубокого обучения для анализа и интерпретации биомедицинских данных.Использовать алгоритмы ИИ для решения конкретных задач в области геномики, протеомики и медицинской диагностики.Критически анализировать результаты, полученные с помощью методов ИИ, и оценивать их надежность и применимость в биомедицинских исследованиях.**Владеть:**Навыками работы с различными типами биомедицинских данных, включая их предварительную обработку, нормализацию и анализ.Умением взаимодействовать в мультидисциплинарных командах, объединяющих специалистов в области биологии, медицины и информатики.Навыками критического мышления и научного анализа для оценки научных исследований и разработки новых исследовательских проектов в области биоинформатики и медицины. |

**4. Объем дисциплины (модуля)** составляет 1 з.е.

**5. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий:**

Раздел 1. Введение: кратко о возможностях методов ИИ и о типах задач, которые могут быть с их помощью решены, но плохо решаются иными методами. Нейросети. Пайплайн машинного обучения на примере простой задачи. Что такое признак, выборка. Переобучение. Обучение и тестирование. Метрики качества: MSE. Как происходит обучение “на пальцах”. Нейросети - как устроено обучение нейросети. Пример применения: Конволюционные нейросети. Медицинские и биологические изображения

Раздел 2. Как устроены сверточные нейросети? Работа свертки в применении к изображениям. Представление изображений в памяти компьютера. Работа с медицинскими изображениями. Возможные биологические задачи на изображениях. Задача сегментации и архитектура U-Net.

Раздел 3. Модели для предсказания геномных треков и экспрессии по последовательности ДНК.

Графовые сверточные сети (message passing). Отбор и генерация веществ с заданными свойствами. Применение методов искусственного интеллекта для разработки новых лекарств и поиска новых способов применения уже существующих.

Раздел 4. Работа с текстом. Токенизация. Проблема представления текстов в памяти компьютера. Эмбеддинги слов. Word2Vec. Специфика научных и биологических данных: задача суммаризации текстов, задача Named Entity Recognition и т.д. Большие языковые модели, их применение в области биологии и медицины.

Раздел 5. RNN. Внимание в RNN. Трансформеры. Идея слоёв внимания. BERT. BioBERT. GPT. ChatGPT. Предсказание структуры и свойств белков. AlphaFold2. Дизайн белков с заданными свойствами. Применение машинного обучения в транскриптомике. Self-supervised подходы в биологии. Мультимодальные данные.

**5.1. Структура дисциплины (модуля) по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий (в строгом соответствии с учебным планом)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Наименование разделов и тем дисциплины (модуля),****Форма промежуточной аттестации по дисциплине (модулю)** | **Номинальные трудозатраты обучающегося**  | **Всего академических часов** | **Форма текущего контроля успеваемости\*** *(наименование)* |
| **Контактная работа (работа во взаимодействии с преподавателем)** **Виды контактной работы, академические часы** | **Самостоятельная работа обучающегося,****академические часы** |
| **Занятия лекционного типа** | **Занятия семинарского типа** |
| *Раздел 1.*  |  |  |  |  |  |
| **Тема 1.**  Введение: кратко о возможностях методов ИИ и о типах задач, которые могут быть с их помощью решены, но плохо решаются иными методами.   | 2 ак. часа |  | 1 ак. час |  |  |
| **Тема 2.**  Нейросети. Пайплайн машинного обучения на примере простой задачи. Что такое признак, выборка. Переобучение. Обучение и тестирование. Метрики качества: MSE. Как происходит обучение “на пальцах”.  | 2 ак. часа |  | 1 ак. час |  |  |
| **Тема 3.** Нейросети - как устроено обучение нейросети. Пример применения: Конволюционные нейросети. Медицинские и биологические изображения | 2 ак. часа |  | 1 ак. час |  |  |
| *Раздел 2.* |  |  |  |  |  |
| **Тема 1.**  Как устроены сверточные нейросети? Работа свертки в применении к изображениям. Представление изображений в памяти компьютера.  | 2 ак. часа |  | 1 ак. час |  |  |
| **Тема 2.** Работа с медицинскими изображениями. Возможные биологические задачи на изображениях. Задача сегментации и архитектура U-Net. | 2 ак. часа |  | 1 ак. час |  |  |
| *Раздел 3.* |  |  |  |  |  |
| **Тема 1.**  Модели для предсказания геномных треков и экспрессии по последовательности ДНК. | 2 ак. часа |  | 1 ак. час |  |  |
| **Тема 2.**  Графовые сверточные сети (message passing). Отбор и генерация веществ с заданными свойствами.  | 2 ак. часа |  | 1 ак. час |  |  |
| **Тема 3.** Применение методов искусственного интеллекта для разработки новых лекарств и поиска новых способов применения уже существующих. | 2 ак. часа |  | 1 ак. час |  |  |
| *Раздел 4.* |  |  |  |  |  |
| **Тема 1.**  Работа с текстом. Токенизация. Проблема представления текстов в памяти компьютера. Эмбеддинги слов. Word2Vec.  | 2 ак. часа |  | 1 ак. час |  |  |
| **Тема 2.**  Специфика научных и биологических данных: задача суммаризации текстов, задача Named Entity Recognition и т.д. | 2 ак. часа |  | 1 ак. час |  |  |
| **Тема 3.** Большие языковые модели, их применение в области биологии и медицины.. | 2 ак. часа |  | 1 ак. час |  |  |
| *Раздел 5* |  |  |  |  |  |
| **Тема 1.**  RNN. Внимание в RNN. Трансформеры. Идея слоёв внимания. BERT. BioBERT. GPT. ChatGPT. Предсказание структуры и свойств белков. AlphaFold2. Дизайн белков с заданными свойствами. Применение машинного обучения в транскриптомике. Self-supervised подходы в биологии. Мультимодальные данные.  | 2 ак. часа |  | 1 ак. час |  |  |
| Другие виды самостоятельной работы (при наличии): | — | — |  |  |  |
| Промежуточная аттестация (зачет) |  | 2 ак часа |  |  | — |
| **Итого** | 24 ак.часов | 12 ак.часов | **36 ак.часов** |  |

