

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования

«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий - РТФ

Кафедра информационных технологий и систем управления

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК

Зав. кафедрой ИТиСУ
 Е.В. Кислицын
(подпись) (Ф.И.О.)
« 04 » 06 2024 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ИЗВЛЕЧЕНИЕ И КЛАССИФИКАЦИЯ ПРИЗНАКОВ ИЗ НАБОРА ДАННЫХ
ОКУЛОГРАФИИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Научный руководитель: Долганов Антон Юрьевич
к.т.н., доцент

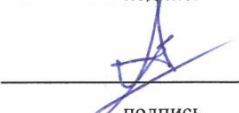
Нормоконтролер: Бредихина Наталья Сергеевна

Консультант Чистяков Сергей Николаевич

Студент группы: РИМ-220906 Касов Артем Михайлович


подпись

подпись

подпись

подпись

Екатеринбург
2024

Институт радиоэлектроники и информационных технологий - РТФ
Кафедра информационных технологий и систем управления
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа Инженерия искусственного интеллекта

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Касова Артема Михайловича группы РИМ-220906
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы Извлечение и классификация признаков из набора данных окулографии методами машинного обучения

Утверждена распоряжением по институту от «4» декабря 2023 г. № 33.02-05/298

2. Научный руководитель Долганов Антон Юрьевич, доцент, кандидат технических наук

(Ф.И.О., должность, ученая степень, ученое звание)

3. Исходные данные к работе Набор данных отслеживания глаз в анализе дислексии

4. Перечень демонстрационных материалов презентация в MS PowerPoint

5. Календарный план

п/п	Наименование этапов выполнения работ	Срок выполнения этапов работы	Отметка о выполнении
1.	Глава 1. Обзор предметной области	до 23.03.2024 г.	✓
2.	Глава 2. Материалы и методы	до 29.04.2024 г.	✓
3.	Глава 3. Эксперименты и результаты	до 19.05.2024 г.	✓
4.	ВКР в целом	до 20.05.2024 г.	✓

Научный руководитель Долганов А.Ю.
Ф.И.О.

Студент задание принял к исполнению 22.02.24

дата

(подпись)

(подпись)

6. Допустить Касова Артема Михайловича к защите выпускной квалификационной работы в экзаменационной комиссии

Зав. кафедрой ИТиСУ

(подпись)

Е.В. Кислицын
Ф.И.О.

РЕФЕРАТ

Выпускная квалификационная работа 83 с., 18 рис., 57 источн., 1 прил.
ОКУЛОГРАФИЯ, EYE-TRACKING, МАШИННОЕ ОБУЧЕНИЕ,
ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ, КЛАССИФИКАЦИЯ, ДИСЛЕКСИЯ, АНАЛИЗ
ДАННЫХ, ПАТТЕРНЫ ДВИЖЕНИЯ ГЛАЗ.

Объект исследования – процесс анализа данных окулографии.

Предметом исследования являются методы извлечения и классификация признаков данных окулографии.

Цель работы – обзор существующих исследований в области окулографии, анализ данных, полученных с помощью технологии Eye-tracking.

В процессе исследования проводились: изучение методов извлечения и классификации признаков с использованием Scikit-Learn, разработка методологии анализа данных и моделей машинного обучения для выявления паттернов движений глаз, проведение экспериментов для сравнения эффективности алгоритмов.

В работе продемонстрирована эффективность подхода с отбором признаков, сделаны предположения о влиянии показателей на отсутствия или присутствия нарушений в области зрения.

Область практического применения: предложенный подход может быть использован для улучшения методов диагностики дислексии, автоматизации анализа данных окулографии.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	6
1 Обзор предметной области.....	8
1.1 Окулография. Эволюция отслеживания глаз	8
1.2 Извлечение и классификация признаков	14
1.3 Методы извлечения признаков в контексте окулографии. Обзор исследований	16
1.4 Анализ статьи скрининга дислексии.....	23
2. Материалы и методы	29
2.1 Описание набора данных	29
2.2 Начало процесса выявления и извлечения признаков	32
2.3 Кодирование классов движения глаз в данных	37
2.4 Визуализация скорости движения глаз для разных групп участников .	39
2.5 Переосмысление подхода к трансформации данных и извлечению признаков	42
2.5.1 Переосмысление подхода к трансформации данных и извлечению признаков	42
2.5.2..... Гипотеза комплексного анализа динамики глазных движений	44
2.6 Применение моделей машинного обучения для анализа данных айтрекинга в диагностике дислексии.....	45
2.7 Выбор инструментария для разработки системы	46
2.7.1 Язык программирования	46
2.7.2 Выбор платформы для разработки.....	47

2.7.3 Выбор библиотек.....	48
2.8 Выбор библиотек.....	50
2.9 Выбор признаков для модели	53
2.10 Обобщение результатов.....	54
3 Результаты.....	55
3.1 Анализ распределения и корреляции целевой переменной	55
3.2 Первичное моделирование и последующая оптимизация.....	57
3.3 Сравнение различных моделей.....	59
3.4 Заключительные замечания по анализу и моделированию	65
ЗАКЛЮЧЕНИЕ	67
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	69
ПРИЛОЖЕНИЕ А	75

ВВЕДЕНИЕ

Современные технологии обработки данных и машинного обучения играют ключевую роль в решении сложных задач анализа информации, особенно в области медицинской диагностики и исследований. Одним из методов, привлекающих все большее внимание, является анализ данных окулографии (eye-tracking) с применением методов машинного обучения для извлечения и классификации признаков.

В контексте данной работы, извлечение и классификация признаков данных окулографии становятся необходимыми этапами для выявления паттернов, закономерностей и характеристик движений глаз, которые могут быть связаны с конкретными состояниями или с определенными ситуациями. Методы машинного обучения предоставляют мощные инструменты для автоматизации этого процесса и выявления скрытых закономерностей в данных.

Это исследование представляет собой анализ данных окулографии с использованием инструментов библиотеки Scikit-Learn.

Целью данной работы является проведение обзора существующих исследований в области окулографии, а также анализ данных, полученных с помощью технологии Eye-tracking.

В рамках данной работы будут поставлены следующие ключевые задачи:

- 1) обзор предметной области и литературы,
- 2) разработка методологии анализа данных и моделей машинного обучения,
- 3) анализ результатов и выводы.

Объектом исследования в данной работе является процесс анализа данных окулографии.

Предметом исследования являются методы извлечения и классификация признаков данных окулографии

Результаты исследования позволят расширить способы выявления новых признаков в контексте окулографии и предоставят важные практические рекомендации.

1 Обзор предметной области

1.1 Окулография. Эволюция отслеживания глаз

Окулография – это определение координат взора («точки пересечения оптической оси глазного яблока и плоскости наблюдаемого объекта или экрана, на котором предъявляется некоторый зрительный раздражитель»). Она представляет собой важный инструмент для изучения заболеваний, когнитивных процессов, психологических состояний, а также многих других аспектов человеческого поведения и функционирования [1]. Этот метод позволяет регистрировать и анализировать направление взгляда, его продолжительность и скорость перемещения глазных яблок, что в свою очередь отражает когнитивные и психофизиологические процессы [1 – 2].

Изучение движений глаз, таких как саккады и фиксации, представляет собой значимое направление, включая аспекты эмоциональной регуляции. Глаза являются ключевым каналом для получения информации о текущем состоянии человека и его реакциях на внешние стимулы. Они предоставляют ценные данные о физиологических процессах, эмоциональных состояниях и когнитивных функциях в реальных, повседневных ситуациях, а также в контексте различных видов межличностного взаимодействия. Использование этой технологии играет ключевую роль в исследованиях мультимодального распознавания эмоций, обеспечивая ценные данные для более глубокого понимания человеческого поведения, эмоциональных реакций, а также психического и физического здоровья [3].

В настоящее время технология отслеживания движения глаз, также известная как айтрекинг, является широко используемым инструментом для анализа визуального внимания человека. Многие психофизиологические процессы отражаются в параметрах движений глаз, таких как динамика морганий и изменение ширины зрачка, которые могут указывать на уровень усталости, когнитивной нагрузки, эмоциональные реакции и другие аспекты.

В настоящее время айтрекинг широко применяется в практических целях, в основном в области исследований юзабилити и нейромаркетинга [4].

В XIX веке исследования, связанные с отслеживанием движений глаз, проводились исключительно методом наблюдения. В 1879 году в Париже Луи Эмиль Жаваль обнаружил, что при чтении текста глазные движения не следуют простому и монотонному шаблону, как предполагалось ранее. Вместо этого происходят короткие остановки, названные им "фиксациями", и резкие передвижения, которые он назвал "саккадами". Это наблюдение привело к появлению важных вопросов о процессе чтения, которые были решены в XX веке: на каких словах сосредотачивается внимание читателя, как долго это происходит, и зачем читатель возвращается к уже прочитанным словам [5].

Пример фиксаций и саккад на рисунке 1 демонстрирует стандартную траекторию движения глаз в тексте. Наблюдается, что глаза не перемещаются равномерно по тексту, а скорее совершают короткие остановки (фиксации) и резкие передвижения (саккады).

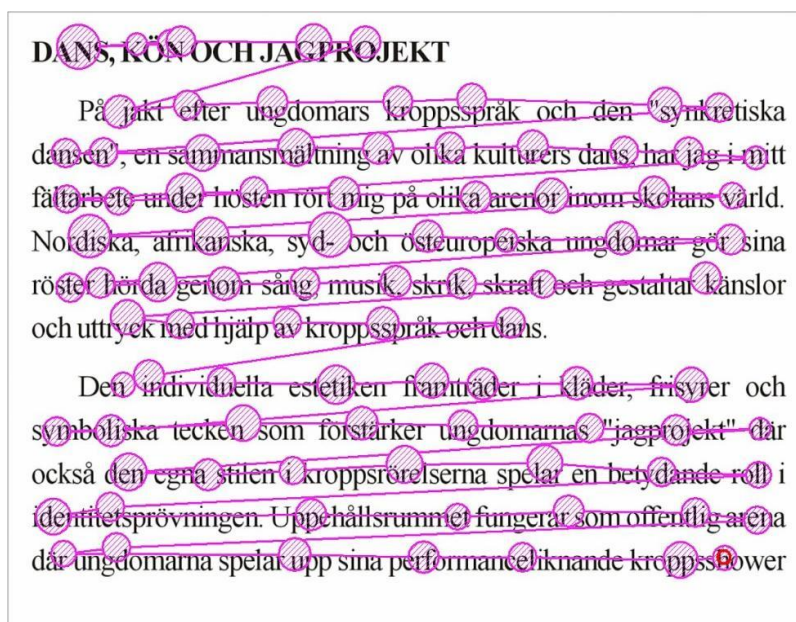


Рисунок 1 – Траектория движения глаз в тексте [56]

Первое устройство для отслеживания глаз было создано Эдмундом Хью. Оно представляло собой контактную линзу с отверстием для зрачка,

связанную с алюминиевой указкой, двигающейся синхронно с глазным яблоком [5].

Первое неинвазивное устройство, т.е. устройство не вмешивается в ткани и структуры организма, для отслеживания глаз был разработан Ги Томасом Бушвеллом в Чикаго [6]. Он использовал отражения световых лучей от глазного яблока на светочувствительную пленку для проведения исследований процессов чтения и анализа статических изображений [6].

В 1950-х годах русский ученый Альфред Ярбус провел значимые исследования в области отслеживания глаз. Его монография 1967 года получила высокую оценку мирового научного сообщества [7]. Ярбус показал, что постановка формальной задачи перед испытуемым оказывает значительное влияние на результаты эксперимента по отслеживанию глаз. Он также выявил взаимосвязь между мотивацией испытуемого и фиксациями его взгляда, показав, что результаты эксперимента зависят не только от визуального стимула, но и от поставленной задачи и ожидаемой информации от визуального стимула.

На рисунке 2 представлены результаты исследования Ярбуса, которые иллюстрируют взаимосвязь между поставленной перед испытуемым задачей и его глазными движениями. На каждую задачу испытуемому отводилось по 3 минуты

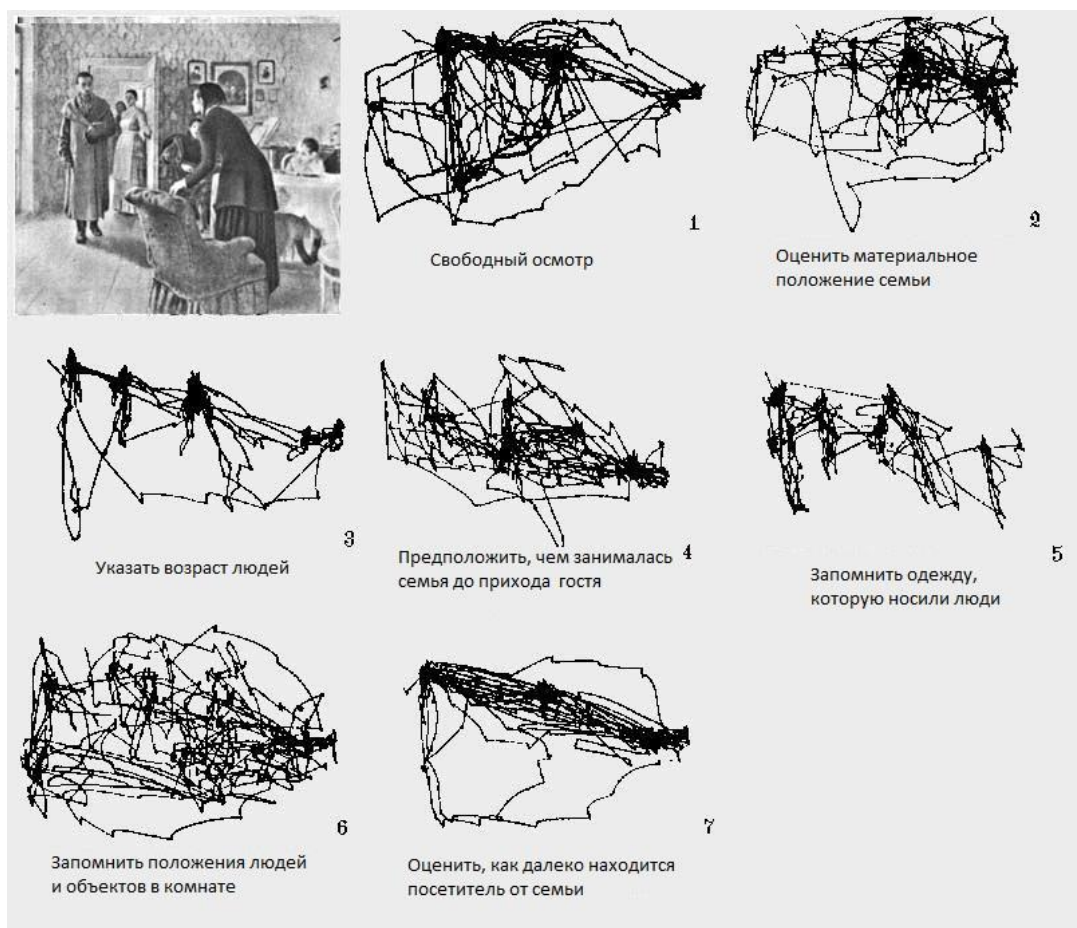


Рисунок 2 – Траектория движения глаз при поставленной задаче [57]

Записи из экспериментов, посвященных оценке движений глаз, свидетельствуют о том, что лишь небольшая часть элементов изображения привлекает внимание испытуемых, приводя к фиксациям глаз на этих элементах. Этот процесс движения глаз отражает внутренние процессы мышления человека. Взгляд, с определенной задержкой, следует за точкой, на которую направлено внимание испытуемого. Поэтому определение, какие элементы изображения привлекают внимание испытуемого, в каком порядке и с какой частотой, является относительно простой задачей.

Часто внимание испытуемых привлекают элементы, которые, по их собственному мнению, могут содержать важную информацию, хотя на самом деле это не так. Глаза испытуемых часто фиксируются на элементах, которые выделяются своей необычностью в данном контексте.

Перемещаясь от одной точки фиксации к другой, глаза испытуемых часто возвращаются к тем элементам изображения, которые они уже видели, что свидетельствует о том, что дополнительное время тратится на повторный осмотр наиболее важных элементов, в ущерб менее значимым.

