МИНОБРНАУКИ РОССИИ
Институт проблем передачи информации им. А.А. Харкевича РАН
Юго-Западный государственный университет
Институт программных систем им. А.К. Айламазяна РАН
Институт информационных технологий, Баку, Азербайджан

ОБЛАЧНЫЕ И РАСПРЕДЕЛЕННЫЕ ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

ОРВС – 2023

В РАМКАХ

НАЦИОНАЛЬНОГО СУПЕРКОМПЬЮТЕРНОГО ФОРУМА (НСКФ – 2023)

Сборник трудов 4-й международной научно-технической конференции

28 ноября – 1 декабря 2023 года

Переславль-Залесский 2024

УДК 621.383.68.3: 681.785
ББК 32.971.35

Облачные и распределенные вычислительные системы в электронном управлении. ОРВС – 2023: сборник трудов 4-й международной научно-технической конференции (28 ноября – 1 декабря 2023 года) / ред. кол.: И.И. Курочкин [и др.]; ИПС РАН. Переславль-Залесский. – Курск: Изд-во ЗАО «Университетская книга», 2024. - 127 с.

ISBN 978-5-00261-018-1
DOI 10.47581/2024.Oblokj-Raspredelenie-OPBC-2023

Сборник содержит труды 4-й международной научно-технической конференции «Облачные и распределенные вычислительные системы» (Переславль-Залесский, 28 ноября – 1 декабря 2023), проводимой в рамках Национального суперкомпьютерного форума (НСКФ – 2023). Целью конференции является ознакомление с имеющимися достижениями по созданию облачных и распределенных вычислительных систем и их внедрение в научные исследования, учебный процесс и промышленность.

Сборник предназначен для научных сотрудников, преподавателей, аспирантов и студентов вузов.

Издание осуществлено с авторских оригиналов.

Редакция не несет ответственности за ошибки авторов.

Материалы для публикации одобрены программным комитетом Международной научно-технической конференции.

ISBN 978-5-00261-018-1

УДК 621.383.68.3: 681.785
ББК 32.971.35

Содержание

Секция «Решение задач оптимизации в среде высокопроизводительных вычислений» .5

Алекперов О.Р. ПРОБЛЕМЫ БЕЗОПАСНОСТИ И КОНФИДЕНЦИАЛЬНОСТИ В МОБИЛЬНЫХ ОБЛАЧНЫХ ВЫЧИСЛЕНИЯХ 5

Волошинов В.В., Соколов А.В. РАЗВИТИЕ МЕТОДОВ КУСОЧНО ЛИНЕЙНЫХ АППРОКСИМАЦИЙ В ОБРАТНЫХ ЗАДАЧАХ С ДИФФЕРЕНЦИАЛЬНЫМИ УРАВНЕНИЯМИ 10

Секция «Искусственный интеллект и машинное обучение» 17

Алгулиев Р.М., Садыгов И.Дж. ПОСТРОЕНИЕ ФОРМУЛ УДОБОЧИТАЕМОСТИ НА ОСНОВЕ МОДЕЛИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ 17

Быков Д.К., Дурманов Н.Н., Курочкин И.И. АНАЛИЗ ИНФРАКРАСНЫХ СПЕКТРОВ БАКТЕРИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ 21

Волков С.С. АНАЛИЗ МЕТОДОВ ВЫЯВЛЕНИЯ ИСКУССТВЕННО СГЕНЕРИРОВАННЫХ ТЕКСТОВ 27

Джафарзаде К.Э. РОЛЬ МОДЕЛЕЙ GPT В ОБРАЗОВАНИИ: ПРОБЛЕМЫ И ИХ РЕШЕНИЯ 31

Елисеев А.Н., Курочкин И.И. РЕШЕНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ СЕЛЬСКОХОЗЯЙСТВЕННЫХ КУЛЬТУР С ИСПОЛЬЗОВАНИЕМ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ 35

Заречнев Д.В., Курочкин И.И. КЛАССИФИКАЦИЯ СПЕКТРОВ РАСТЕНИЙ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ 41

Кабанов А.Ю., Домрачева А.Б., Посевин Д.П. ИССЛЕДОВАНИЕ МЕТОДОВ И ТЕХНОЛОГИЙ АЙТРЕКИНГА ДЛЯ РЕАЛИЗАЦИИ ИНТЕРФЕЙСА ЗАПОЛНЕНИЯ ВЕБ-ФОРМ ПОСРЕДСТВОМ ГЛАЗНЫХ ЖЕСТОВ 44