|  |
| --- |
| **5.2. Содержание разделов (тем) дисциплины** |
| **№ п/п** | *Наименование разделов (тем) дисциплины* | **Содержание разделов (тем) дисциплин** |
|  | *Раздел 1.*  |  |
| 1 | **Тема 1.**  Введение: кратко о возможностях методов ИИ и о типах задач, которые могут быть с их помощью решены, но плохо решаются иными методами.  | Обзор возможностей методов искусственного интеллекта, особенно тех, которые превосходят традиционные подходы. Рассмотрение типов задач, решаемых с помощью ИИ, с акцентом на уникальные свойства и преимущества этих методов. |
| 2 | **Тема 2.**  Нейросети. Пайплайн машинного обучения на примере простой задачи. Что такое признак, выборка. Переобучение. Обучение и тестирование. Метрики качества: MSE. Как происходит обучение “на пальцах”.  | Введение в нейросети и пайплайн машинного обучения через решение простой задачи. Объяснение понятий признака, выборки, переобучения, обучения и тестирования. Рассмотрение метрик качества, включая MSE, и основы процесса обучения. |
| 3 | **Тема 3.** Нейросети - как устроено обучение нейросети. Пример применения: Конволюционные нейросети. Медицинские и биологические изображения | Детальное изучение процесса обучения нейросетей. Примеры применения конволюционных нейросетей в анализе медицинских и биологических изображений. |
|  | *Раздел 2.* |  |
| 4 | **Тема 1.**  Как устроены сверточные нейросети? Работа свертки в применении к изображениям. Представление изображений в памяти компьютера.  | Исследование структуры сверточных нейросетей, механизма работы свертки в применении к изображениям и представления изображений в памяти компьютера. |
| 5 | **Тема 2.** Работа с медицинскими изображениями. Возможные биологические задачи на изображениях. Задача сегментации и архитектура U-Net. | Обзор методов работы с медицинскими изображениями. Рассмотрение биологических задач на основе изображений, включая сегментацию и архитектуру U-Net. |
|  | *Раздел 3.* |  |
| 6 | **Тема 1.**  Модели для предсказания геномных треков и экспрессии по последовательности ДНК. | В этой лекции исследуются модели ИИ, используемые для предсказания геномных треков и экспрессии генов из последовательностей ДНК. Обсуждаются различные подходы и методы, включая последние достижения в этой области. Подчеркивается важность точности и способности моделей обрабатывать большие объемы данных. |