В 1980 году Джаст и Карпентер сформулировали влиятельную гипотезу - Strong eye-mind hypothesis, согласно которой "между тем, что зафиксировано взглядом, и тем, что обрабатывается когнитивно, нет заметной задержки". Если эта гипотеза верна, то когда испытуемый смотрит на слово или объект, он или она также думает о нем (процессирует когнитивно) и ровно столько же времени, сколько длится зафиксированный взгляд [8]. Эту гипотезу часто принимают на веру исследователи, использующие технику отслеживания глаз. Однако, методика, основанная на направленном взгляде, предлагает интересный вариант для разделения явного и скрытого внимания, для разграничения того, что фиксируется взглядом и что обрабатывается когнитивно.

В 1980-е годы гипотеза о взаимосвязи между глазом и умом часто ставилась под сомнение в свете скрытого внимания, то есть внимания к чему-то, на что человек не смотрит, что люди часто делают. Если скрытое внимание является распространенным явлением во время записей отслеживания глаз, то результирующие пути сканирования и узоры фиксации часто показывают не то, где было внимание, а только где был направлен взгляд, что не отражает когнитивную обработку.

В 1980-е годы также началось использование отслеживания глаз для ответа на вопросы, связанные с взаимодействием человека с компьютером. В частности, исследователи изучали, как пользователи ищут команды в меню компьютера. Кроме того, компьютеры позволили исследователям использовать результаты отслеживания глаз в реальном времени, преимущественно для помощи людям с ограниченными возможностями [9].

В последнее время наблюдается рост использования отслеживания глаз для изучения взаимодействия пользователей с различными интерфейсами

компьютера. Исследователи задаются конкретными вопросами о том, насколько легко разные интерфейсы воспринимаются пользователями. Результаты исследований по отслеживанию глаз могут привести к изменениям в дизайне интерфейса [10].

Еще одной актуальной областью исследований является разработка веб-сайтов. Сюда входит, например, реакция пользователей на выпадающие меню или тот факт, куда они сосредотачивают внимание на веб-сайте, чтобы разработчик знал, куда разместить рекламу [10].

Благодаря прогрессу в портативных электронных устройствах, портативные головные трекеры в настоящее время способны достигать отличных показателей и все более часто используются в исследованиях и прикладных задачах, направленных на повседневные сценарии. Этот же прогресс привел к увеличению исследований небольших глазных движений, которые происходят во время фиксации, как в лабораторных условиях, так и в прикладных ситуациях [11].

В XXI веке использование искусственного интеллекта (ИИ) и искусственных нейронных сетей стало жизнеспособным способом выполнения задач и анализа отслеживания глаз. В частности, сверточная нейронная сеть подходит для отслеживания глаз, поскольку она предназначена для задач, связанных с изображениями. С помощью ИИ задачи исследования и отслеживания глаз могут предоставлять дополнительную информацию, которая не могла бы быть обнаружена наблюдателями. Практика глубокого обучения также позволяет определенной нейронной сети улучшаться в выполнении задачи при наличии достаточного количества обучающих данных. Однако для этого требуется относительно большое количество данных для обучения [12].

Потенциальные применения ИИ в отслеживании глаз охватывают широкий спектр тем от медицинских приложений до безопасности водителей, игровой теории, а также образовательных и тренировочных приложений.

1.2 Извлечение и классификация признаков

Извлечение признаков (feature extraction) представляет собой важный этап в анализе данных, который заключается в создании новых признаков из существующих данных. Целью этого процесса является сокращение размерности данных и выделение наиболее информативных аспектов для последующего анализа или обучения моделей машинного обучения. Извлечение признаков может быть выполнено вручную или автоматически [13].

В контексте анализа данных айтрекинга, ручное извлечение признаков играет ключевую роль в определении характеристик, которые могут быть использованы для классификации и прогнозирования различных состояний, включая дислексию. Этот процесс требует глубокого понимания предметной области и способности идентифицировать те аспекты данных, которые наиболее значимы для решения конкретной задачи. Во многих ситуациях хорошее понимание фона или области может помочь принять обоснованные решения относительно того, какие признаки могут быть полезными. За десятилетия исследований инженеры и ученые разработали методы извлечения признаков для изображений, сигналов и текста.

Автоматическое извлечение признаков использует специализированные алгоритмы, машинное обучение или глубокие нейронные сети для автоматического извлечения признаков из наборов данных без необходимости в человеческом вмешательстве. Эта техника может быть очень полезной, когда требуется быстро перейти от сырых данных к разработке алгоритмов машинного обучения. Вейвлет-преобразование является примером автоматического извлечения признаков [14].

С расцветом глубокого обучения ручное извлечение признаков в значительной степени было заменено первыми слоями глубоких сетей, но в основном для изображений. Для сигналов и временных рядов извлечение

признаков остается первой проблемой, требующей значительных знаний перед тем, как можно построить эффективные прогностические модели.

Основные методы извлечения признаков включают в себя преобразование данных, создание новых признаков на основе комбинации существующих, а также использование алгоритмов машинного обучения для автоматического извлечения информации из данных.

Преобразование данных может включать в себя такие методы, как преобразование категориальных переменных в числовые, масштабирование признаков для обеспечения их сопоставимости, а также уменьшение размерности данных с использованием методов, таких как метод главных компонент (PCA) [15].

Создание новых признаков может основываться на анализе взаимодействий между существующими признаками или на введении дополнительных признаков на основе знания о предметной области.

Использование алгоритмов машинного обучения для извлечения признаков может включать в себя такие методы, как автоэнкодеры или глубокие нейронные сети, которые могут автоматически извлекать важные признаки из данных.

Все эти методы извлечения признаков играют важную роль в обработке данных, помогая обеспечить лучшую производительность моделей машинного обучения и выявлять скрытые закономерности в данных для более эффективного принятия решений.

Пример извлечения признаков можно рассмотреть на задаче обработки изображений для классификации объектов. Предположим, у нас есть набор изображений с различными видами фруктов, и мы хотим создать модель машинного обучения, которая сможет автоматически определять тип каждого фрукта на изображении.

Для начала мы можем извлечь характеристики изображений, которые будут полезны для классификации. Например, мы можем извлечь следующие признаки из каждого изображения:

- 1) цветовые характеристики: средние значения интенсивности красного, зеленого и синего каналов,
- 2) текстурные характеристики: статистики по текстурным фильтрам, таким как градиенты или гистограммы ориентированных градиентов (HOG),
- 3) форма и размер объекта: площадь, периметр, соотношение сторон,
- 4) геометрические характеристики: координаты ключевых точек или дескрипторы углов.

После извлечения этих признаков мы можем использовать их как входные данные для модели машинного обучения. Модель будет обучаться на этих признаках, чтобы определять, какой тип фрукта изображен на входном изображении.

Таким образом, это позволяет нам конвертировать изображения в числовые характеристики, которые могут быть использованы для обучения модели и принятия решений.

1.3 Методы извлечения признаков в контексте окулографии. Обзор исследований

Извлечение признаков является важным этапом в анализе данных, особенно в окулографии. В данной главе рассматривается методика извлечения признаков из данных окулографии, которая позволяет выявить ключевые характеристики движения глаз и использовать их для классификации и анализа. Будут рассмотрены основные признаки, используемые при анализе окулографических данных, а также методы их извлечения. Это позволит лучше понять процессы, происходящие во время визуального восприятия, и разработать более эффективные методы анализа данных.

В рамках отслеживания движений глаз будут рассмотрены несколько основных признаков, которые достаточно часто используются при анализе данных окулографии. Эти признаки включают в себя:

- 1) размер зрачка,
- 2) саккады (быстрые движения глазных яблок),
- 3) фиксация взгляда,
- 4) скорость движения глаз,
- 5) моргание,
- 6) позиция зрачка,
- 7) электроокулограмма (EOG),
- 8) точка взгляда.

Каждый из этих признаков предоставляет ценную информацию о состоянии и поведении человека в различных ситуациях и условиях. Рассмотрение этих признаков позволяет лучше понять процессы внимания, эмоциональные реакции, уровень утомления и многие другие аспекты, связанные с функционированием человеческого зрительного восприятия.

Sao и др. авторы предложили подход для системы распознавания намерений с использованием эндоскопического манипулятора, используя размер зрачка и скорость вращения глаза [16]. Исследование показало, что изменение размера зрачка имеет значительное влияние на управление эндоскопом. В работе Labibah авторы провели эксперимент, чтобы определить, может ли человек, используя изменения диаметра зрачка и движение глазного яблока влево или вправо [17]. Определение лжи классифицируется с использованием алгоритма дерева решений. Технология отслеживания глаз также широко используется в области безопасности. В работе Vozkir и др. предложена схема распознавания когнитивной нагрузки по направлению взгляда для водителей в критических ситуациях на основе диаметра зрачка в виртуальной реальности [18]. В исследовании Orlosky и др. авторы провели эксперимент по распознаванию понимания словарного запаса

и языка пользователя в интерфейсах дополненной реальности (AR) и виртуальной реальности (VR) [19].

В работе Slanzi авторы оценили эксперимент по прогнозированию намерения пользователя кликнуть при просмотре веб-сайта с использованием комбинации расширения зрачка и сигналов электроэнцефалограммы (ЭЭГ). Исследования показали, что у пользователей больше размер зрачка, когда они собираются кликнуть [20]. В другом исследовании Guo и др. предложено исследование мультимодального распознавания эмоций с использованием трех модальностей: движения глаз, таких как диаметр и моргание зрачка, ЭЭГ и изображений глаз (EIG) [21].

В исследовании, проведенном Zhang и Juhola, исследователи изучали биометрическую идентификацию и верификацию с использованием сигналов саккадического глазного движения [22]. Эффективность была высокой, с самым высоким уровнем идентификации в 90%. В другом исследовании, проведенном Sargezeh и др., авторы исследовали параметры глазного движения, такие как амплитуда саккады, количество саккад и отношение общей продолжительности фиксации к общей продолжительности саккад (RFDSD), для классификации полов [23]. В исследовании Tamuly и др. предложено исследование классификации изображений для прогнозирования эмоциональной окраски сцены с использованием множества признаков глазных движений, таких как количество саккад и частота саккад, а также методов машинного обучения [24].

Данные о движениях глаз широко используются в медицинских и клинических исследованиях. В работе Ahmed и Noble проводилась классификация изображений ультразвукового сканирования (УЗИ) для получения стандартизированных кадров живота, головы и бедра плода с использованием извлеченных фиксаций глаз [25]. Результаты показали, что создание набора моделей слов с использованием фиксаций является многообещающим методом для идентификации изображений УЗИ плода [26]. Zhou и др. представили исследование классификации изображений с

использованием метода двухэтапного выбора признаков. Основой для этого метода является область интереса (ROI), которая определяется из извлеченных данных о фиксации. Qi и др. использовали точки фиксации и траектории взгляда для распознавания материалов и улучшения производительности путем создания карт выделения [27]. Исследование Singh и др. анализирует шаблоны чтения с использованием фиксаций и саккад с использованием нескольких классификаторов для определения наилучших для оценки [28]. В другом исследовании Luo и др. предложен метод на основе взгляда для определения намерений с использованием длительности фиксации и радиуса фиксации [29]. Startsev и Dorr классифицировали аутистический спектральный расстройство (АСР) на основе траектории сканирования и особенностей выделения [30]. В работе Zhu и др. проведены два отдельных эмоциональных эксперимента на основе данных электроэнцефалографии (ЭЭГ) и отслеживания глаз с использованием метода ансамбля на основе содержания (CAM) в качестве классификатора [31]. Фиксации захватывают визуальное внимание человека, когда он сосредоточен на интересном объекте. Viduapu и др. [32] предложили прогнозирование внимания на изображениях веб-страниц с использованием мультиклассификации. В работе Roy и др. разработан метод когнитивной обработки для определения неоднозначных изображений [33].

Кроме того, есть 3 статьи, которые использовали комбинацию ЭЭГ с фиксациями в своих исследованиях. В работе Shi и др. предложен метод оценки внимания с использованием ЭЭГ, фиксаций и траектории сканирования для автоматической классификации данных высокого и низкого качества с использованием пространственно-временного анализа траектории сканирования [34].

Согласно исследованию Czyzewski и др, они предложили мультимодальный подход для полисенсорного лечения и стимуляции некоммуникативных субъектов и классифицировали реальное и воображаемое движение конечностей, используя точки фиксации взгляда и ЭЭГ [35]. В исследовании Jiang и др. предложен метод классификации аутизма на основе

особенностей лица и фиксаций глаз, в то время как в работе Tharaliya и др. авторы использовали комбинацию времени фиксации и ЭЭГ для классификации диагноза аутизма [36 – 37]. Исследования показали, что совместное использование данных дает лучшие результаты по сравнению с данными ЭЭГ или глаза в отдельности. Кроме того, имеется исследование Ding и др. о выявлении большого депрессивного расстройства (БДР) с использованием трех комбинаций данных: данных отслеживания взгляда, ЭЭГ и данных гальванических кожных реакций (ГКР) [38]. Более того, имеется исследование Abdelrahman, которое предложило метод классификации типов внимания с использованием данных о фиксации и тепловом изображении [39].

Из исследования Koza и др. авторы предложили программу распознавания назначения эндоскопического манипулятора на основе 2 классификаторов: метода опорных векторов (SVM) и вероятностной нейронной сети (PNN) [40]. В исследовании использовались признаки изменения диаметра зрачка и скорости вращения глаза. Исследования показали, что изменение диаметра зрачка оказывает значительное влияние на сроки активации эндоскопического манипулятора для перемещения рабочего поля в центр визуального поля монитора.

В исследовании Borys и др. был проведен эксперимент по классификации производительности пользователей на основе теста фигуральной свободной флуэнции (RFFT) с использованием машинного обучения и признаков глаз, таких как моргания, размер зрачка, фиксации и саккады. RFFT оценивает способность человека генерировать новые фигуры с использованием пяти различных конфигураций точек [41]. В исследовании Guo и др. проведен эксперимент по оценке нагрузки и исследованию использования технологии отслеживания глаз в оценке производительности в телеманипуляции в космосе с учетом двух факторов влияния, таких как задержка и временное давление [42]. В этом исследовании использовались признаки глаз, такие как моргания, фиксации, саккады, взгляд и диаметр зрачка. Производительность распознавания нагрузки оценивается с

использованием предложенного метода и протокола leave-one-subject-out (LOCO). Классификация четырех классов нагрузки выполнена, и лучшая достигнутая точность составила 49,32%.

В исследовании Guo и соавторов (2019) были использованы моргания и диаметр зрачка для классификации эмоций пяти классов с использованием комбинации EEG и EIG [42]. В работе На и др. авторы создали устройство-помощник при приеме пищи для повышения самооценки и улучшения качества жизни людей с ограниченными возможностями и пожилых людей [43]. Предложенная система мозг-компьютер (BCI) разработана на основе признаков трех морганий, EEG и EMG. В этом исследовании были получены положительные результаты с точностью в диапазоне от 83 до 97%.

Также технология отслеживания позиции глаз стала активно применяться в диагностике медицинских состояний. В исследовании Касир и др. (2019) был представлен метод диагностики шизофрении с использованием позиции взгляда с тестом Роршаха [44]. В работе Yoo и др. предложена техника классификации записей с числовыми данными, введенными с помощью взгляда [45].

Существуют исследования, которые использовали комбинацию позиции зрачка и сигналов функциональной магнитно-резонансной томографии (fMRI) для декодирования восприятия двойной решетки (Wilbertz и др.) [46]. Двойное восприятие описывает состояние, при котором сознательное восприятие непрерывно перемещается между двумя потенциальными восприятиями физически согласованного, но воспринимаемого как неоднозначного стимула. В исследовании Lin и др. предложена система ментального написания, использующая комбинацию позиции взгляда и особенностей стабильных вызванных потенциалов зрительной системы (SSVEP) из электроэнцефалограммы с анализом канонической корреляции фильтрового банка (FBCCA) для классификации [47].

Электроокулография — это метод, используемый для измерения потенциала между роговицей и сетчаткой глаза человека. В исследовании

Lagodzinski и др. был проведен эксперимент по распознаванию когнитивной активности с использованием сигналов ЭОГ и алгоритма машинного обучения, метода кодового книжного подхода [48].