Казимов Т.Г., Меликова Н.Дж. ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ПРИ ТЕСТИРОВАНИИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ 47

Курбанова К.Ш. СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ РАСПОЗНАВАНИЯ ЖЕСТОВ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ КАМЕР ГЛУБИНЫ 51

Mammadova L.R. A COMPARATIVE ANALYSIS OF RNN, LSTM, AND GRU FOR TEXT CLASSIFICATION 56

Махмудова Р.Ш. УГРОЗЫ ЗАЩИТЕ ПЕРСОНАЛЬНЫХ ДАННЫХ, СОЗДАВАЕМЫЕ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ, И МЕТОДЫ ИХ СНИЖЕНИЯ 60

Минина П.С., Нагимов Т.Р. ИСПОЛЬЗОВАНИЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ АНАЛИЗА СНИМКОВ КОМПЬЮТЕРНОЙ ТОМОГРАФИИ 65

Окунев Д.А. ИССЛЕДОВАНИЕ РАЗЛИЧНЫХ ИСКАЖЕНИЙ ИЗОБРАЖЕНИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ 70

Секция «Интеграция высокоуровневых ресурсов в распределенной вычислительной среде для решения научных и инженерных задач» 75

Авакьянц А.В. РАЗРАБОТКА МЕТОДА ОРГАНИЗАЦИИ СВЯЗИ МЕЖДУ КОМПОНЕНТАМИ РАСПРЕДЕЛЁННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ ЧЕРЕЗ ВИРТУАЛЬНЫЕ СЕТЕВЫЕ КАНАЛЫ НА ОСНОВЕ ИНКАПСУЛЯЦИИ ДАННЫХ В СЛУЖЕБНЫЕ ПРОТОКОЛЫ 75

Baghirov E. CRITICAL ANALYSIS AND REVIEW OF CURRENT RESEARCH ON GANs FOR MALWARE DETECTION 81

Востокин С.В., Русин М.А. ПРОЕКТИРОВАНИЕ АРХИТЕКТУРЫ СЕРВИСА СИНХРОНИЗАЦИИ ГЛОБАЛЬНОГО СОСТОЯНИЯ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ 84

Гашимов М.А. ПРОБЛЕМЫ ПРИМЕНЕНИЯ FOG COMPUTING ТЕХНОЛОГИЙ В СРЕДЕ УМНОГО ГОРОДА 87

Секция «Гриды из рабочих станций и комбинированные гриды» 93

Балабаев С.А., Лупин С.А. ВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ ВЫЧИСЛЕНИЯ НА КЛАСТЕРЕ ИЗ СМАРТФОНОВ 93

Болгак А.В., Ватутин Э.И. ОЦЕНКА РЕАЛЬНОЙ ПРОИЗВОДИТЕЛЬНОСТИ ПРОЦЕССОРОВ СЕМЕЙСТВА INTEL CORE РАЗЛИЧНЫХ ПОКОЛЕНИЙ В ЗАДАЧЕ УМНОЖЕНИЯ ВЕЩЕСТВЕННЫХ МАТРИЦ ДЛЯ ОДНОПОТОЧНОЙ ПРОГРАММНОЙ РЕАЛИЗАЦИИ 98

Ватутин Э.И., Никитина Н.Н., Манзюк М.О., Курочкин И.И., Альбертьян А.М. О ЧИСЛЕ ТРАНСВЕРСАЛЕЙ В ДИАГОНАЛЬНЫХ ЛАТИНСКИХ КВАДРАТАХ ЧЕТНЫХ ПОРЯДКОВ 101

Вердиева Н.Н. ПРИМЕНЕНИЕ МЕТОДА МАТРИЧНОЙ ФАКТОРИЗАЦИИ ДЛЯ УЛУЧШЕНИЯ РЕКОМЕНДАЦИЙ ПРОЕКТОВ ГРАЖДАНСКОЙ НАУКИ НА ПЛАТФОРМЕ CITSCI.ORG 106

Жиронкин А.В., Ватутин Э.И. СПЕЦИАЛИЗИРОВАННОЕ ИТЕРАЦИОННОЕ ВЫЧИСЛИТЕЛЬНОЕ УСТРОЙСТВО УМНОЖЕНИЯ БИНАРНЫХ МАТРИЦ 110

Колесникова Д.П., Курочкин И.И. ГЕНЕРАЦИЯ МОТИВИРУЮЩИХ ФРАЗ МЕТОДАМИ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРОЕКТА ДОБРОВОЛЬНЫХ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛЕНИЙ 113