|  |  |  |
| --- | --- | --- |
| 7 | **Тема 2.**  Графовые сверточные сети (message passing). Отбор и генерация веществ с заданными свойствами.  | Введение в концепции графовых сверточных сетей и их использование в биоинформатике и химии. Обсуждаются принципы работы и применение этих сетей в отборе и генерации веществ с заданными свойствами. Примеры из реальной практики демонстрируют их потенциал в открытии новых лекарственных средств. |
| 8 | **Тема 3.** Применение методов искусственного интеллекта для разработки новых лекарств и поиска новых способов применения уже существующих. | Обзор использования методов ИИ в фармацевтике для разработки новых лекарств и поиска новых способов применения существующих препаратов. Обсуждаются успехи и вызовы, а также перспективы ИИ в ускорении и повышении эффективности фармацевтических исследований. |
|  | *Раздел 4.* |  |
| 9 | **Тема 1.**  Работа с текстом. Токенизация. Проблема представления текстов в памяти компьютера. Эмбеддинги слов. Word2Vec.  | Глубокий анализ методов обработки текста в контексте ИИ, включая токенизацию и создание эмбеддингов слов. Рассматриваются различные подходы, такие как Word2Vec, и их применение в анализе текстовых данных. Обсуждаются проблемы представления текстов и способы их решения с помощью современных технологий ИИ. |
| 10 | **Тема 2.**  Специфика научных и биологических данных: задача суммаризации текстов, задача Named Entity Recognition и т.д. | Обсуждаются особенности и вызовы, связанные с обработкой научных и биологических текстов. Рассматриваются задачи суммаризации текстов, распознавание именованных сущностей (NER) и другие специфические задачи в области научных данных. Подчеркивается важность точности и адаптации методов ИИ к специфике научной лексики. |
| 11 | **Тема 3.** Большие языковые модели, их применение в области биологии и медицины.. | Исследование применения больших языковых моделей в биологии и медицине. Обсуждаются последние достижения в этой области, включая использование моделей, таких как BERT, GPT и других, для анализа и интерпретации научных текстов. Рассматриваются практические примеры их использования в научных исследованиях. |
|  | *Раздел 5* |  |
| 12 | **Тема 1.**  RNN. Внимание в RNN. Трансформеры. Идея слоёв внимания. BERT. BioBERT. GPT. ChatGPT. Предсказание структуры и свойств белков. Применение машинного обучения в транскриптомике. Self-supervised подходы в биологии. Мультимодальные данные.  | Введение в рекуррентные нейронные сети (RNN) и их применение в биологических данных.Механизмы внимания в RNN и их роль в обработке последовательностей.Основы и принципы работы трансформеров.Понятие слоёв внимания и их вклад в развитие ИИ.Обзор моделей BERT и BioBERT для биологических приложений.Архитектура и применение моделей GPT и ChatGPT в биоинформатике.AlphaFold2: инновации в предсказании структуры белков.Подходы к дизайну белков с заданными свойствами.Применение машинного обучения в транскриптомике.Self-supervised подходы в биологии.Анализ мультимодальных данных в биоинформатике. |

**6. Фонд оценочных средств (ФОС, оценочные и методические материалы) для оценивания результатов обучения по дисциплине (модулю).**

Оценивание результатов обучения по дисциплине проводится в виде устного или письменного зачета.

**6.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости, критерии и шкалы оценивания (в отсутствие утвержденных соответствующих локальных нормативных актов на факультете)**

Примерный список вопросов к зачету:

1. Какие основные принципы лежат в основе работы рекуррентных нейронных сетей (RNN)?
2. В чем заключается ключевое отличие механизмов внимания в RNN от традиционных RNN?
3. Опишите архитектуру трансформера и основные принципы его работы.
4. Как слои внимания в трансформерах улучшают обработку последовательностей данных?
5. В чем состоит особенность модели BERT по сравнению с традиционными методами обработки естественного языка?
6. Каковы ключевые особенности BioBERT и его применение в биоинформатике?
7. Объясните, как модель GPT отличается от предыдущих поколений языковых моделей.
8. Какие задачи биоинформатики могут быть решены с использованием ChatGPT?
9. Опишите принцип работы AlphaFold2 в предсказании структуры белков.
10. Какие методы используются для дизайна белков с заданными свойствами?
11. В чем особенности применения машинного обучения в транскриптомике?
12. Как self-supervised подходы используются в биологии и каковы их преимущества?
13. Что такое мультимодальные данные в биоинформатике и как они анализируются?
14. Какие проблемы решает использование ИИ в анализе геномных данных?
15. Какие алгоритмы машинного обучения наиболее эффективны в протеомике?
16. Объясните, как модели глубокого обучения могут улучшить разработку лекарств.
17. Какие этические вопросы возникают при использовании ИИ в биомедицинских исследованиях?
18. Как ИИ может быть использован для персонализированной медицины?
19. В чем заключается вклад механизмов внимания в точность диагностических моделей?
20. Какие типы нейронных сетей наиболее подходят для анализа медицинских изображений?
21. Как ИИ может помочь в предсказании реакции на лечение у разных пациентов?
22. Каковы преимущества использования ИИ для мониторинга эпидемиологических данных?
23. Какие основные вызовы существуют в интеграции ИИ в клиническую практику?
24. Какие методы машинного обучения используются для анализа временных рядов в медицинских данных?
25. Какие преимущества и ограничения имеют сверточные нейронные сети в анализе биомедицинских данных?
26. В чем различие между supervised и unsupervised подходами в машинном обучении, применительно к биоинформатике?
27. Как ИИ влияет на скорость и точность диагностики редких заболеваний?
28. Какие подходы используются для обеспечения безопасности и конфиденциальности данных при использовании ИИ в медицине?
29. Какие возможности открывает анализ мультимодальных данных для понимания сложных биологических систем?
30. Какие новые направления исследований в биоинформатике наиболее обещающие в контексте развития ИИ?