В исследовании Kubacki автор предложил новую гибридную систему BCI для сортировки предметов. Исследование проводилось с использованием промышленного робота в виртуальной модели [49]. Управление роботом осуществлялось с помощью признаков SSVEP, ЭОГ, позиции глаз и обратной связи по силе. Тесты были воспроизведены на реальном промышленном роботе. Проверка точности позиционирования кончика робота проводилась с помощью системы обратной связи. Это исследование показало, что предложенная система может сортировать предметы, используя сигналы человеческого тела, и поддерживать точность на уровне 90%. В работе Song и др. предложили глубокую рекуррентную автоэнкодерную модель (DCRA) для оценки бдительности с использованием комбинации ЭЭГ и ЭОГ [50]. В этом исследовании дизайн автоэнкодера был основан на блоках управляемых рекуррентных единиц (GRUs). Данные ЭЭГ и ЭОГ записывались с помощью нейросканера, а данные о движениях глаз собирались с помощью очков для отслеживания взгляда, включая моменты моргания, закрытия глаз, фиксацию и саккады. Для моделирования эксперимента использовался набор данных о бдительности водителя SEED-VIG (Zheng и Lu) [51].

В исследовании Horng и Lin авторы разработали эксперимент по прогнозированию и классификации сонливости, используя мультимодальные биосигналы, такие как движения глаз, GSR, сигналы мозговой активности и пульс [52]. В работе Karessli и др. авторы использовали признаки взгляда и встраивания, такие как признаки взгляда с сеткой (GFG), признаки взгляда с последовательностью (GFS) и гистограмма взгляда (GH) для классификации изображений без обучающих примеров [53]. Извлеченные данные о взгляде включали точки взгляда, местоположение взгляда, длительность взгляда, последовательность взгляда, а также диаметр зрачка. В данном исследовании в качестве классификатора использовался метод опорных векторов (SVM), и

наивысшая достигнутая точность составила 78,2%. В работе Saab и др. авторы собрали данные о взгляде для классификации медицинских изображений с использованием сверточных нейронных сетей (CNN). Полученные результаты показали положительный эффект и указали на то, что данные о взгляде могут использоваться для предоставления потенциальных сигналов наблюдения за медицинским изображением [54].

В этой главе были рассмотрены основные признаки, используемые в контексте окулографии и анализа данных отслеживания глаз. Каждый признак имеет свою уникальную значимость и может быть использован для различных целей.

1.4 Анализ статьи скрининга дислексии

Дислексия — это расстройство чтения, вызванное развитием нервной системы, которое отрицательно влияет на скорость и точность распознавания слов, затрудняя беглость чтения и понимание текста.

В эксперименте используются данные отслеживания глаз 185 испытуемых, участвующих в проекте развития чтения Kronoberg, проведенном с 1989 по 2010 год. Из первоначальной когорты из 2165 человек, 103 испытуемых были выявлены в 1989 году как группа с высоким риском дислексии [55]. Отбор испытуемых не включал определение дислексии на основе расхождений между уровнем чтения и IQ. Контрольная группа испытуемых с низким риском была сопоставлена с группой высокого риска по нескольким критериям. Когнитивное, образовательное и социальное развитие было оценено в течение 20 лет. Все испытуемые читали один и тот же текст, представленный на одной странице белой бумаги с высокой контрастностью.

Исследователи анализировали записи движения глаз, используя алгоритм порога динамической дисперсии, чтобы выделить особенности движения глаз для классификации между субъектами высокого и низкого

риска. Также, они стремились сохранить как можно больше исходного сигнала о движении глаз, используя широкий и систематический отбор признаков. Был определен простой набор низкоуровневых признаков, охватывающих фиксацию и саккады. Фиксации и саккады не были объединены в составные показатели. Саккады были разделены на поступательные и регрессивные движения. Для каждого типа события определилось несколько параметров, отражающих их количественные свойства, включая продолжительность, амплитуду, направление и стабильность. Итоговый набор признаков состоял из 168 параметров, отражающих различные аспекты движения глаз при чтении.

Для выявления периодов фиксации, саккадических движений и других событий в записях движения глаз авторы использовали алгоритм порога динамической дисперсии для анализа необработанных данных о положении глаз с течением времени (см. Рисунок 3). На основе этого анализа были выделены характеристики движения глаз, которые затем использовали как входные данные для обучения классификационной модели с целью различения между субъектами HR и LR. Авторы избегали ориентации процесса выделения признаков на предполагаемые различия в поведении при движении глаз между группами, стремясь к широкому, систематическому и объективному подходу к отбору признаков, чтобы сохранить как можно больше информации о движении глаз. В результате был определен простой набор низкоуровневых признаков, охватывающих как фиксацию, так и саккадические события.

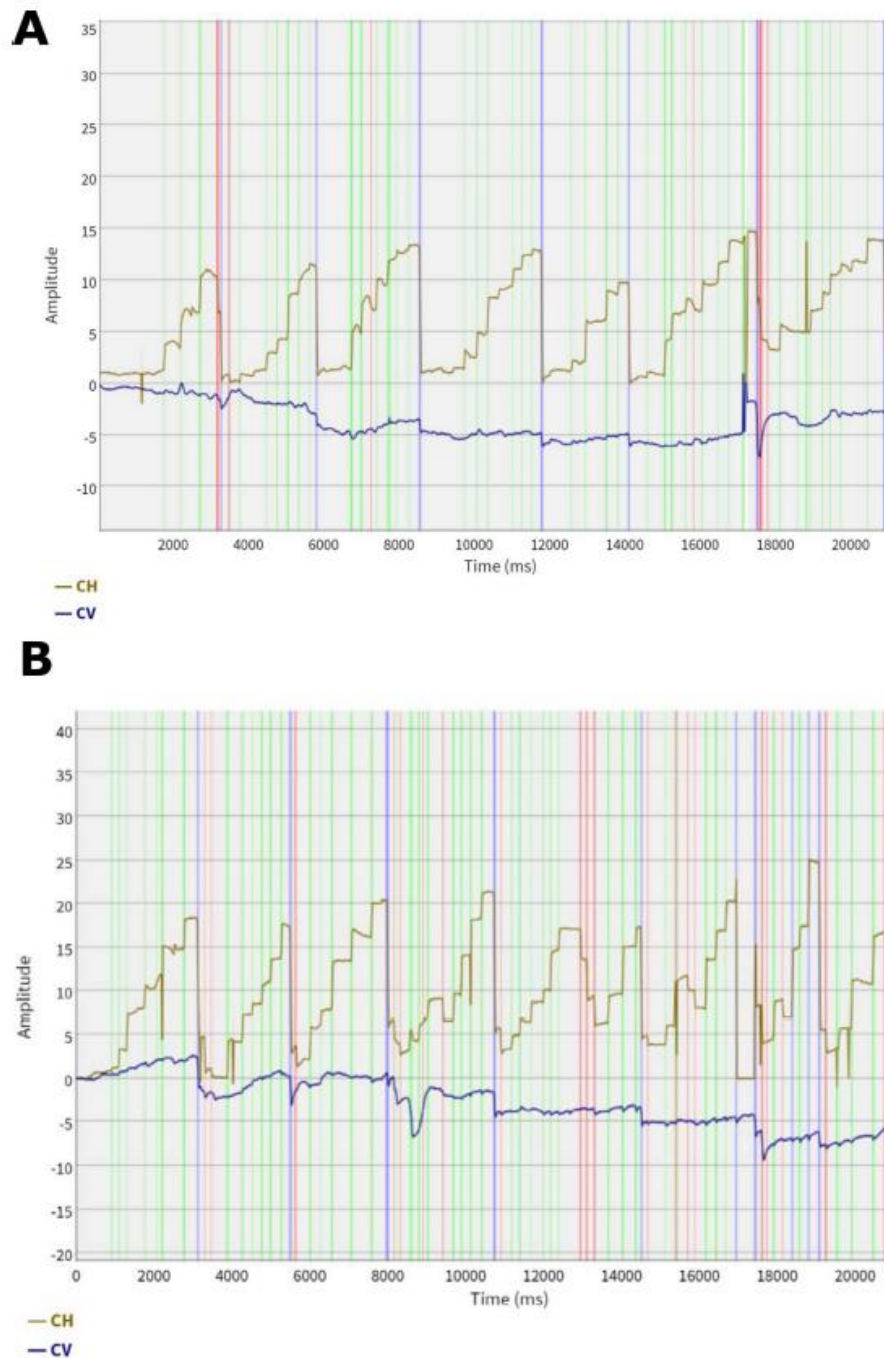


Рисунок 3 – Пример анализа движения глаз, где сигнал движения глаз по горизонтали (CH) и вертикали (CV) отображается во времени [55]

В соответствии с анализом, представленным на графике А и графике В, светло-зеленые полосы указывают на саккады, светло-серые области - на фиксации, светло-синие полосы - на движения взгляда (sweeps), чаще всего обратные движения (returning-sweeps), а красные полосы - на переходные процессы.

Для проведения анализа использовался алгоритм динамического порога дисперсии, основанный на физиологических свойствах фовеального и парафовеального полей зрения. Алгоритм работает, анализируя сигнал отслеживания выборка за выборкой и переключаясь между четырьмя взаимоисключающими состояниями: искажениями, переходными процессами, фиксациями и саккадами.

Саккады разделены на движения вперёд (слева направо) и назад (справа налево), а фиксации определены в соответствии с предыдущим направлением саккады. Для каждого типа фиксации и саккады измерены следующие параметры: продолжительность, расстояние, среднее положение глаз, стандартное отклонение среднего положения, максимальный диапазон между двумя позициями и накопленное расстояние по всем последующим позициям. Параметры 2-6 измерены по горизонтали и вертикали для версии и вергенции, и вычислены как среднее положение двух глаз и разница в положениях двух глаз соответственно. В итоге было получено 168 признаков, отражающих различные характеристики движений глаз при чтении, такие как их продолжительность, амплитуда, направление, стабильность и симметрия.

Были использованы линейные машины опорных векторов с последовательной минимальной оптимизацией для обучения классификаторов с максимальным запасом. Для оценки их прогностической эффективности сравнились результаты с произвольным выбором признаков. Классификаторы обучались с использованием рекурсивного исключения признаков, чтобы удалить нерелевантные и избыточные признаки. Эффективность классификации была оценена с использованием 10-кратной перекрестной проверки и повторена 100 раз для стабилизации оценок. Для демонстрации эффективности все классификаторы были протестированы на рандомизированных данных. Чтобы оценить способность алгоритма рекурсивного исключения признаков выбирать предиктивные признаки для классификации, сравнилась производительность всех классификаторов с результатами, полученными при простом выборе признаков случайным

образом. Для этого авторы обучали классификаторы с использованием n случайно выбранных признаков из полного набора признаков в каждом тренировочном наборе. Все эксперименты проводились с использованием одной и той же схемы тренировки и тестирования, которая представлена на рисунке 4.

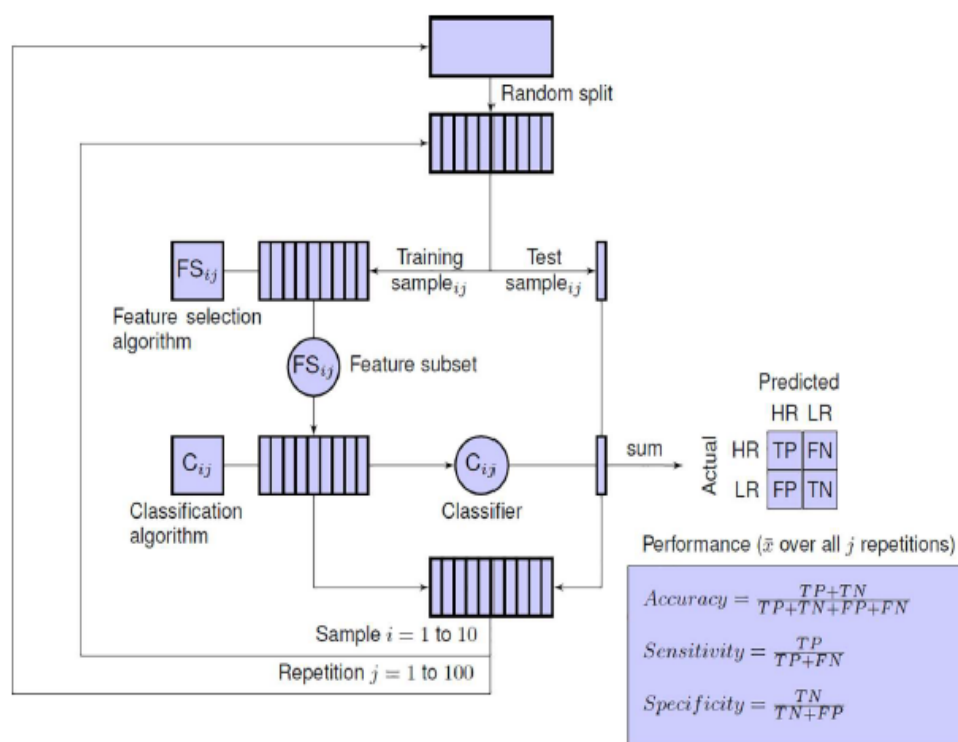


Рисунок 4 – Схема экспериментального тестирования, основанная на повторной перекрестной валидации с внутренним выбором [55]

Вся база данных случайным образом делится на 10 подмножеств, отводя одно подмножество (10% всех испытуемых) как тестовую выборку, а оставшиеся девять подмножеств (90% всех испытуемых) как обучающую выборку. На обучающей выборке применяется алгоритм отбора признаков для выбора подмножества из n признаков. Используя это подмножество признаков, на обучающей выборке применяется алгоритм классификации, который производит классификатор с параметрами в качестве выходных данных. Затем этот классификатор используется для классификации

испытуемых в тестовой выборке, и предсказанные результаты сравниваются с фактической идентичностью (HR или LR) испытуемых в тестовой выборке. Этот шаг повторяется 10 раз с различными обучающими и тестовыми наборами для каждой итерации. После завершения одного полного цикла перекрестной проверки 10 раз, каждый испытуемый во всей базе данных был протестирован ровно один раз, при этом строгая разница между обучающими и тестовыми испытуемыми сохранялась. Для уменьшения дисперсии оценки производительности перекрестной проверки весь процесс повторяется 100 раз с различными первоначальными случайными делениями исходной базы данных. Окончательная оценка ожидаемой прогностической производительности вычисляется путем усреднения производительности перекрестной проверки за все 100 повторений. Эта оценка представляет собой ожидаемую точность прогноза финальной модели. Финальная модель - та, которую мы бы использовали на практике, — это классификатор, который мы бы построили из всей базы данных с использованием метода отбора признаков m для выбора n признаков.

2. Материалы и методы

2.1 Описание набора данных

В данной главе, посвященной обзору данных, будут рассмотрены исходные данные статьи, опубликованной в PLOS ONE, о применении технологии отслеживания взгляда для скрининга дислексии [55].

Набор данных представлен в форме архива, структурированного по именам папок, таких как "111GM3", "111JA2", "111RP1", и так далее. Каждая папка содержит файлы формата txt, где первая строка определяет структуру данных и содержит следующие столбцы: T, LX, LY, RX, RY. Здесь T обозначает время, а LX, LY, RX и RY соответственно представляют положение левого глаза по координате X и Y, а также правого глаза по координате X и Y. Пример содержания такого файла приведен на рисунке 5.