Секция «Прикладное программное обеспечение» 121

Штейников А.А., Пенкин А.Д., Иванов И.П., Посевин Д.П. ПРОГРАММНО-АППАРАТНЫЙ КОМПЛЕКС БЕСПРОВОДНОЙ ПЕРЕДАЧИ ДАННЫХ 121

АЛФАВИТНЫЙ УКАЗАТЕЛЬ 126

9. Zhou Z., Peng Y. The locally Chen-Harker-Kanzow-Smale smoothing functions for mixed complementarity problems. Journal of Global Optimization, 2019, 74(1), pp. 169-193.
10. Misener R., Floudas C.A. Piecewise-Linear Approximations of Multidimensional Functions. J. Optim. Theory Appl., 145, pp. 120–147, 2010.
11. Huchette J., Vielma J.P. Nonconvex Piecewise Linear Functions: Advanced Formulations and Simple Modeling Tools. Operations Research, 2022, 71(5), pp. 1835-1856.

Секция «Искусственный интеллект и машинное обучение»

Алгулиев Р.М., Садыгов И.Дж.

Институт информационных технологий, Баку, Азербайджан

ПОСТРОЕНИЕ ФОРМУЛ УДОБОЧИТАЕМОСТИ НА ОСНОВЕ МОДЕЛИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Аннотация. В статье представлена методика создания формул удобочитаемости со многими параметрами на основе существующих формул удобочитаемости и используемых в них параметров (факторов) для оценки степени сложности текстов. Для этой цели описано построение модели множественной линейной регрессии, способ определения регрессоров и коэффициентов регрессии. На примерах показано, что существующие формулы удобочитаемости являются частными случаями предложенной формулы удобочитаемости со многими параметрами.

Ключевые слова: формулы удобочитаемости, модель множественной линейной регрессии, шкала оценивания, формулы удобочитаемости со многими параметрами.

Известно, что к настоящему времени разработаны сотни формул удобочитаемости для оценки сложности текстов (в основном для текстов на английском языке). Правда, не все из них стали популярными, однако некоторые из этих формул широко используются и сегодня. Также известно, что почти во всех формулах удобочитаемости используются 1-3 параметра. Каждый параметр оценивает удобочитаемость (сложность) текста с одной стороны. Поэтому даже для одного и того же текста разные формулы удобочитаемости, считающиеся надежными, могут выдавать разные результаты. Возникает естественный вопрос: нельзя ли за счет увеличения числа параметров в формулах добиться более полных, более «точных» оценок? Конечно, это возможно. Тогда возникает второй вопрос: почему не создаются такие формулы удобочитаемости со многими параметрами?

Дело в том, что «бум формул удобочитаемости» начался в середине прошлого века и продолжался до 1980-х годов [1]ю Из алгоритмов применения этих формул также понятно, что они предназначены для «ручного» расчета. Так, чтобы упростить и сократить время расчета, в алгоритмах многих формул отмечается взятие фрагмента из 100 слов оцениваемого текста. Однако, поскольку персональные компьютеры сегодня широко используются и создание необходимого программного обеспечения для анализа текста не является проблемой, мы считаем, что нет серьезных препятствий для создания формул удобочитаемости со многими параметрами. Для этого необходимо решить два вопроса: 1) какие факторы, влияющие на сложность текстов, следует включить в создаваемые формулы удобочитаемости; 2) как должны быть определены коэффициенты параметров, входящих в формулы?

Предлагаем такую методику для решения поставленной задачи:

1) Взять наиболее широко используемые формулы удобочитаемости;

2) Выбрать параметры (коэффициенты), используемые в каждой из этих формул;

3) Определить частоту использования каждого параметра, т. е. в скольких формулах удобочитаемости используется каждый параметр, выбрать несколько (например, 4-5) параметров с более высокой частотой использования и на основе этих параметров построить уравнение регрессии;

4) Определить шкалу оценки;

5) Определить коэффициенты параметров так, чтобы значения, полученные из уравнения регрессии, попали в заданную шкалу.

С 1920 по 1980 годы было разработано более 200 формул удобочитаемости английских текстов. Конечно, не все эти формулы стали популярными, но некоторые формулы широко используются и сегодня: [2]

- Формула легкости чтения Флэша (Flesch Reading Ease Formula);
- Формула уровня образования Флэша-Кинкейда (Flesch-Kincaid Grade Level Formula);
- Формула удобочитаемости Фрая (Fry Readability Formula);
- Индекс туманности Ганнинга (Gunning Fog Index);
- Формула удобочитаемости Дейла-Челла (Dale-Chall Readability Formula);
- Формула удобочитаемости SMOG (SMOG Readability Formula);
- Формула удобочитаемости Спеша (Spache Readability Formula);
- Формула удобочитаемости Пауэрса-Самнера-Кёрла (Powers-Sumner-Kearl Readability Formula);
- Формула удобочитаемости FORCAST (FORCAST Readability Formula);
- Автоматизированный индекс удобочитаемости (Automated Readability Index).