**6.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации по дисциплине (модулю), критерии и шкалы оценивания (в отсутствие утвержденных соответствующих локальных нормативных актов на факультете)**

1. Какой, на ваш взгляд, самый большой успех применения ИИ в биологии за последние 5 лет? Обоснуйте свой ответ.
2. Приведите 3 примера биологических задач, с которыми ИИ способен справиться лучше, чем классические алгоритмы. Опишите методы ИИ, используемые для их решения.
3. Как устроен пайплайн машинного обучения? Расскажите его на примере простой задачи.
4. Терминология ML: что такое признак, выборка. Концепция переобучения. Обучение и тестирование.
5. Какие существуют задачи машинного обучения?
6. Архитектура конволюционной нейросети. Как устроена свертка? Почему конволюция работает с изображениями намного более эффективно, чем полносвязные слои?
7. Опишите, в чем заключается задача сегментации? Как устроена архитектура U-Net?
8. Опишите постановку задачи предсказания экспрессии по последовательности ДНК. Какие методы используются для получения исходных данных для такого анализа?
9. Расскажите про концепцию message passing в графовых нейронных сетях. Для каких биологических задач GNN показали наибольшие успехи в биологии?
10. Как осуществляется работа нейронных сетей с текстовыми объектами? Токенизация и проблема представления текстов в памяти компьютера. Эмбеддинги слов. Word2Vec.
11. Как устроена архитектура RNN? Что представляет собой концепция внимания в RNN?
12. Как используется архитектура Attention Transformer? Идея слоёв внимания. Наиболее популярные модели и архитектуры: BERT, BioBERT, GPT, ChatGPT.
13. Задача предсказания структуры и свойств белков. Успехи модели AlphaFold2.
14. Как с помощью AlphaFold2 осуществляется дизайн белков с заданными свойствами?
15. Методы снижения размерности для обработки и визуализации транскриптомных данных. На чем основан метод PCA?
16. Объясните, какие слабые стороны ChatGPT следует учитывать при ее применении в области научных исследований?
17. Какие проблемы решает использование ИИ в анализе геномных данных?
18. Объясните, как модели глубокого обучения могут улучшить разработку лекарств.
19. Какие типы нейронных сетей наиболее подходят для анализа медицинских изображений?

**Шкала и критерии оценивания результатов обучения по дисциплине**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Результаты** **обучения** | **«Неудовлетворительно»** | **«Удовлетворительно»** | **«Хорошо»** | **«Отлично»** |
| Знания  | Знания отсутствуют | Фрагментарные знания | Общие, но не структурированные знания | Сформированные систематические знания |
| Умения  | Умения отсутствуют | В целом успешное, но не систематическое умение | В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера) | Успешное и систематическое умение |
| Владения  | Навыки владения отсутствуют | Наличие отдельных навыков (наличие фрагментарного опыта) | В целом, сформированные навыки (владения), но используемые не в активной форме | Сформированные навыки (владения), применяемые при решении задач |

**7. Ресурсное обеспечение:**

**7.1. Перечень основной и дополнительной литературы**

- Нейронные сети, генетические алгоритмы и нечеткие системы. Рутковская Д., Пилиньский М., Рутковский Л. Пер.с польск.И.Д.Рудинского. Издательство "Горячая линия-Телеком"

ISBN 978-5-9912-0320-3 2013 2-е изд. 384 стр.

- Как учится машина: Революция в области нейронных сетей и глубокого обучения. Лекун Ян.

Издательство "Альпина PRO" ISBN 978-5-907394-92-6 2021 335 стр.

**7.1. Перечень лицензионного программного обеспечения, в том числе отечественного производства (подлежит обновлению при необходимости)**

нет

**7.2. Перечень профессиональных баз данных и информационных справочных систем (подлежит обновлению при необходимости)**

База данных научной литературы PubMed

**7.3. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»**

[**https://pubmed.ncbi.nlm.nih.gov/**](https://pubmed.ncbi.nlm.nih.gov/)

**7.4. Описание материально-технического обеспечения.**

Аудитория, проектор, компьютер для запуска презентации со слайдами

**8. Соответствие результатов обучения по данному элементу ОПОП результатам освоения ОПОП указано в Общей характеристике ОПОП.**

**9. Разработчик (разработчики) программы.**

Фенюк Б.А., д.б.н., профессор ФББ МГУ

Пензар Д.Д., преподаватель ФББ МГУ

Зинкевич А.О., научный сотрудник ФББ МГУ.