T	LX	LY	RX	RY	
0	0	0	0	0	
20	0,65535	-9,9999999960687E-6		0,65535999999999	-0,65536
40	0,65534	-9,9999999960687E-6		0,65535999999999	-0,65536
60	0,65534	-9,9999999960687E-6		0,65535	-0,65536
80	0,65534	-9,9999999960687E-6		0,65534	0
100	0,65533	-9,9999999960687E-6		0,65534	-0,65536
120	1,31069	-9,9999999960687E-6		0,65534	-0,65536
140	1,31069	-0,65537		1,31069	-0,65536
160	1,31068	-0,65537		1,31069	-0,65536
180	0,65532	-9,9999999960687E-6		1,31069	0
200	1,31068	0,65536	1,31068	9,9999999960687E-6	
220	-0,65543		3,27683	-1,96614	5,89831
240	-1,31079		7,86439	-1,31078	7,86439
260	-1,31079		7,86439	-0,65543	7,20903
280	-0,65543		7,20902	-1,31078	7,20903
300	-0,65543		6,55366	-1,31078	6,55367
320	-0,65543		5,8983	-0,65542	5,24294
340	-0,65543		5,89829	-1,31078	5,24294
360	-0,65543		5,24293	-1,31077	4,58757
380	-0,65543		4,58756	-1,31077	4,58757
400	-0,65543		3,9322	-1,31077	3,93221
420	-0,65543		3,27684	-1,31077	2,62148
440	-1,31078		3,27683	-1,31077	2,62148
460	-1,31078		2,62147	-1,31077	2,62148
480	-0,65542		3,27683	-0,65541	3,27684
500	-1,31078		2,62147	-1,31077	2,62147
520	-0,65542		2,62147	3,93216	2,62146
540	4,58751	3,93217	3,2768	3,93219	
560	3,27678	3,93218	3,2768	3,27682	
580	3,27678	3,93217	3,27679	3,93219	
600	3,27678	4,58754	2,62143	3,93218	
620	2,62141	3,93218	2,62143	3,27682	
640	1,96606	2,62145	1,96607	1,9661	
660	1,96607	2,62144	1,96607	1,96609	
680	1,31071	1,96608	1,96608	1,31072	
700	1,31071	1,31071	1,31073	0,65535	
720	1,31071	0,65534	1,31073	-1,0000000003174E-5	
740	1,31072	0,65534	1,31073	-1,0000000003174E-5	
760	1,31073	-1,999999999243E-5		1,31074	-1,999999999243E-5
780	1,31073	-2,9999999995312E-5		1,31074	-0,65538
800	1,31073	-0,65539		0,65538	-0,65538

Рисунок 5 – Содержание файла с данными

Кроме указанных файлов, в наличии также имеется дополнительный текстовый файл-справочник, содержащий интерпретацию названий папок. В этом файле описано, что если название папки завершается на числа 1 или 2, это указывает на группу с нарушением чтения, а если на числа 3 или 4, то на

контрольную группу. Кроме того, числа 1 или 3 указывают на мужчин, а числа 2 или 4 – на женщин.

Для последующего анализа данные будут преобразованы в формат DataFrame библиотеки Pandas с целью улучшения их структурирования и обеспечения более эффективного анализа. Дальнейшие усилия будут направлены на выявление закономерностей и создание новых признаков на основе временных различий в положении глаз. В процессе будут выделены основные признаки, такие как саккады и фиксации.

Будет проведен анализ временных рядов положения глаз с использованием методов машинного обучения для выявления характерных паттернов и различий между группами с нарушениями чтения и контрольной группой. Дополнительные признаки будут созданы на основе статистических характеристик временных рядов, а также путем выделения особенностей саккад и фиксаций. Этот анализ позволит лучше понять особенности движений глаз у лиц с нарушениями чтения и предоставит новые данные для диагностики и лечения данного расстройства.

Далее будет проведен анализ качества данных, включая проверку на наличие пропущенных значений и аномальных выбросов. Затем будет осуществлено создание дополнительных признаков на основе временных рядов, таких как скорость и ускорение движения глаз, а также временные интервалы между событиями. После этого данные будут разделены на обучающий и тестовый наборы для последующего обучения и валидации модели машинного обучения. Выбранные алгоритмы машинного обучения будут обучены на обучающем наборе данных и затем протестированы на тестовом наборе для оценки их производительности и способности предсказания. Полученные результаты будут анализироваться с целью выявления значимых закономерностей и паттернов, что позволит лучше понять особенности движений глаз и их связь с нарушениями чтения.

2.2 Начало процесса выявления и извлечения признаков

Первым этапом работы с набором данных стало извлечение исходных данных из архива формата zip. Далее информация была переведена в удобный формат для анализа с добавлением колонки `dir_name`, для последующего его использования в анализе (Приложение A1). В результате был получен `DataFrame`, представленный на рисунке 6.

	T	LX	LY	RX	RY	dir_name
0	0	0	0	0	0	712K03
1	20	-0,65537	0	-0,65536	0	712K03
2	40	-1,31073	0	-7,86443	-0,65536	712K03
3	60	-12,45204	0	-12,45202	0	712K03
4	80	-12,45202	1,00000000003174E-5	-11,79666	0	712K03
...	...	...	...	...	...	...
323245	39900	193,33378	36,70079	191,36769	22,93817	622FD1
323246	39920	182,84795	-20,97129	172,36199	-94,37216	622FD1
323247	39940	150,07937	-109,44495	161,87569	-83,23048	622FD1
323248	39960	157,94324	-62,91429	148,76786	-51,11779	622FD1
323249	39980	140,24795	-42,59803	129,10665	-34,73362	622FD1
323250 rows x 6 columns						

Рисунок 6 – Данные переведены в формат `DataFrame`

Для упрощения последующей обработки и анализа были введены дополнительные колонки, обогащающие данные о каждом участнике. Конкретно, добавлены колонки `sex` и `status`, где `sex` обозначает пол участников (1 — мужской, 0 — женский), а `status` отражает наличие трудностей с чтением (1 — наличие трудностей, 0 — отсутствие трудностей). Эти данные были выведены на основе анализа идентификаторов директорий (`dir_name`), что позволило классифицировать участников для последующего детального

анализа и выявления признаков, коррелирующих с возможными нарушениями чтения (Приложение А2).

Для детального изучения механизмов, потенциально влияющих на эффективность чтения, были введены новые признаки: SpeedL и SpeedR. Эти признаки представляют собой скорости движения левого (SpeedL) и правого (SpeedR) глаз соответственно (Приложение А3). Выбор скорости движения глаз как признака обусловлен несколькими факторами:

- скорость движения глаз является ключевым показателем в оценке визуальной системы участника, поскольку задержки или аномалии в этих значениях могут указывать на проблемы в визуальной обработке информации, что критично при чтении,

- сравнительный анализ движения обоих глаз дает возможность оценить симметричность и координацию взгляда, что может быть важно для понимания фундаментальных нарушений в механизмах чтения, особенно у лиц с дислексией.

Таким образом, анализ этих параметров позволяет не только выявить особенности движения глаз у читающих с трудностями, но и предложить объективные метрики для последующей классификации и более глубокого изучения нарушений чтения.

В рамках углубленного анализа данных, особое внимание было уделено наблюдениям в группе с нарушениями чтения. В качестве основы для анализа были выбраны данные из директории 111RP1, представляющей субъекта с явными трудностями чтения. Набор данных после преобразований и добавлении новых колонок для 111RP1 представлен на рисунке 7.

	T	LX	LY	RX	RY	dir_name	status	sex	SpeedL	SpeedR
245750	0	0.00000	0.00000	0.00000	0.00000	111RP1	1	1	NaN	NaN
245751	20	0.00000	0.00000	0.00000	-0.65536	111RP1	1	1	0.000000e+00	3.276800e-02
245752	40	0.00000	-0.00001	0.65536	-0.00001	111RP1	1	1	5.000000e-07	4.634060e-02
245753	60	0.65536	-0.00001	0.65537	-0.00001	111RP1	1	1	3.276800e-02	5.000000e-07
245754	80	0.65536	0.65535	0.65537	-0.00001	111RP1	1	1	3.276800e-02	0.000000e+00
...	...	...	...	...	...	...	...	...	...	...
247745	39900	-53.74032	51.77425	-53.74031	52.42962	111RP1	1	1	3.276800e-02	1.310730e-01
247746	39920	-53.74032	54.39570	-53.74031	53.74035	111RP1	1	1	1.310725e-01	6.553650e-02
247747	39940	-53.74032	55.05108	-53.74031	55.05108	111RP1	1	1	3.276900e-02	6.553650e-02
247748	39960	-53.74032	57.01717	-53.74031	57.01718	111RP1	1	1	9.830450e-02	9.830500e-02
247749	39980	-53.74032	58.98327	-53.74031	57.67255	111RP1	1	1	9.830500e-02	3.276850e-02

2000 rows x 10 columns

Рисунок 7 – DataFrame 111RP1

Для демонстрации динамики скорости движения глаз во времени были построены графики (Приложение А4), которые отражают изменения скорости левого и правого глаза по отдельности. Визуализация данных выполнялась с использованием библиотеки `matplotlib` в Python, что позволило наглядно представить зависимость скорости движения глаз от времени. Графики были построены с целью облегчить интерпретацию результатов и выявить потенциальные аномалии или особенности в данных.



Рисунок 8 - График скорости движения левого глаза 111RP1



Рисунок 9 - График скорости движения правого глаза 111RP1

На рисунке 8 представлена зависимость скорости движения левого глаза (SpeedL) от времени (T). График позволяет оценить, как изменяется скорость в зависимости от времени, моменты ускорения и замедления движения.

Аналогично, для правого глаза был построен второй график, представленный на рисунке 9, отражающий зависимость его скорости (SpeedR) от времени. Этот график также служит для анализа динамики скоростных характеристик глаз.

Эти визуализации играют важную роль в последующем анализе, поскольку они предоставляют наглядное представление о поведении глазных движений в различных условиях чтения.

В процессе анализа данных окулографии особое внимание было уделено идентификации саккад и широких взмахов глаз, которые являются важными компонентами для понимания механизмов чтения и возможных нарушений. Для этого были определены пороговые значения скорости движения глаз:

- порог саккад установлен на уровне 0.3, что позволяет отделить более медленные движения от саккад, характеризующихся быстрым и резким перемещением глаз,

- порог широких взмахов установлен на уровне 2.5, что помогает идентифицировать особенно быстрые и широкие движения глаз, превышающие обычную активность при чтении.

Для наглядности результатов анализа был построен график (Приложение А5), показанный на рисунке 10, на котором представлены скорости движения левого и правого глаз во времени, а также выделены саккады и широкие взмахи. Это позволяет визуально оценить динамику и характер движений глаз у участников исследования, выявляя возможные аномалии или особенности, связанные с нарушениями чтения. Эта визуализация и классификация движений глаз служат основой для дальнейшего более глубокого анализа и понимания взаимосвязей между характеристиками взгляда и когнитивными процессами при чтении.

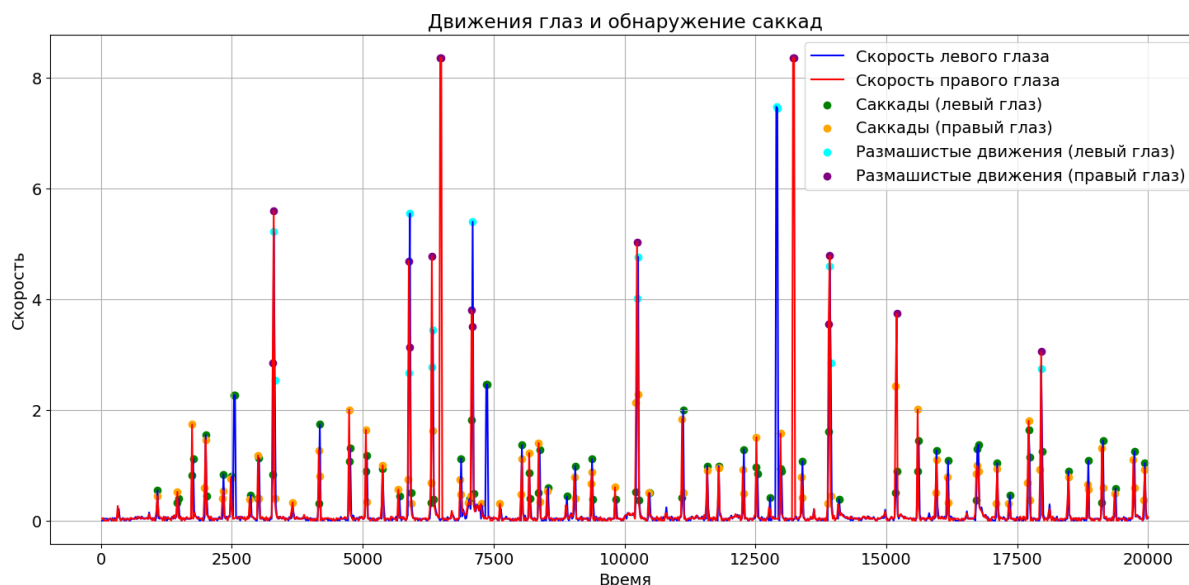


Рисунок 10 - Выделение основных признаков

2.3 Кодирование классов движения глаз в данных

Для анализа данных окулографии были созданы новые бинарные признаки (Приложение А6) в наборе данных, которые классифицируют каждое наблюдение как саккаду, широкий взмах или фиксацию в зависимости от скорости движения глаз. Эти признаки служат для уточнения классификации движений глаз и улучшения аналитической точности моделирования.

Для обоих глаз были созданы переменные `Saccade_L` и `Saccade_R`, которые обозначают наличие саккады, когда скорость движения глаза превышает порог саккад (`saccade_threshold`) и остается ниже порога широких взмахов (`sweep_threshold`):

— `Saccade_L`: Истинно (True), если скорость левого глаза (`SpeedL`) находится в интервале саккад,

— `Saccade_R`: Истинно для правого глаза, когда скорость (`SpeedR`) удовлетворяет тем же критериям.

Широкие взмахи определяются как моменты, когда скорость движения глаз значительно превышает порог широких взмахов:

— `Saccade_L`: Истинно (True), если скорость левого глаза (`SpeedL`) находится в интервале саккад,

— `Saccade_R`: Истинно для правого глаза, когда скорость (`SpeedR`) удовлетворяет тем же критериям.

Широкие взмахи определяются как моменты, когда скорость движения глаз значительно превышает порог широких взмахов:

— `Sweep_L`: Флаг устанавливается в истинное значение (True), если скорость левого глаза превышает порог (`sweep_threshold`),

— `Sweep_R`: Аналогично для правого глаза.

Также были определены фиксации — состояние, когда глаза практически неподвижны, что обозначается очень низкой скоростью движения:

— `Fixation`: Устанавливается в истинное значение (True), когда скорости обоих глаз одновременно меньше порога саккад.

Эти признаки интегрированы в `DataFrame`, что позволяет последовательно анализировать каждый тип движения глаз отдельно или в комбинации, предоставляя более широкие возможности для изучения взаимосвязей между различными видами глазных движений и когнитивными процессами, связанными с чтением и восприятием информации. Пример для 111RP1 представлен на рисунке 11.

	T	LX	LY	RX	RY	dir_name	status	sex	SpeedL	SpeedR	SpeedDiff	Saccade_L	Saccade_R	Sweep_L	Sweep_R	Fixation
245750	0	0.00000	0.00000	0.00000	0.00000	111RP1	1	1	NaN	NaN	NaN	False	False	False	False	False
245751	20	0.00000	0.00000	0.00000	-0.65536	111RP1	1	1	0.000000e+00	3.276800e-02	-3.276800e-02	False	False	False	False	True
245752	40	0.00000	-0.00001	0.65536	-0.00001	111RP1	1	1	5.000000e-07	4.634060e-02	-4.634010e-02	False	False	False	False	True
245753	60	0.65536	-0.00001	0.65537	-0.00001	111RP1	1	1	3.276800e-02	5.000000e-07	3.276750e-02	False	False	False	False	True
245754	80	0.65536	0.65535	0.65537	-0.00001	111RP1	1	1	3.276800e-02	0.000000e+00	3.276800e-02	False	False	False	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
247745	39900	-53.74032	51.77425	-53.74031	52.42962	111RP1	1	1	3.276800e-02	1.310730e-01	-9.830500e-02	False	False	False	False	True
247746	39920	-53.74032	54.39570	-53.74031	53.74035	111RP1	1	1	1.310725e-01	6.553650e-02	6.553600e-02	False	False	False	False	True
247747	39940	-53.74032	55.05108	-53.74031	55.05108	111RP1	1	1	3.276900e-02	6.553650e-02	-3.276750e-02	False	False	False	False	True
247748	39960	-53.74032	57.01717	-53.74031	57.01718	111RP1	1	1	9.830450e-02	9.830500e-02	-5.000000e-07	False	False	False	False	True
247749	39980	-53.74032	58.98327	-53.74031	57.67255	111RP1	1	1	9.830500e-02	3.276850e-02	6.553650e-02	False	False	False	False	True

2000 rows x 16 columns

Рисунок 11 - Выделение новых признаков

2.4 Визуализация скорости движения глаз для разных групп участников

Для детального сравнения поведения взгляда между участниками с нарушениями чтения и без таковых, были созданы два графика (Приложение А7), отражающие динамику скорости движения глаз, которые представлены на рисунке 12. Эти графики предоставляют визуальное представление о различиях в глазных движениях, что может способствовать выявлению специфических паттернов, связанных с нарушениями чтения.