Большинство существующих формул удобочитаемости основаны на модели линейной регрессии, а переменными этой модели выступают статистические параметры текста:

$$f(x, b) = b_0 + \sum_{i=1}^{k} b_i x_i$$

где b_i – коэффициенты регрессии; x_i – регрессоры (факторы сложности модели); k – количество факторов модели.

В формулах удобочитаемости обычно используются 1–3 параметра, связанных с лексикой и синтаксисом. В этих формулах, основными областями применения которых является оценка учебных текстов, коэффициенты регрессии выбраны таким образом, чтобы полученные результаты указывали на уровень образования или возраст читателя для понимания предлагаемого текста.

Покажем в виде таблицы, какие коэффициенты используются в каждой из 10 приведенных формул удобочитаемости (таблица 1).

Таблица 1

Параметры, используемые в формулах удобочитаемости

№	Формула удобочитаемости	Используемые параметры
1	Формула легкости чтения Флэша	ASL, ASW
2	Формула уровня образования Флэша-Кинкейда	ASL, ASW
3	Формула удобочитаемости Фрая	ASL, ASW
4	Индекс туманности Ганнинга	ASL, NHW
5	Формула удобочитаемости Дейла-Челла	ASL, PDW
6	Формула удобочитаемости SMOG	NST, NHW
7	Формула удобочитаемости Спеша	ASL, PDW
8	Формула удобочитаемости Пауэрса-Самнера-Кёрла	ASL, NSL
9	Формула удобочитаемости FORCAST	NSW
10	Автоматизированный индекс удобочитаемости	ASL, AWL

Здесь:
ASL – average sentence length – средняя длина предложения в словах,
ASW – average number of syllables per word – средняя длина слова в слогах,
AWL – average word length in symbols – средняя длина слова в символах,
NHW – hard words (number of words of more than two syllables) – количество трудных слов (количество слов с 3 и более слогами),
NST – number of sentences – количество предложений,
NSL – number of syllables – количество слогов,
NSW – number of single-syllable words in a 150-word sample – количество односложных слов в выборке из 150 слов,
PDW – percentage of difficult words (words not on the Dale-Chall word list) – процент сложных слов (слов, которых нет в списке Дейла-Челла).

Выделим среди этих параметров 4 наиболее часто используемых параметра. Как видно из таблицы, параметр средней длины предложения (*ASL*) участвует практически во всех из упомянутых 10 самых популярных формул. Далее идут следующие параметры: средняя длина слова (*ASW*), процент сложных слов (*PDW*), количество многосложных слов (*NHW*). Если принять эти параметры за основу, приведенная выше регрессионная модель будет иметь вид:

$$f(x, b) = b_0 + b_1 \times ASL + b_2 \times ASW + b_3 \times PDW + b_4 \times NHW$$

Почти все упомянутые выше формулы удобочитаемости можно рассматривать как частные случаи этой формулы. Например, при $b_0 = 206.835$, $b_1 = –1.015$, $b_2 = –84.6$, $b_3 = b_4 = b_5 = 0$ получается формула легкости чтения Флэша для английских текстов:

$$RE = 206.835 – 1.015 \times ASL – 84.6 \times ASW.$$

Кстати, соответствующие коэффициенты в модифицированной версии этой формулы для текстов на азербайджанском языке будет $b_0 = 206.835$, $b_1 = –1.318$, $b_2 = –44.3$ [3], а для русских текстов $b_0 = 206.836$, $b_1 = –1.3$, $b_2 = –60.1$ [4].

При значениях коэффициентов регрессии $b_0 = –15.59$, $b_1 = 0.39$, $b_2 = 11.8$, $b_3 = b_4 = b_5 = 0$ получается формула уровня образования Флэша-Кинкейда (для текстов на английском языке):

$$GL = 0.39 \times ASL + 11.8 \times ASW – 15.59;$$

При значениях $b_0 = 3.6365$ (если значение x_3 выше 5%, иначе $b_0 = 0$), $b_1 = 0.496$, $b_3 = 0.1579$, $b_2 = b_4 = b_5 = 0$ получается формула удобочитаемости Дейла-Челла:

$$Adjusted\ Score = \begin{cases} 0.1579 \times PDW + 0.496 \times ASL + 3.6365, & PDW > 5 \\ 0.1579 \times PDW + 0.496 \times ASL, & PDW \le 5 \end{cases}$$

Остается определить шкалу оценки и коэффициенты параметров, чтобы значения, полученные из регрессионной модели, попали в эту шкалу. Для этой цели можно принять шкалу, предложенную Р. Флэшем в 1949 г. (табл. 2) [5].

Таблица 2

Значения легкости чтения по формуле Флэша

Значения легкости чтения	Описание
0–30	очень сложный
30–50	сложный
50–60	достаточно сложный
60–70	средний
70–80	достаточно простой
80–90	простой
90–100	очень простой

Для определения коэффициентов параметров считаем целесообразным использовать такую методику: для выбранных формул удобочитаемости необходимо рассчитать значения степени сложности текста при различных возможных значениях используемые там параметры. Например, соответствующие значения удобочитаемости для различных возможных значений параметров средней длины предложения и средней длины слова в формуле Флэша приведены в таблице ниже (табл. 3).

Таблица 3

Зависимость значений легкости чтения по формуле Флэша от средней длины предложения и средней длины слова

Средняя длина слова (*ASW*)	**Средняя длина предложения (*ASL*)**							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1.2	100	99	98	97	96	95	94	93
1.3	92	91	90	89	88	87	86	85
1.4	83	82	81	80	79	78	77	76
1.5	75	74	73	72	71	70	69	68
1.6	66	65	64	63	62	61	60	59
1.7	58	57	56	55	54	53	52	51
1.8	49	48	47	46	45	44	43	42
1.9	41	40	39	38	37	36	35	34
2.0	33	32	31	30	29	27	26	25
2.1	24	23	22	21	20	19	18	17

Аналогичные таблицы следует подготовить и для других формул удобочитаемости, а затем провести экспертную оценку удобочитаемости в различных случаях, когда задействованы все параметры *ASL*, *ASW*, *PDW*, *NHW*. В результате такой оценки можно определить функцию зависимости (линию тренда) легкости чтения от указанных 4 параметров.

Библиографический список

1. DuBay W.H. The Classic Readability Studies // Costa Mesa, California: Impact Information, 2006.
2. DuBay W.H. The Principles of Readability. Costa Mesa, California: Impact Information, 2004.
3. Sadigov I.J. Mathematical and information models for evaluating readability of texts in Azerbaijani language. “El-Cezeri Journal of Science and Engineering”, v. 5, no. 3, 2018, pp.888–903.
4. Оборнева И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров. Дис. канд. пед. наук, Москва: 2006.
5. Алгулиев Р.М., Садыгов И.Дж. Оценивание удобочитаемости учебников на азербайджанском языке // "ScienceRise", 2018, № 11(52), s. 50–57.

Быков Д.К.[1], Дурманов Н.Н.[2], Курочкин И.И.[3]

1 Университет науки и технологий МИСИС, Москва
2 Институт биохимической физики им. Н.М. Эмануэля РАН, Москва
3 Институт проблем передачи информации им. А.А Харкевича РАН, Москва

АНАЛИЗ ИНФРАКРАСНЫХ СПЕКТРОВ БАКТЕРИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Аннотация. Данная работа исследует применение методов машинного обучения для классификации бактерий на основе анализа их инфракрасных спектров. Исследование включает анализ данных с применением методов уменьшения размерности, таких как PCA, t-SNE и LDA, для эффективного извлечения ключевых признаков из сложных инфракрасных спектров. Далее на данных с уменьшенной размерностью была обучена модель случайного леса библиотеки TensorFlow Decision Forests для классификации бактерий. Оценка качества классификации по метрике accuracy показала высокие результаты для метода LDA, достигнув 100% точности. Выводы исследования указывают на перспективность применения линейного дискриминантного анализа совместно с моделью случайного леса для анализа инфракрасных спектров бактерий, что имеет важное значение для разработки эффективных методов диагностики в биомедицинских исследованиях.

Ключевые слова: инфракрасная спектроскопия, машинное обучение, случайный лес, PCA, LDA, t-SNE, TensorFlow Decision Forests.

Инфракрасная спектроскопия представляет собой метод анализа, который исследует воздействие инфракрасного излучения на вещество. Его основа заключается в том, что молекулы вещества абсорбируют определенные частоты инфракрасного излучения, что приводит к изменению их колебательных и вращательных состояний. Этот метод широко применяется в химии с целью идентификации веществ, определения их структуры и изучения химических реакций.