На первом графике представлены скорости движения левого и правого глаз для участников с нарушениями чтения (df1). Скорости левого и правого глаз показаны синим и красным цветами соответственно. График помогает наблюдать за изменениями скорости во времени, выделяя возможные аномалии или особенности движений глаз, которые могут быть связаны с трудностями в чтении.

Аналогично, для контрольной группы без нарушений чтения (df2) был построен второй график, где скорости движений левого и правого глаз обозначены зеленым и оранжевым цветами. Этот график служит для сравнения и выявления нормативных паттернов движения глаз, предоставляя базу для анализа отклонений, наблюдаемых у первой группы.

На обоих графиках указаны оси времени (Time) и скорости (Speed), что позволяет точно отслеживать изменения скорости в различные моменты времени. Для лучшей наглядности и сопоставимости результатов на графиках присутствует сетка, а также легенда, которая помогает различать данные для левого и правого глаза каждой группы. Эти визуализации играют ключевую роль в анализе и могут помочь выявить, как изменения в скорости глазных движений связаны с процессом чтения и его нарушениями. Визуальный анализ также способствует формированию гипотез о механизмах, которые могут приводить к различиям в чтении между двумя группами.

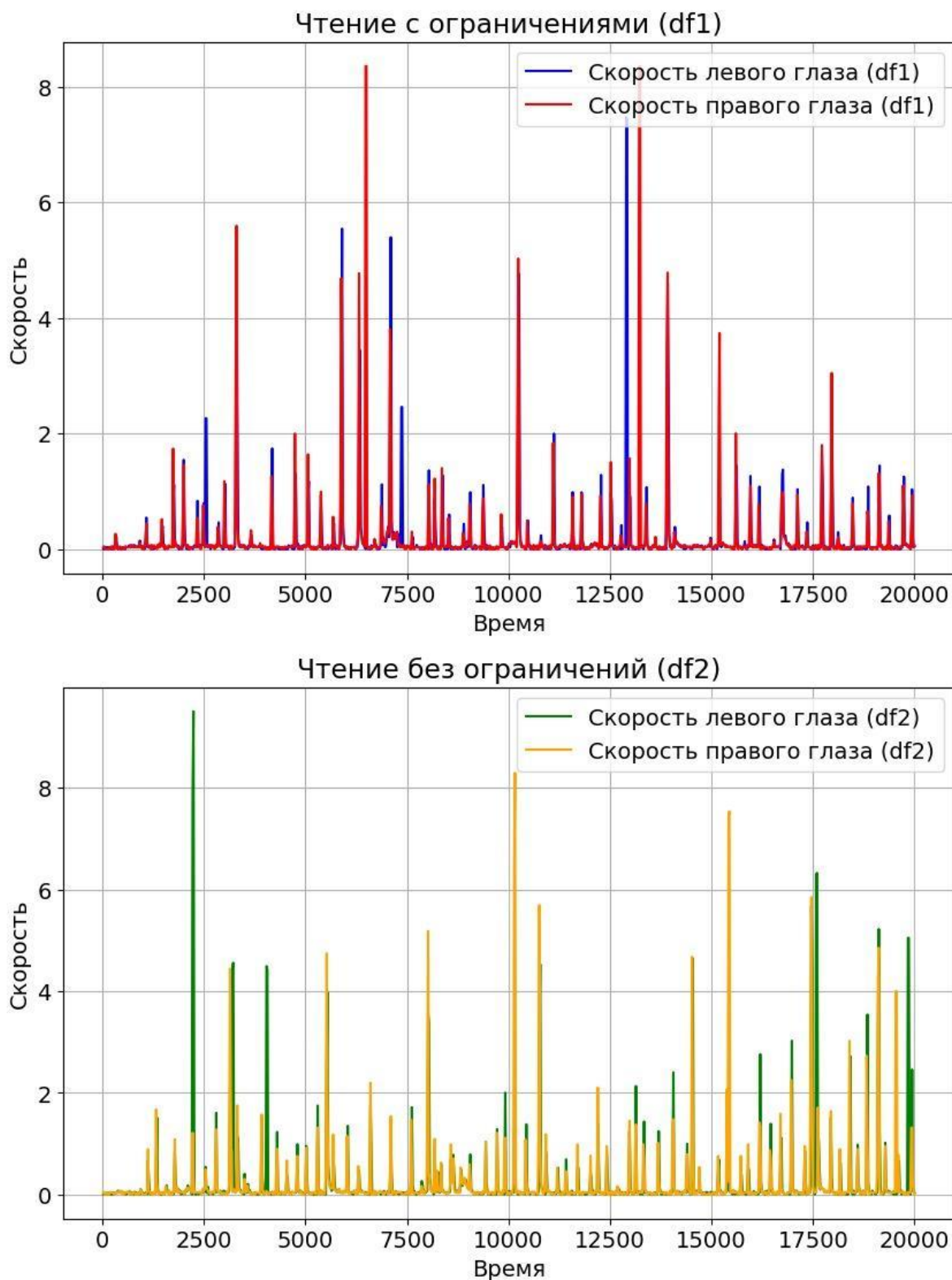


Рисунок 12 - Различия в движениях глаз

На основе построенных графиков были замечены некоторые отклонения в скорости движений глаз, которые потенциально могут указывать на наличие трудностей при чтении. Однако, несмотря на визуально наблюдаемые

различия, стоит отметить, что данные графики не позволяют с уверенностью определить, насколько скорость движения глаз является надежным индикатором нарушений чтения.

Широкие взмахи определяются как моменты, когда скорость движения глаз значительно превышает порог широких взмахов:

— субъективность интерпретации: Визуальный анализ может быть подвержен субъективным оценкам и трактовкам, что может привести к различным выводам от разных исследователей. Sweep_R: Аналогично для правого глаза,

— необходимость дополнительного анализа: Наблюдаемые отклонения требуют подтверждения через статистический анализ и могут потребовать дополнительных экспериментальных данных для уточнения связей между скоростью движения глаз и специфическими нарушениями чтения.

Для более точной диагностики и анализа нарушений чтения необходимо использовать комплекс подходов, включая статистические методы и машинное обучение, которые могут обеспечить более глубокий и объективный анализ данных. Графики могут служить в качестве предварительного инструмента для выявления потенциальных паттернов, но окончательные выводы должны базироваться на более строгих методах анализа и проверке гипотез.

2.5 Переосмысление подхода к трансформации данных и извлечению признаков

2.5.1 Переосмысление подхода к трансформации данных и извлечению признаков

Транспонирование матрицы данных в ходе анализа движений глаз — таких как саккады, фиксации и другие — с целью создания матрицы, где

каждое наблюдение представлено отдельными признаками, индексированными по времени, изначально рассматривалось как метод повышения аналитической точности. Этот подход позволял интегрировать временные метки в анализ, предоставляя подробное представление о динамике и последовательности движений глаз во времени. В результате, была создана матрица с более чем 24 тысячами признаков (Приложение А8).

Однако, несмотря на предполагаемые преимущества, данный подход также продемонстрировал ряд существенных недостатков:

- высокая размерность данных: Такое огромное количество признаков значительно усложняет процесс моделирования. Это ведет к увеличению вычислительных затрат и риска переобучения моделей. В машинном обучении переобучение — это ситуация, когда модель слишком точно описывает тренировочные данные, потеряв при этом способность обобщения на новые данные,

- управление данными: Обработка и анализ такого большого массива данных становится трудоемким заданием, особенно при ограниченных вычислительных ресурсах. Манипулирование столь громоздкими наборами данных может потребовать другого программного обеспечения и аппаратуры, а также может вызвать проблемы с производительностью и масштабируемостью системы.

Эти сложности могут затруднить или даже сделать невозможным эффективное использование стандартных алгоритмов машинного обучения, а также требуют дополнительных методов для сокращения размерности и оптимизации обработки данных, например, понижения размерности.

2.5.2 Гипотеза комплексного анализа динамики глазных движений

В рамках переосмысления подхода к анализу данных окулографии были внедрены усовершенствованные методы для детализированной оценки и классификации не только фиксаций глаз, но и саккад (Saccade_L и Saccade_R). Эти методы предназначены для более глубокого понимания динамики глазных движений при различных условиях чтения.

Разработанные функции позволяют анализировать следующие аспекты движений глаз:

- расстояние охвата (Distance Span): оценка общего перемещения глаз для каждого события, что помогает определить амплитуду движений во время фиксаций и саккад,

- средняя позиция (Average Position) и стандартное отклонение позиции (Std Position): эти метрики вычисляются для оценки положения и стабильности взгляда, соответственно, во время фиксаций и саккад,

- максимальный диапазон движения (Max Range) и Накопленное расстояние (Accumulated Distance): Предоставляют информацию о динамике и общей активности глазных движений.

Процесс включает создание подробных записей каждого события движения глаз, включая фиксации и саккады (пример с фиксациями представлен в Приложении А9). Для каждого типа события анализируются следующие параметры:

- продолжительность,
- расстояние охвата,
- средняя позиция,
- стандартное отклонение,
- максимальный диапазон,
- накопленное расстояние движений.

Собранные данные подвергаются статистическому анализу, который включает расчет средних значений, стандартных отклонений и медианных значений для каждого из измеренных параметров. Это позволяет провести сравнение и анализ поведения глаз между различными группами участников и сценариями чтения.

Этот усовершенствованный подход позволяет не только более точно классифицировать движения глаз, но и глубже понять механизмы, влияющие на различные виды глазных движений. Интеграция данных о фиксациях и саккадах обеспечивает более полное представление о когнитивных процессах, происходящих во время чтения и восприятия визуальной информации.

Эти методы анализа данных окулографии представляют значительный шаг в направлении улучшения понимания взаимосвязей между визуальным восприятием и когнитивной деятельностью, что является важным для разработки методик коррекции и обучения лиц с нарушениями чтения.

2.6 Применение моделей машинного обучения для анализа данных айтрекинга в диагностике дислексии

Процесс подготовки данных охватывает следующие этапы:

- загрузка данных: Исходные данные айтрекинга, содержащие предобработанные метрики, загружаются из CSV-файла,
- масштабирование признаков: для обеспечения одинакового масштаба признаков и предотвращения смещения в сторону переменных с большими абсолютными значениями применяется стандартизация.

Выбор и обучение модели включают несколько ключевых шагов:

- выбор признаков и разделение данных: набор данных разделяется на признаки (X) и целевую переменную (y), которые затем распределяются между обучающей и тестовой выборками,

— выбор модели: производится оценка нескольких алгоритмов, включая случайный лес, логистическую регрессию и системы опорных векторов (SVM).

— настройка гиперпараметров: используется GridSearchCV для оптимизации гиперпараметров, особенно для модели случайного леса.

Эффективность модели оценивается через несколько методов:

— кросс-валидация: Применение кросс-валидации для верификации обобщающей способности модели,

— интерпретация результатов: использование инструментов SHAP для объяснения предсказаний модели, выделение ключевых признаков,

— производительность на тестовой выборке: анализ точности, полноты, прецизионности и F1-меры, а также визуализация результатов с помощью матрицы ошибок.

Этот усовершенствованный подход позволяет не только более точно классифицировать движения глаз, но и глубже понять механизмы, влияющие на различные виды глазных движений. Интеграция данных о фиксациях и саккадах обеспечивает более полное представление о когнитивных процессах, происходящих во время чтения и восприятия визуальной информации.

2.7 Выбор инструментария для разработки системы

2.7.1 Язык программирования

Выбор инструментария для реализации основывается на требованиях к функциональности, производительности и доступности современных технологических решений.

В качестве языка программирования для разработки информационной системы был выбран язык Python по двум основным причинам:

— доступность: Python является бесплатным, с открытым исходным кодом, что делает его доступным для широкого круга разработчиков и исследователей,

— богатая экосистема: Python поддерживается обширной библиотекой пакетов, специализированных для анализа данных, машинного обучения и научных вычислений.

Этот усовершенствованный подход позволяет не только более точно классифицировать движения глаз, но и глубже понять механизмы, влияющие на различные виды глазных движений. Интеграция данных о фиксациях и саккадах обеспечивает более полное представление о когнитивных процессах, происходящих во время чтения и восприятия визуальной информации.

2.7.2 Выбор платформы для разработки

Google Colab был выбран как основная среда разработки благодаря своим уникальным возможностям и интеграции с широким спектром инструментов, доступных для исследователей и разработчиков.

Google Colab предлагает ряд значительных преимуществ, которые делают его идеальным выбором для анализа и работы в сфере машинного обучения:

— доступ к мощным вычислительным ресурсам: Google Colab предоставляет бесплатный доступ к вычислительным мощностям, включая GPU и TPU, что позволяет ускорить обработку данных и обучение моделей машинного обучения,

— нет необходимости в установке: как веб-платформа, Google Colab не требует установки программного обеспечения на локальный компьютер, что упрощает начало работы и исключает возможные конфликты зависимостей,

— совместимость с Jupyter Notebook: Colab основан на Jupyter Notebook, что обеспечивает удобную интеграцию с большинством инструментов анализа данных, применяемых в научном сообществе.

Выбор Google Colab обусловлен не только его техническими возможностями, но и удобством использования в академических и образовательных проектах. Платформа позволяет быстро перейти к анализу данных без затрат времени на подготовку и настройку среды разработки, что является ключевым фактором в динамичной научной среде.

Использование Google Colab ускоряет процесс разработки и тестирования моделей, предоставляя исследователям интуитивно понятные инструменты для непосредственного исполнения кода, визуализации данных. Эта платформа стимулирует новые подходы к анализу данных.

2.7.3 Выбор библиотек

Библиотеки Pandas и NumPy являются основными инструментами для обработки и анализа данных в Python, благодаря их мощным и гибким возможностям.

Библиотека Pandas предоставляет высокоуровневые структуры данных и широкие возможности для манипуляции данными, включая удобные средства для фильтрации, трансформации и агрегации данных. Она особенно полезна для работы с табличными данными, где необходимо эффективно управлять большими объемами данных, выполнять сложные операции слияния и группировки, а также предоставлять интуитивно понятный интерфейс для анализа временных рядов.

Библиотека NumPy является фундаментом для научных вычислений в Python. Она предоставляет поддержку мощных многомерных массивов и матриц, а также большой набор математических функций для операций с этими массивами. NumPy оптимизирует вычисления и значительно ускоряет

операции благодаря векторизации, что критически важно при обработке и анализе больших данных.

Выбор этих библиотек обусловлен не только их функциональностью, но и хорошей интеграцией с другими библиотеками и инструментами, используемыми в проекте, такими как Matplotlib и Seaborn для визуализации данных и Scikit-learn для машинного обучения. Это позволяет создавать мощные и масштабируемые решения для анализа данных айтирекинга.

Использование Pandas и NumPy способствует ускорению разработки за счет предоставления готовых к использованию функций и методов, которые значительно сокращают количество кода, необходимого для выполнения стандартных задач обработки данных. Это упрощает итерации проекта, позволяя разработчикам сосредоточиться на решении более сложных аналитических задач.

Scikit-learn является одним из наиболее популярных и мощных инструментов для машинного обучения в Python, в рамках разработки информационной системы для анализа данных. Scikit-learn предоставляет широкий спектр алгоритмов машинного обучения, предварительной обработки данных, выбора признаков и оценки моделей, что делает её идеальным выбором для проекта.

Scikit-learn известен своей высокой эффективностью и простотой в использовании, предоставляя следующие ключевые возможности:

- обширный набор алгоритмов: включает классификацию, регрессию, кластеризацию и уменьшение размерности, поддерживая как обучение с учителем, так и без учителя,

- инструменты для предварительной обработки данных: нормализация, стандартизация, обработка пропущенных значений и кодирование категориальных признаков,

— выбор и оценка моделей: поддержка кросс-валидации, настройки гиперпараметров (например, с помощью GridSearchCV) и оценки моделей с использованием различных метрик,

— интеграция с другими библиотеками: отлично работает с NumPy, Pandas и Matplotlib, создавая удобный рабочий поток для исследований и разработки.

В контексте анализа данных айтекинга Scikit-learn используется для следующих задач:

— классификация данных айтекинга: использование алгоритмов, таких как случайный лес, логистическая регрессия и машины опорных векторов (SVM), для идентификации и классификации паттернов, связанных с дислексией,

— оценка и сравнение моделей: проведение кросс-валидации для проверки надежности моделей, настройка гиперпараметров для оптимизации производительности и использование метрик, таких как точность, отзыв, F1-мера, для оценки результатов.