Инфракрасное излучение представляет собой электромагнитное излучение с длиной волны от 0,78 до 1000 микрометров, расположенное в спектре между видимым светом и микроволнами. Это излучение обладает уникальными свойствами, которые придает ему значимость в различных областях наук. Когда ИК-излучение взаимодействует с веществом, оно вызывает колебания и вращения молекул. Молекулы, в свою очередь, состоят из атомов, объединенных химическими связями. Эти колебания и вращения приводят к изменению энергетических состояний молекул. Каждая молекула обладает уникальной структурой, что приводит к уникальным колебательным и вращательным модам. Колебательные моды отвечают за изменение расстояний между атомами в молекуле, а вращательные моды – за изменение ориентации молекулы в пространстве. ИК-спектроскопия использует это взаимодействие для анализа состава и структуры вещества. При прохождении через образец ИК-излучение поглощается молекулами, и характер поглощения зависит от типа связей и функциональных групп в молекуле.

ИК-спектр представляет собой график интенсивности поглощения ИК-излучения в зависимости от его частоты или волнового числа. Характеристические пики на спектре соответствуют функциональным группам и типам связей в молекуле, что позволяет анализировать состав и структуру вещества.

В контексте медицины анализ инфракрасных спектров с применением методов машинного обучения представляет собой перспективный подход, который может значительно сократить время диагностики, обеспечивая более быстрое и точное выявление бактериальных видов. Значимость данного исследования подтверждается результатами нескольких предыдущих работ, в том числе статьи [1], где проводился анализ инфракрасных спектров воздуха, выдыхаемого различными группами волонтеров. В данном исследовании были рассмотрены группы, страдающие диабетом первого типа, бронхиальной астмой и пневмонией. Результаты указывают на то, что инфракрасная спектроскопия может быть эффективным инструментом для различения характерных молекулярных профилей в выдыхаемом воздухе у различных пациентов. Также данный метод анализа широко используется в пищевой, фармацевтической, нефтехимической, сельскохозяйственной промышленности и в производстве продуктов питания, что описано в статье [2].

Целью данного исследования было осуществить классификацию различных классов бактерий на основе их инфракрасных спектров с помощью машинного обучения.

Данные предоставлены Институтом биохимической физики им. Н.М. Эмануэля РАН и представляют собой инфракрасные спектры 4 классов бактерий. Распределение количества спектров по классам приведено на рис. 1. Инфракрасные спектры, представленные в виде числовых значений интенсивности сигнала для каждого волнового числа находятся в заданном диапазоне от 500 до 4000 $см^{-1}$, что соответствует области электромагнитного спектра, где колебания связаны с основными функциональными группами и типичными связями органических молекул. Для каждого класса бактерий имеется набор спектров. Каждый спектр содержит 1800 значений. Эти входные данные представляют собой основу для проведения анализа, уменьшения размерности и классификации бактерий с использованием методов машинного обучения. Пример изображения спектров на графике приведен на рис. 2.

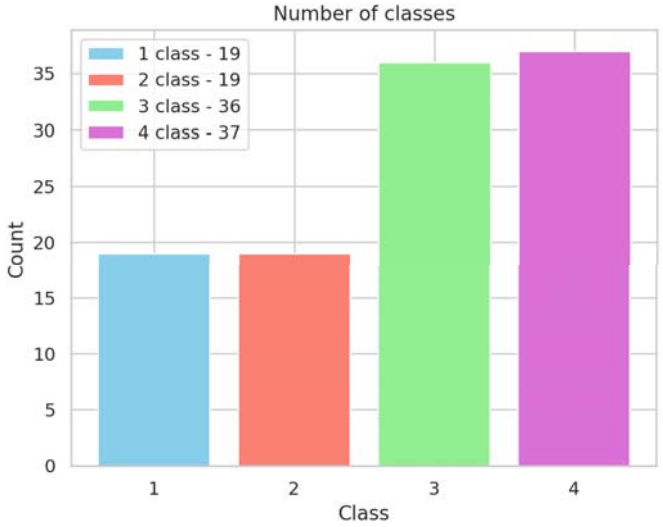

Рис. 1. Распределение количества спектров по классам

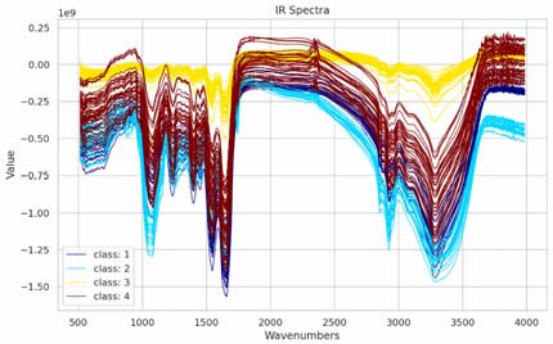

Рис. 2. Визуализация ИК-спектров 4 классов бактерий

Для улучшения производительности модели было использовано уменьшение размерности данных. Уменьшение размерности данных становится необходимым для решения ряда проблем, таких как повышение производительности моделей, снижение вычислительной сложности, и улучшение обобщающей способности на малых выборках. Этот процесс также обеспечивает более эффективное использование ресурсов при обучении моделей машинного обучения, улучшает интерпретируемость результатов и облегчает визуализацию данных. В качестве методов уменьшения размерности были выбраны три самых распространенных метода: PCA – анализ главных компонент, t-SNE – стохастическое вложение соседей с t распределением и LDA – линейный дискриминантный анализ.

Важно учитывать, что выбор количества признаков – это компромисс между сохранением информации и уменьшением размерности. В методах PCA, t-SNE и LDA было выбрано по три признака, так как объем данных пока что небольшой и также в некоторых методах (например, в методе LDA) количество признаков должно быть меньше количества классов. Однако выбор количества признаков может подвергаться корректировкам по мере поступления новых данных и изменения задач исследования.

Метод главных компонент (PCA) – это техника уменьшения размерности данных, направленная на преобразование исходных признаков в новый набор переменных, называемых главными компонентами. Главные компоненты упорядочены по убыванию дисперсии данных, что позволяет сохранить наибольшее количество информации в первых компонентах. PCA основан на вычислении собственных векторов и собственных значений ковариационной матрицы данных. Проекция данных на подпространство главных компонент позволяет эффективно уменьшить размерность, сохраняя основные характеристики данных и облегчая обработку и анализ. Результат уменьшения размерности данных методом PCA приведен на рис. 3.

t-SNE (Стохастическое вложение соседей с t распределением) – это метод снижения размерности, который позволяет сохранить локальные структуры данных и обнаруживать нелинейные зависимости. Основная идея заключается в том, чтобы преобразовать исходные данные таким образом, чтобы схожие объекты в исходном пространстве сохраняли свою схожесть и в новом, сниженном пространстве. Результат уменьшения размерности данных методом t-SNE приведен на рис. 4.

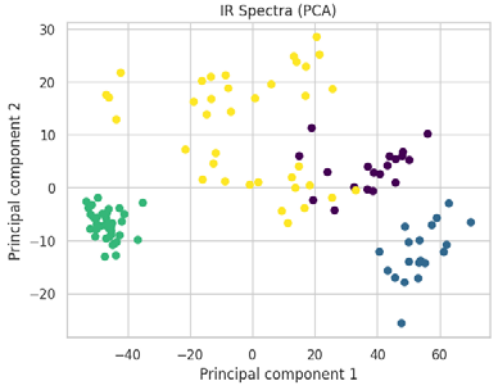

Рис. 3. Результат уменьшения размерности данных методом PCA

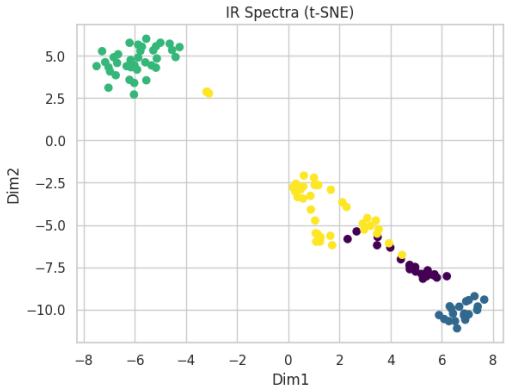

Рис. 4. Результат уменьшения размерности данных методом t-SNE

Методом, который показал лучшие результаты, стал линейный дискриминантный анализ (LDA) – метод уменьшения размерности данных и одновременной максимизации различия между классами. LDA нацелен на проекцию данных на подпространство таким образом, чтобы разброс между классами был максимален, а внутриклассовый разброс – минимален. Основан на вычислении матрицы разброса между и внутри классами. Результатом LDA являются линейные дискриминанты, которые представляют собой новые признаки, максимально разделяющие классы данных. Этот метод часто используется в задачах классификации и позволяет улучшить эффективность многоклассовой классификации за счет учета структуры классов. В данном методе размерность вывода должна быть меньше, чем количество классов. Результат уменьшения размерности данных методом LDA приведен на рис. 5. Можно увидеть на этом графике, что кластеры более чётко отделены и сгруппированы по сравнению с предыдущими методами. Одной из причин является то, что методы PCA и t-SNE не требуют информации о классах, и свои плюсы проявляет как раз там, где неизвестны разметки данных.