2.8 Выбор библиотек

На схеме, представленной ниже (рисунок 13) показана блок-схема алгоритма обработки данных

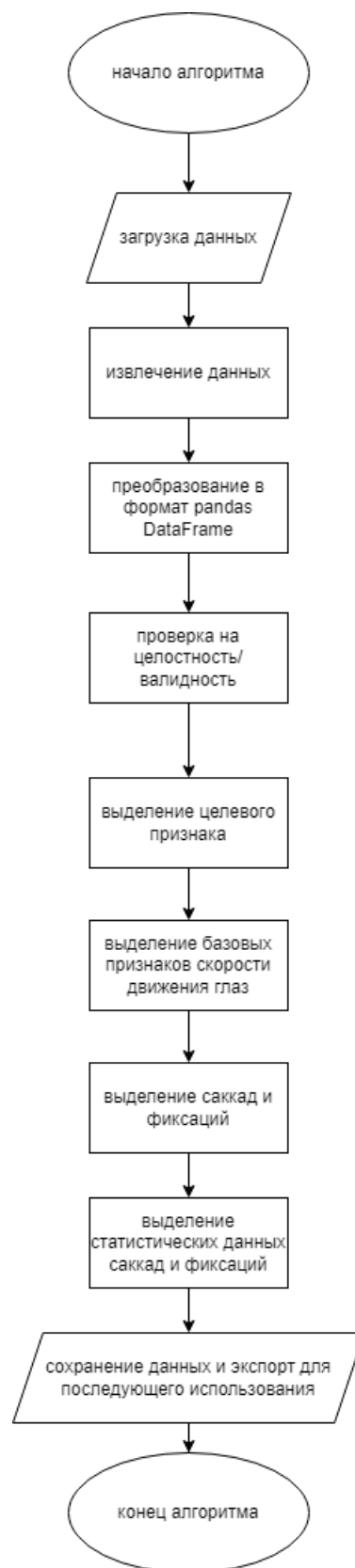


Рисунок 13 - Блок-схема обработки данных

Изображение представляет собой блок-схему алгоритма, которая последовательно демонстрирует этапы обработки и анализа данных айтрекинга. Ниже представлен обзор шагов, изображённых на схеме

- 1) начало алгоритма,
- 2) загрузка данных: этап, на котором происходит импорт данных.
- 3) извлечение данных: процесс извлечения необходимых данных из архива с загруженных источников,
- 4) преобразование в формат pandas DataFrame: конвертация данных в формат DataFrame библиотеки pandas для удобства последующей обработки и анализа,
- 5) проверка на целостность/валидность: проверка данных на наличие ошибок, пропущенных значений или других аномалий, которые могут повлиять на качество анализа,
- 6) выделение целевого признака: “status”,
- 7) выделение базовых признаков скорости движения глаз: определение и извлечение основных признаков, связанных со скоростью движения глаз, которые впоследствии будут использоваться для анализа и предсказаний.
- 8) выделение саккад и фиксаций: более специфический анализ данных айтрекинга для идентификации саккад (быстрых движений глаз) и фиксаций (моменты, когда взгляд зафиксирован),
- 9) выделение статистических данных саккад и фиксаций: расчет статистических показателей для данных саккад и фиксаций, что включает: средние значения, стандартные отклонения и медианные значения,
- 10) охранение данных и экспорт для последующего использования: Завершающий этап, на котором обработанные данные сохраняются в формате csv для дальнейшего использования и экспортируются во внешние системы,
- 11) конец алгоритма: завершение работы алгоритма.

Эта блок-схема является отличным инструментом для визуализации и демонстрации последовательности шагов в алгоритме обработки данных,

позволяя легко отслеживать процесс и убедиться, что все ключевые аспекты анализа учтены.

2.9 Выбор признаков для модели

В ходе анализа данных окулографии было выбрано 162 признака, которые обеспечивают комплексное представление о различных аспектах движения глаз. Эти признаки охватывают как параметры саккад (быстрые движения глаз между фиксациями), так и параметры фиксаций (периоды, когда взгляд остается статичным). Основные используемые признаки включают среднее значение, стандартное отклонение и медиану для длительности и расстояния саккад и фиксаций, а также средние позиции, стандартные отклонения и медианные значения для каждого из четырех измерений позиции глаз (LX, LY, RX, RY). Кроме того, были извлечены признаки, отражающие максимальный диапазон и накопленное расстояние движения глаз, что позволяет детально анализировать динамику взгляда и его стабильность в различных условиях наблюдения.

Эти данные были обработаны и подготовлены для дальнейшего использования в моделях машинного обучения, что позволит выявить потенциальные биомаркеры для диагностики дислексии, связанных с процессами визуального восприятия и когнитивной обработки информации. Выбор и обработка признаков основывалась на предварительном анализе и литературном обзоре, что обеспечило включение всех значимых переменных для достижения наивысшей точности и эффективности исследования.

2.10 Обобщение результатов

В данной части работы описывается процесс подготовки и анализа данных, использование методов машинного обучения и их применение для диагностики и изучения дислексии на основе данных окулографии. Исходные данные, взятые из статьи в PLOS ONE, структурированы и преобразованы в удобный для анализа формат DataFrame, что значительно упрощает обработку и позволяет эффективнее выявлять закономерности. В процессе предварительной обработки данных создаются новые признаки, которые критичны для оценки визуальной системы участников и могут помочь отличить лица с нарушениями чтения от контрольной группы.

С помощью визуализации данных удалось наглядно продемонстрировать динамику скорости движения глаз, что помогает выявить потенциальные биомаркеры дислексии. Особое внимание уделено переосмыслению подходов к анализу данных в свете сложностей, связанных с высокой размерностью данных, и предложены методы для оптимизации анализа. Выбор инструментов и платформы, таких как Python, Google Colab, Pandas и Scikit-learn, был обусловлен их функциональностью, производительностью и доступностью, что играет ключевую роль в эффективном анализе и разработке.

3 Результаты

3.1 Анализ распределения и корреляции целевой переменной

Цель этой главы – исследовать распределение целевой переменной 'status', которая указывает на наличие или отсутствие дислексии у участников, и анализировать связь этой переменной с признаками, извлеченными из данных окулографии.

Первым шагом является анализ распределения 'status' в исследуемой выборке. Для этого будет использована столбчатая диаграмма, которая показывает количество участников с дислексией и без. Это позволит оценить баланс классов и определить, есть ли необходимость в методах балансировки данных перед обучением моделей машинного обучения. На рисунке 14 видно, что распределение целевой переменной status почти равномерное, с небольшим преобладанием единиц.

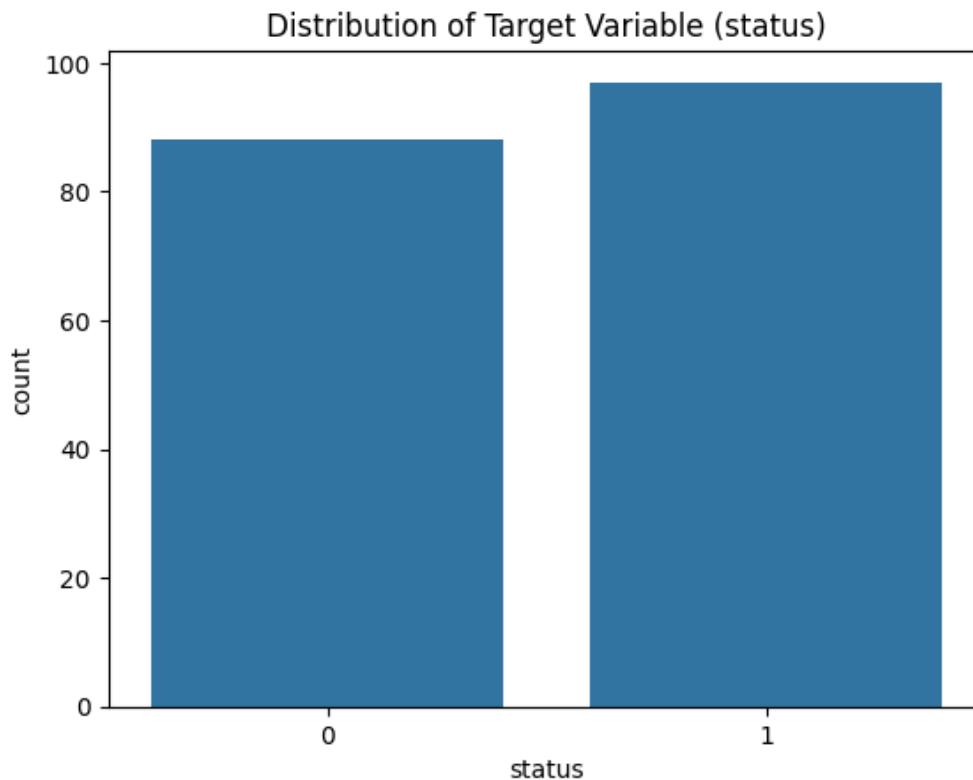


Рисунок 14 – Распределение переменной status

Для анализа связи между 'status' и признаками будет использован корреляционный анализ. Корреляционный анализ, представленный на рисунке 15 выявил, что только первые 14 признаков с начала и 12 признаков с конца списка заметно влияют на результат модели.

Для примера: чем больше длительность фиксации, тем выше вероятность того, что у наблюдаемого будет присутствовать нарушение.

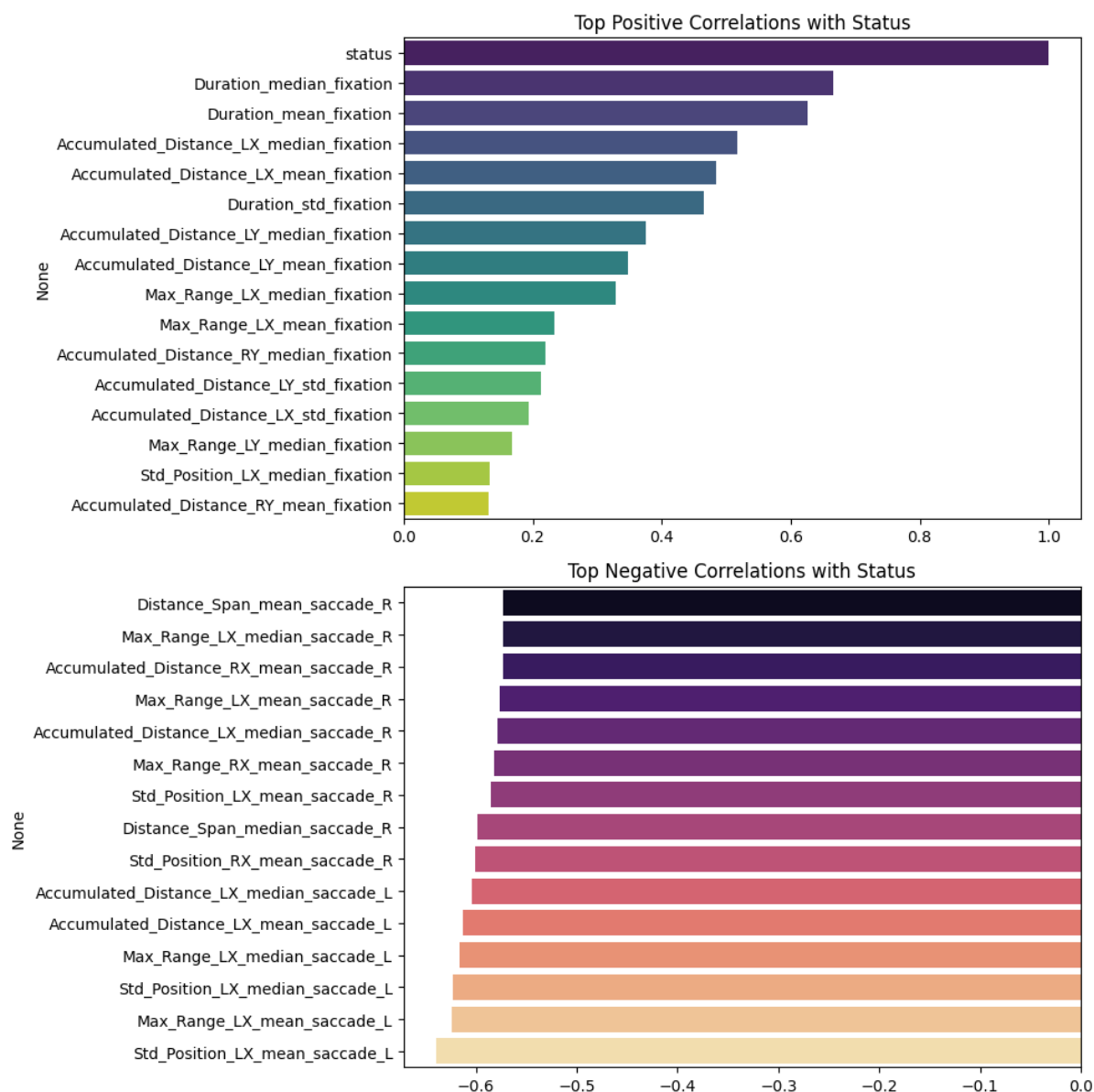


Рисунок 35 – корреляционный анализ

3.2 Первичное моделирование и последующая оптимизация

В ходе первоначального этапа, без углубленного применения разведочного анализа данных (EDA), была предложена логистическая регрессия - baseline-модель, без предварительной обработки данных, и подбора гиперпараметров, которая поможет установить минимальный уровень производительности, который будущие модели должны превзойти, чтобы быть полезными. Модель продемонстрировала удовлетворительные результаты. Результаты модели логистической регрессии представлены на рисунке 16.

```
Accuracy: 0.8378378378378378
      precision    recall  f1-score   support

     0       0.86      0.86      0.86         22
     1       0.80      0.80      0.80         15

   accuracy          0.84         37
  macro avg       0.83      0.83      0.83         37
weighted avg       0.84      0.84      0.84         37

Confusion Matrix:
[[19  3]
 [ 3 12]]
```

Рисунок 16 – Результат работы логистической регрессии. Без выделения основных признаков

Таким образом хорошо обученную модель можно использовать для помощи в разметке набора данных.

Однако, при анализе результатов, включая использование SHAP (результаты представлены на рисунке 17), было установлено, что большинство признаков не оказывают значимого влияния на результат модели и,

следовательно, только ухудшают целевую метрику точности (accuracy). Поскольку это был начальный этап, было решено провести повторный анализ данных с углубленным использованием EDA. Как и показал SHAP, большинство признаков оказывали незначительное влияние на целевую переменную.