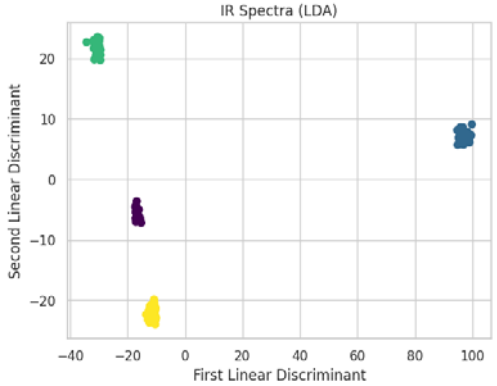

Рис. 5. Результат уменьшения размерности данных методом LDA

Основываясь на том, что набор данных небольшой, была выбрана модель случайного леса, так как данная модель эффективна на небольшом датасете и также может являться основой для будущего ансамбля моделей для решения более сложных задач.

TensorFlow Decision Forests (TF-DF) – это библиотека для обучения, запуска и интерпретации моделей леса принятия решений (например, случайных лесов, деревьев с градиентным усилением) в TensorFlow. Данная библиотека является мощным инструментом для обработки неструктурированных данных, что соответствует характеру инфракрасных спектров, представленных в виде числовых значений, что показано в статье [3]. Библиотека специально адаптирована для задач классификации с использованием лесов решений. Это позволяет эффективно работать с множеством классов бактерий, предоставляя точные результаты классификации.

Получившиеся данные были разбиты на тренировочную (81 спектр) и тестовую (30 спектров) выборку. Далее была скомпилирована модель с использованием оптимизатора Adam и функции потерь MSE (Mean Squared Error). Также были указаны метрики для оценки производительности модели, включая точность accuracy.

На основе проведенного анализа данных с использованием различных методов уменьшения размерности, включая PCA, t-SNE и LDA, результаты показали, что линейный дискриминантный анализ (LDA) наилучшим образом подходит для задачи классификации бактерий на основе инфракрасных спектров. Предложенный метод не только обеспечивает точность в выделении характерных особенностей спектров, но также позволяет применять их для успешной классификации различных видов бактерий. Оценка классификации приведена в таблице 1.

Таблица 1

Оценка классификации данных по метрике accuracy

Метод	Accuracy
PCA + TF-DF	0.91
t-SNE + TF-DF	0.93
LDA + TF-DF	1.0

В ходе исследования был проведен анализ инфракрасных спектров бактерий с использованием методов машинного обучения с целью эффективной классификации микроорганизмов. Рассмотрены различные методы уменьшения размерности данных, включая PCA, t-SNE и LDA, а также использована модель случайного леса для классификации.

Появление новых объемов данных может дать перспективу для будущих исследований:

1. Улучшение обобщающей способности модели.
2. Разработка методов для определения характеристик и состава среды, в которой собираются образцы для анализа.
3. Тестирование на различных типах образцов для применения в других областях.

Результаты исследования предоставляют важную информацию о потенциале инфракрасной спектроскопии и методов машинного обучения для точной диагностики и классификации бактерий. Этот подход может иметь практическое применения в области медицины, промышленности и научных исследований.

Финансирование: работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 122040800139-4).

Библиографический список

1. Голяк, И. С., Гутырчик, Т. А., Небритова, О. А., Морозов, А. Н., Анфимов, Д. Р., Винтайкин, И. Б., Коноплева, А. А., Фуфурин, И. Л. Применение машинного обучения для диагностики некоторых социально значимых заболеваний по выдыхаемому человеком воздуху методом инфракрасной лазерной спектроскопии // Оптика и спектроскопия. — 2023. — № 6. — С. 825-831.
2. Zhang, W., Kasun, L. C., Wang, Q. J., Zheng, Y. & Lin, Z. (2022). A review of machine learning for near-infrared spectroscopy. Sensors, 22(24), 9764-. https://dx.doi.org/10.3390/s22249764.
3. Analysis of infrared spectra using tensorflow 2.0 [Электронный ресурс] – URL: https://www.kaggle.com/code/gtteixeira/analysis-of-infrared-spectra-using-tensorflow-2-0 Дата обращения: 30.10.2023.