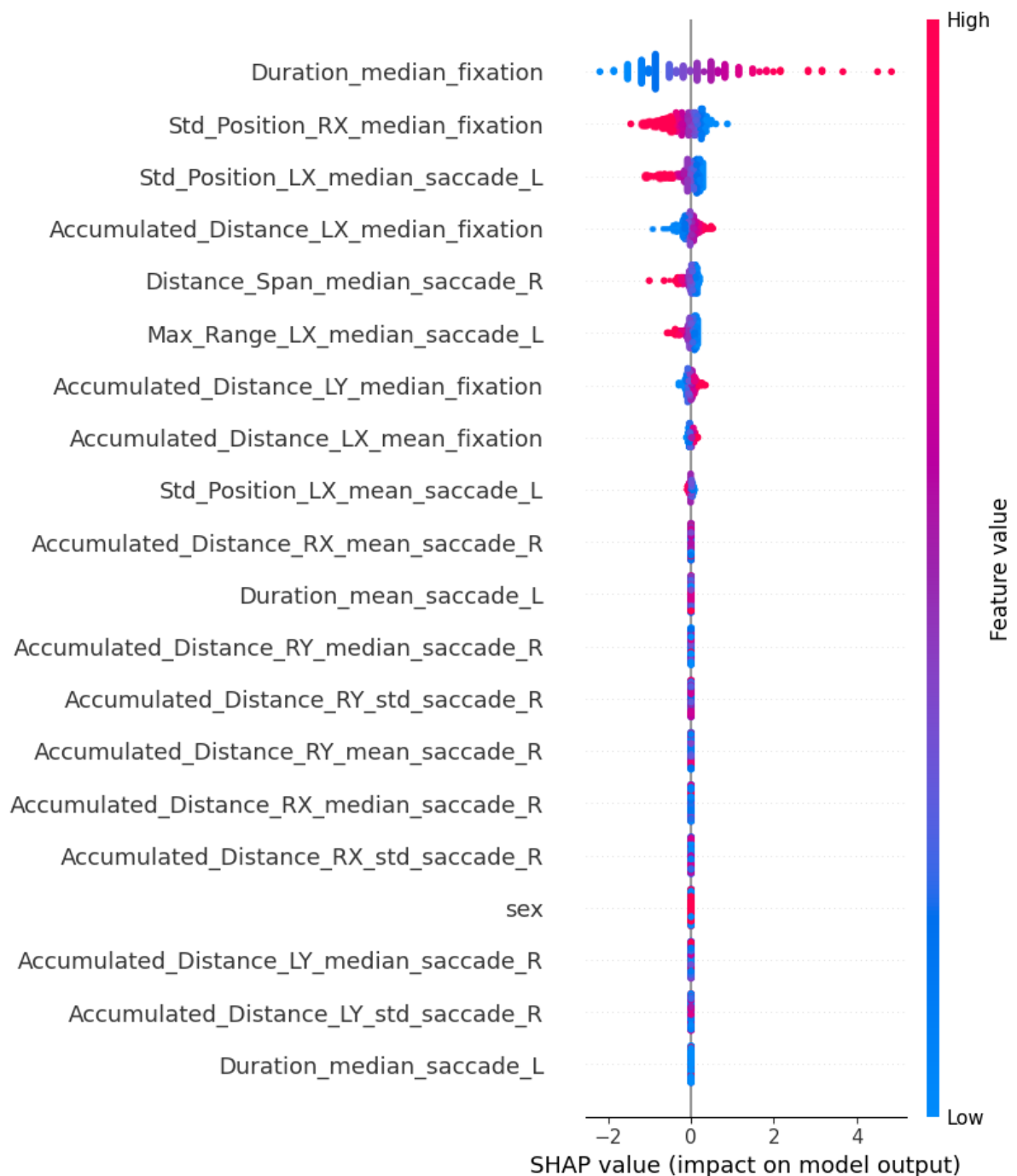


Рисунок 17 – SHAP анализ

Были также запущены модели (случайный лес, SVM, логистическая регрессия) с различными гиперпараметрами, но уже с ограниченным количеством признаков. В результате, показатели моделей значительно улучшились (Пример использования модели логистической регрессии с лучшими показателями представлен в приложении A10). Результаты логистической регрессии продемонстрированы на рисунке 18.

```
0.9107142857142857
[[26  1]
 [ 4 25]]
```

	precision	recall	f1-score	support
0	0.87	0.96	0.91	27
1	0.96	0.86	0.91	29
accuracy			0.91	56
macro avg	0.91	0.91	0.91	56
weighted avg	0.92	0.91	0.91	56

Рисунок 18 - Результат работы логистической регрессии. После
глубокого анализа

3.3 Сравнение различных моделей

Перед применением инструментов машинного обучения были выдвинуты следующие предположения:

— Базовые модели без предварительной обработки данных показывают ниже производительности по сравнению с моделями после разведочного анализа данных (EDA) и настройки гиперпараметров.

— Применение EDA и подбор гиперпараметров улучшают производительность моделей, что позволяет повысить точность, полноту и F1-меру.

— Различные типы моделей (логистическая регрессия, случайный лес, SVM) реагируют по-разному на одни и те же методы обработки данных.

Методы проверки:

— Базовая производительность измерялась сразу после деления данных на обучающую и тестовую выборки без какой-либо предварительной обработки.

— Применение EDA включало в себя анализ распределений, и нормализацию признаков.

— Настройка гиперпараметров проводилась с использованием методов, таких как кросс-валидация и поиск по сетке, для оптимизации параметров моделей.

Для анализа данных по движению глаз были отобраны признаки, которые больше всего влияли на целевую переменную. Включены медианные и средние значения длительности фиксаций, а также стандартные отклонения этих длительностей. Также анализировались накопленные расстояния по осям X и Y для медианных и средних фиксаций, а также для медианных саккад. Кроме того, учитывались максимальные диапазоны движений по осям X и Y для фиксаций и саккад, и стандартные отклонения позиций по оси X для медианных фиксаций. Среди признаков также присутствуют накопленные расстояния для средней саккады по оси X и медианной саккады по оси X. Все эти признаки были выбраны на основе их высокой корреляции с изучаемыми характеристиками и необходимы для проведения детального анализа данных.

Для гиперпараметров был использован scikit-learn GridSearchCV:

Для логистической регрессии лучшими гиперпараметрами вышли: {'C': 1, 'penalty': 'l2', 'solver': 'newton-cg'}

Для случайного леса: {'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 80}

Для опорных векторов: {'C': 1, 'kernel': 'rbf', 'gamma': 'scale'}

Сводная информация по всем моделям продемонстрирована в таблице 1.

Таблица 1 – Сводная таблица по моделям

		precision	recall	f1-score
Logistic Regression без обработки	accuracy			0,84 [0.75, 0.94]
	macro avg	0,83 [0.73, 0.92]	0,83 [0.73, 0.93]	0,83 [0.74, 0.93]
	weighted avg	0,84 [0.75, 0.93]	0,84 [0.73, 0.93]	0,84 [0.74, 0.93]
Logistic Regression отбор признаков	accuracy			0,9 [0.80, 1.00]
	macro avg	0,9 [0.80, 1.00]	0,9 [0.80, 1.00]	0,9 [0.80, 1.00]
	weighted avg	0,9 [0.80, 1.00]	0,9 [0.80, 1.00]	0,9 [0.80, 1.00]
Logistic Regression отбор признаков + подбор гиперпараметров	accuracy			0,91 [0.82, 1.00]
	macro avg	0,91 [0.81, 1.00]	0,91 [0.80, 1.00]	0,91 [0.81, 1.00]
	weighted avg	0,92 [0.82, 1.00]	0,91 [0.80, 1.00]	0,91 [0.82, 1.00]
Random Forest без обработки	accuracy			0,86 [0.76, 0.95]
	macro avg	0,86 [0.75, 0.95]	0,86 [0.76, 0.96]	0,86 [0.76, 0.95]
	weighted avg	0,86 [0.77, 0.96]	0,86 [0.76, 0.95]	0,86 [0.76, 0.95]
Random Forest отбор признаков	accuracy			0,89 [0.80, 0.99]
	macro avg	0,9 [0.80, 0.99]	0,9 [0.80, 1.00]	0,89 [0.80, 0.99]
	weighted avg	0,9 [0.81, 1.00]	0,89 [0.80, 0.99]	0,89 [0.79, 1.00]
Random Forest подбор гиперпараметров + отбор признаков	accuracy			0,9 [0.81, 1.00]
	macro avg	0,91 [0.81, 0.98]	0,91 [0.80, 0.98]	0,91 [0.81, 0.98]
	weighted avg	0,91 [0.81, 0.98]	0,9 [0.81, 0.98]	0,9 [0.81, 0.98]
SVM без обработки	accuracy			0,84 [0.74, 0.94]
	macro avg	0,84 [0.75, 0.93]	0,84 [0.75, 0.94]	0,84 [0.74, 0.94]
	weighted avg	0,84 [0.74, 0.94]	0,84 [0.75, 0.93]	0,84 [0.75, 0.94]
SVM отбор признаков	accuracy			0,86 [0.77, 0.96]
	macro avg	0,86 [0.76, 0.96]	0,86 [0.77, 0.95]	0,86 [0.76, 0.96]
	weighted avg	0,86 [0.77, 0.96]	0,86 [0.76, 0.96]	0,86 [0.76, 0.95]
SVM подбор гиперпараметров + отбор признаков	accuracy			0,89 [0.79, 0.99]
	macro avg	0,89 [0.79, 0.98]	0,89 [0.79, 0.99]	0,89 [0.78, 0.98]
	weighted avg	0,9 [0.80, 1.00]	0,89 [0.80, 0.99]	0,89 [0.79, 0.99]

После обработки данных был произведен анализ важности признаков, почему именно они влияют на целевую переменную. Были выдвинуты следующие предположения:

1) `Duration_median_fixation` и `Duration_mean_fixation` (1.741365, 1.523287): Эти показатели относятся к продолжительности фиксаций глаз. Результаты показали, что у людей с дислексией могут наблюдаться длинные фиксации из-за трудностей с обработкой текста. Большие значения этих коэффициентов подтверждают их важность в диагностировании дислексии.

2) `Accumulated_Distance_LX_median_fixation`, `Accumulated_Distance_LY_median_fixation`, и т.д. (0.731293, 0.698386): Эти признаки измеряют общее накопленное расстояние перемещения глаз по горизонтали (LX) и вертикали (LY) во время фиксаций. Такие движения могут указывать на необходимость частых коррекций взгляда при чтении, что является типичным для дислексии.

3) `Duration_std_fixation` (0.570489): Стандартное отклонение продолжительности фиксации может указывать на нестабильность в процессе чтения, что также может быть связано с дислексией.

4) `Distance_Span_mean_saccade_R` `Distance_Span_mean_saccade_L` (-0.586041 -0.615081): Среднее расстояние саккад вправо и влево. Отрицательный коэффициент может указывать на то, что меньшие саккады ассоциируются с дислексией, что может быть связано с трудностями в предварительной обработке текста перед чтением.

5) `Max_Range_LX_median_saccade_R`, `Max_Range_LX_median_saccade_L` (-0.491942, -0.305881): Максимальные амплитуды саккад, показывающие наибольшие перемещения глаз влево и вправо за одну саккаду. Отрицательные коэффициенты могут свидетельствовать о меньшей амплитуде движений глаз у дислектиков, что подтверждает проблемы с визуальной интерпретацией текста.

В ходе исследования на лучшей модели был проведен анализ важности признаков с использованием метода SHAP, представленный на рисунке 19, который помог идентифицировать наиболее значимые параметры для прогнозирования. Результаты показали, что длительность фиксации (Duration_median_fixation и Duration_mean_fixation) оказывает значительное влияние на предсказания модели, что подчеркивает ее роль в идентификации паттернов поведения глаз при дислексии.

Параметры, отражающие накопленное перемещение глаз (Accumulated_Distance_LX_median_fixation и Accumulated_Distance_LY_median_fixation), также демонстрировали высокую степень влияния, указывая на значимость анализа пространственных паттернов движения глаз для классификации состояний.

Эти результаты подтверждают предположение о том, что определенные характеристики глазных движений могут быть ключевыми индикаторами при скрининге дислексии, и предоставляют ценную информацию для дальнейшего уточнения алгоритмов машинного обучения в области биомедицинских исследований.

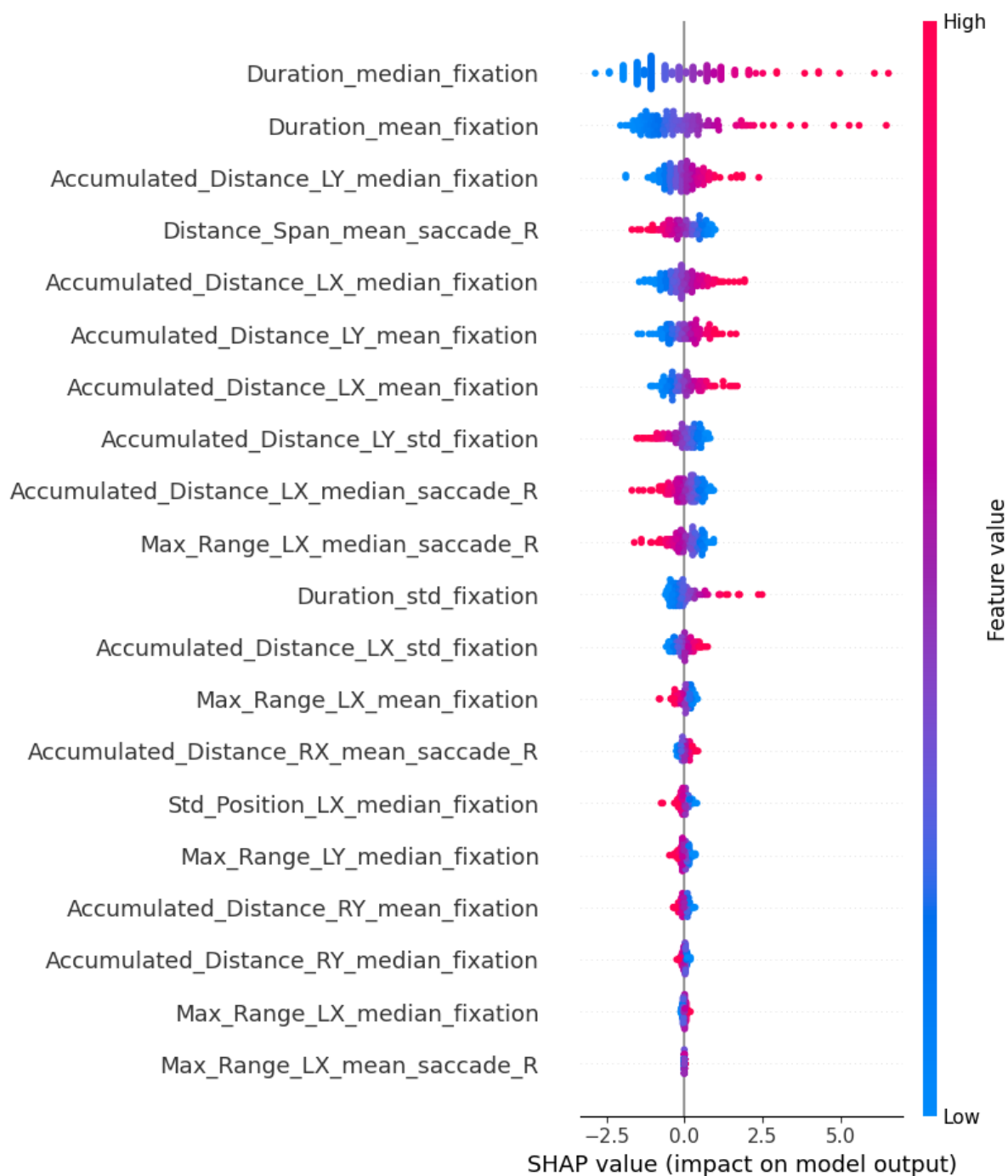


Рисунок 18 - SHAP анализ на лучшей модели.

Все модели показали улучшение производительности после EDA и настройки гиперпараметров. Это подтверждает гипотезу, что предварительная обработка данных и оптимизация параметров могут значительно повысить качество моделей.

Наибольшее улучшение наблюдается в моделях с подбором гиперпараметров и EDA, что указывает на важность комплексного подхода к подготовке данных и настройке моделей.

Различные модели показали разное улучшение, что подчеркивает необходимость выбора адекватной модели и подхода к обработке данных в зависимости от задачи.

3.4 Заключительные замечания по анализу и моделированию

В третьей главе проанализирована целевая переменная 'status', что позволило оценить распределение классов среди участников. Важно знать, что баланс классов важен для точности моделей, поскольку дисбаланс может привести к ошибочным предсказаниям, особенно в контексте медицинских диагностических задач. Корреляционный анализ выявил ряд ключевых признаков, оказывающих значимое влияние на статус дислексии, подчеркивая их значимость при формировании надежных и обоснованных прогностических моделей.

Оптимизация моделей демонстрирует значительное улучшение их производительности. Применение SHAP позволило не только идентифицировать признаки с наибольшим влиянием на модели, но и улучшить интерпретируемость моделей. Оптимизированные модели демонстрируют улучшенную способность точно классифицировать случаи дислексии.

Сравнение различных подходов к моделированию подчеркивает важность разведочного анализа данных (EDA) и настройки гиперпараметров. Базовые модели без предварительной обработки данных показывали относительно низкую производительность, в то время как модели, оптимизированные с помощью EDA и настройки гиперпараметров, улучшили свои показатели. Это подтверждает, что подход к подготовке данных и

настройке моделей должен быть комплексным и тщательно продуманным для достижения наилучших результатов.

ЗАКЛЮЧЕНИЕ

В ходе проведения обзора предметной области и литературы были рассмотрены научные исследования и публикации по применению методов машинного обучения в анализе данных окулографии. Рассмотрение существующих методик позволило принять во внимание основные инструменты исследований, что позволило разобраться в сопутствующей теме.

В процессе работы была разработана методология для обработки и анализа данных айтрекинга с использованием техник машинного обучения. Основное внимание было уделено созданию и выявлению новых признаков, таких параметров, как длительность фиксаций, частота и амплитуда саккад, а также паттерны перемещения взгляда между точками интереса на экране. Эти параметры были проанализированы с целью определения их влияния на уровень внимания и когнитивные процессы, связанные с чтением и восприятием текста, что помогло эффективно классифицировать и предсказывать наличие или отсутствие дислексии.

В рамках исследования было проверено четыре основных парадигм работы с признаками:

1) Исключение неинформативных признаков: Отбор признаков путем удаления тех, которые имеют малую вариативность или низкую корреляцию с целевой переменной.

2) Инжиниринг признаков: Создание новых признаков из существующих данных.

3) Нормализация и стандартизация: Преобразование признаков для обеспечения одинакового масштаба.

4) Балансировка классов: Применение методов анализа количества объектов в классах.

И три модели машинного обучения с последующей их модификацией, для повышения результатов:

- 1) Логистическая регрессия
- 2) Случайный лес (Random Forest)
- 3) Метод опорных векторов (SVM)

Лучшие результаты были достигнуты благодаря оптимизации моделей с использованием разведочного анализа данных и настройки гиперпараметров, что позволило улучшить метрики по сравнению с базовыми моделями.

Анализ полученных результатов показал, что разработанные модели машинного обучения справляются с задачей идентификации дислексии на основе данных айтрекинга. Особенно важным стало применение методов интерпретации машинного обучения, таких как SHAP, которые позволили не только подтвердить значимость отдельных признаков, но и улучшить понимание механизмов, лежащих в основе дислексии.

Это исследование способствует пониманию механизмов взаимосвязи когнитивных функций и движения глаз, раскрывая новые аспекты в понимании дислексии.

Для дальнейших исследований следует расширить объём данных, наблюдений, включая новые метрики и параметры окулографии, а также протестировать альтернативные модели машинного обучения для повышения точности классификации. Кроме того, следует изучить влияние различных условий тестирования на показатели движения глаз, что может предоставить дополнительные данные для более точного определения когнитивных нарушений.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Anwar S. N. S. S., Abd Aziz A., Adil S. H. Development of real-time eye tracking algorithm //2021 4th International Conference on Computing & Information Sciences (ICCIS). – IEEE, 2021. – С. 1-6.
2. Wei X., Li C. T. Fixation and saccade based face recognition from single image per person with various occlusions and expressions //Proceedings of the IEEE conference on computer vision and pattern recognition workshops. – 2013. – С. 70-75.
3. Alkan Y. et al. Functional connectivity in vergence and saccade eye movement tasks assessed using Granger Causality Analysis //2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. – IEEE, 2011. – С. 8114-8117.
4. Julian I. et al. Combining UEQ and Eye-Tracking Method as Usability Evaluation for Mobile Apps //2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS). – IEEE, 2021. – С. 1-6.
5. Huey E. B. The psychology and pedagogy of reading. 1908. Reprint. – 1968.
6. Buswell G. T. Fundamental reading habits: A study of their development. – University of Chicago, 1922. – №. 21.
7. Yarbus A. L. Eye movements and vision. – Springer, 2013.
8. Just M. A., Carpenter P. A. A theory of reading: from eye fixations to comprehension //Psychological review. – 1980. – Т. 87. – №. 4. – С. 329.
9. Posner, M.I. // Please Scroll down for Article the Quarterly Journal of Experimental Psychology Orienting of Attention Orienting of Attention.
10. Radach R., Hyona J., Deubel H. (ed.). The mind's eye: Cognitive and applied aspects of eye movement research. – Elsevier, 2003.
11. Cognolato M., Atzori M., Müller H. Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances //Journal of

rehabilitation and assistive technologies engineering. – 2018. – T. 5. – C. 2055668318773991.

12. Zhao L. et al. Eye state recognition based on deep integrated neural network and transfer learning //Multimedia Tools and Applications. – 2018. – T. 77. – C. 19415-19438.

13. Rahamathunnisa U., Sudhakar K. Analysis on texture feature extraction methods for face recognition in new born //2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). – IEEE, 2021. – C. 894-897.

14. Mitiche L., Adamou-Mitiche A. B. H., Naimi H. Medical image denoising using dual tree complex thresholding wavelet transform //2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). – IEEE, 2013. – C. 1-5.

15. Rehman A. et al. Performance analysis of pca, sparse pca, kernel pca and incremental pca algorithms for heart failure prediction //2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). – IEEE, 2020. – C. 1-5.

16. Cao Y. et al. Pupil variation applied to the eye tracking control of an endoscopic manipulator //IEEE Robotics and Automation Letters. – 2016. – T. 1. – №. 1. – C. 531-538.

17. Labibah Z., Nasrun M., Setianingsih C. Lie detector with the analysis of the change of diameter pupil and the eye movement use method Gabor wavelet transform and decision tree //2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS). – IEEE, 2018. – C. 214-220.

18. Bozkir E., Geisler D., Kasneci E. Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup //2019 IEEE conference on virtual reality and 3D user interfaces (VR). – IEEE, 2019. – C. 1834-1837.

19. Orlosky J., Huynh B., Hollerer T. Using eye tracked virtual reality to classify understanding of vocabulary in recall tasks //2019 IEEE International

Conference on Artificial Intelligence and Virtual Reality (AIVR). – IEEE, 2019. – C. 66-667.

20. Slanzi G., Balazs J. A., Velásquez J. D. Combining eye tracking, pupil dilation and EEG analysis for predicting web users click intention //Information Fusion. – 2017. – T. 35. – C. 51-57.

21. Guo J. J. et al. Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks //2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). – IEEE, 2019. – C. 3071-3074.

22. Zhang Y., Juhola M. On biometrics with eye movements //IEEE journal of biomedical and health informatics. – 2016. – T. 21. – №. 5. – C. 1360-1366.

23. Sargezeh B. A., Tavakoli N., Daliri M. R. Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study //Physiology & behavior. – 2019. – T. 206. – C. 43-50.

24. Tamuly S., Jyotsna C., Amudha J. Tracking eye movements to predict the valence of a scene //2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). – IEEE, 2019. – C. 1-7.

25. Ahmed M., Noble J. A. Fetal ultrasound image classification using a bag-of-words model trained on sonographers' eye movements //Procedia Computer Science. – 2016. – T. 90. – C. 157-162.

26. Zhou X. et al. Eye tracking data guided feature selection for image classification //Pattern Recognition. – 2017. – T. 63. – C. 56-70.

27. Qi L. et al. Fusing visual saliency for material recognition //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. – 2018. – C. 1965-1968.

28. Singh M., Walia G. S., Goswami A. Using supervised learning to guide the selection of software inspectors in industry //2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). – IEEE, 2018. – C. 12-17.

29. Luo X. et al. Interested object detection based on gaze using low-cost remote eye tracker //2019 9th International IEEE/EMBS Conference on Neural Engineering (NER). – IEEE, 2019. – C. 1101-1104.
30. Startsev M., Dorr M. Classifying autism spectrum disorder based on scanpaths and saliency //2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). – IEEE, 2019. – C. 633-636.
31. Zhu J. et al. Toward depression recognition using EEG and eye tracking: an ensemble classification model CBEM //2019 IEEE international conference on bioinformatics and biomedicine (BIBM). – IEEE, 2019. – C. 782-786.
32. Vidyapu S., Vedula V. S., Bhattacharya S. Attention prediction on webpage images using multilabel classification //2019 IEEE international conference on systems, man and cybernetics (SMC). – IEEE, 2019. – C. 1975-1980.
33. Roy A. K. et al. A novel technique to develop cognitive models for ambiguous image identification using eye tracker //IEEE Transactions on Affective Computing. – 2017. – T. 11. – №. 1. – C. 63-77.
34. Shi Z. F. et al. Attention evaluation with eye tracking glasses for EEG-based emotion recognition //2017 8th International IEEE/EMBS Conference on Neural Engineering (NER). – IEEE, 2017. – C. 86-89.
35. Czyzewski A. et al. Multimodal approach for polysensory stimulation and diagnosis of subjects with severe communication disorders //Procedia computer science. – 2017. – T. 121. – C. 238-243.
36. Jiang M. et al. Classifying individuals with ASD through facial emotion recognition and eye-tracking //2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). – IEEE, 2019. – C. 6063-6068.
37. Thapaliya S., Jayarathna S., Jaime M. Evaluating the EEG and eye movements for autism spectrum disorder //2018 IEEE international conference on big data (Big Data). – IEEE, 2018. – C. 2328-2336.

38. Ding X. et al. Classifying major depression patients and healthy controls using EEG, eye tracking and galvanic skin response data //Journal of affective Disorders. – 2019. – T. 251. – C. 156-161.
39. Abdelrahman Y. et al. Classifying attention types with thermal imaging and eye tracking //Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. – 2019. – T. 3. – №. 3. – C. 1-27.
40. Koza J. R. et al. Automated design of both the topology and sizing of analog electrical circuits using genetic programming //Artificial intelligence in design'96. – 1996. – C. 151-170.
41. Borys M. et al. Using machine learning models to classify user performance in the ruff figural fluency test from eye-tracking features //2017 International Conference on Electromagnetic Devices and Processes in Environment Protection with Seminar Applications of Superconductors (ELMECO & AoS). – IEEE, 2017. – C. 1-4.
42. Guo Y. et al. Eye-tracking for performance evaluation and workload estimation in space telerobotic training //IEEE Transactions on Human-Machine Systems. – 2021. – T. 52. – №. 1. – C. 1-11.
43. Ha J. et al. A hybrid brain–computer interface for real-life meal-assist robot control //Sensors. – 2021. – T. 21. – №. 13. – C. 4578.
44. Kacur J., Polec J., Csóka F. Eye tracking and KNN based detection of schizophrenia //2019 International Symposium ELMAR. – IEEE, 2019. – C. 123-126.
45. Yoo S., Jeong D. K., Jang Y. The study of a classification technique for numeric gaze-writing entry in hands-free interface //IEEE Access. – 2019. – T. 7. – C. 49125-49134.
46. Wilbertz G. et al. Combined fMRI-and eye movement-based decoding of bistable plaid motion perception //NeuroImage. – 2018. – T. 171. – C. 190-198.
47. Lin X. et al. Development of a high-speed mental spelling system combining eye tracking and SSVEP-based BCI with high scalability //2019 41st

annual international conference of the IEEE engineering in medicine and biology society (EMBC). – IEEE, 2019. – C. 6318-6322.

48. Lagodzinski P., Shirahama K., Grzegorzec M. Codebook-based electrooculography data analysis towards cognitive activity recognition //Computers in biology and medicine. – 2018. – T. 95. – C. 277-287.

49. Kubacki A. Use of force feedback device in a hybrid brain-computer interface based on SSVEP, EOG and eye tracking for sorting items //Sensors. – 2021. – T. 21. – №. 21. – C. 7244.

50. Song K., Zhou L., Wang H. Deep coupling recurrent auto-encoder with multi-modal EEG and EOG for vigilance estimation //Entropy. – 2021. – T. 23. – №. 10. – C. 1316.

51. Zheng W. L., Lu B. L. A multimodal approach to estimating vigilance using EEG and forehead EOG //Journal of neural engineering. – 2017. – T. 14. – №. 2. – C. 026017.

52. Horng G. J., Lin J. Y. Using multimodal bio-signals for prediction of physiological cognitive state under free-living conditions //IEEE Sensors Journal. – 2019. – T. 20. – №. 8. – C. 4469-4484.

53. Karessli N. et al. Gaze embeddings for zero-shot image classification //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2017. – C. 4525-4534.

54. Saab K. et al. Observational supervision for medical image classification using gaze data //Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. – Springer International Publishing, 2021. – C. 603-614.

55. Nilsson Benfatto M. et al. Screening for dyslexia using eye tracking during reading //PloS one. – 2016. – T. 11. – №. 12. – C. e0165508.

56. Hanisch S. et al. Privacy-protecting techniques for behavioral data: A survey //arXiv preprint arXiv:2109.04120. – 2021.

57. Yarbus A. L. Eye movements and vision. – Springer, 2013.

ПРИЛОЖЕНИЕ А

A1 Преобразование исходных данных в DataFrame pandas

```
import pandas as pd
import os

zip_file = "/content/RecordingData.zip"
extract_to = "/content/eye-tracking"

with zipfile.ZipFile(zip_file, 'r') as zip_ref:
    zip_ref.extractall(extract_to)

directory = '/content/eye-tracking/Recording Data/'
combined_df = pd.DataFrame()

for dir_name in os.listdir(directory):
    subdir_path = os.path.join(directory, dir_name)
    if os.path.isdir(subdir_path):
        file_path = os.path.join(subdir_path, 'A1R.txt')
        if os.path.exists(file_path):
            df = pd.read_csv(file_path, delimiter='\t')
            df['dir_name'] = dir_name
            combined_df = pd.concat([combined_df, df],
                                    ignore_index=True)

combined_df.to_csv('recorded_pandas.csv', index=False)
```

A2 Добавление колонок status и sex

```
filter_condition = df['dir_name'].str.endswith(('1', '2'))
df.loc[filter_condition, 'status'] = 1 # Reading disable
df.loc[~filter_condition, 'status'] = 0
df['status'] = df['status'].astype(int)

filter_condition = df['dir_name'].str.endswith(('1', '3'))
df.loc[filter_condition, 'sex'] = 1 # Male
df.loc[~filter_condition, 'sex'] = 0
df['sex'] = df['sex'].astype(int)

df[['LX', 'LY', 'RX', 'RY']] = df[['LX', 'LY', 'RX',
'RY']].replace(',', '.', regex=True).astype(float)
df
```

A3 Добавление колонок Speed_L и Speed_R для 1 наблюдения

```
df1 = df[df['dir_name'] == '111RP1'].copy() # reading disable
# амплитуда
df1['SpeedL'] = ((df1['LX'].diff() ** 2 + df1['LY'].diff() ** 2)
** 0.5) / df1['T'].diff()
df1['SpeedR'] = ((df1['RX'].diff() ** 2 + df1['RY'].diff() ** 2)
** 0.5) / df1['T'].diff()
df1
```

A4 Построение графиков скорости левого и правого глаза для 1 наблюдения

```
# Построение графика скорости движения левого глаза
plt.figure(figsize=(10, 6))
plt.plot(df1['T'], df1['SpeedL'], color='blue', label='Speed of
left eye')
plt.xlabel('Time')
plt.ylabel('Speed')
plt.title('Speed of left eye over time')
plt.legend()
plt.grid(True)
plt.show()

# Построение графика скорости движения правого глаза
plt.figure(figsize=(10, 6))
plt.plot(df1['T'], df1['SpeedR'], color='red', label='Speed of
right eye')
plt.xlabel('Time')
plt.ylabel('Speed')
plt.title('Speed of right eye over time')
plt.legend()
plt.grid(True)
plt.show()
```

A5 Скорости движения левого и правого глаз с выделением признаков

```
saccade_threshold = 0.3
sweep_threshold = 2.5

saccades_L = df1[(df1['SpeedL'] > saccade_threshold) &
(df1['SpeedL'] < sweep_threshold)]
saccades_R = df1[(df1['SpeedR'] > saccade_threshold) &
(df1['SpeedR'] < sweep_threshold)]
```

```

sweep_move_L = df1[df1['SpeedL'] > sweep_threshold]
sweep_move_R = df1[df1['SpeedR'] > sweep_threshold]

plt.figure(figsize=(12, 6))
plt.plot(df1['T'], df1['SpeedL'], color='blue', label='Speed of
Left Eye')
plt.plot(df1['T'], df1['SpeedR'], color='red', label='Speed of
Right Eye')
plt.scatter(saccades_L['T'], saccades_L['SpeedL'],
color='green', label='Saccades (Left Eye)')
plt.scatter(saccades_R['T'], saccades_R['SpeedR'],
color='orange', label='Saccades (Right Eye)')
plt.scatter(sweep_move_L['T'], sweep_move_L['SpeedL'],
color='cyan', label='Sweep_movements (Left Eye)')
plt.scatter(sweep_move_R['T'], sweep_move_R['SpeedR'],
color='purple', label='Sweep_movements (Right Eye)')
plt.xlabel('Time')
plt.ylabel('Speed')
plt.title('Eye Movement and Detected Saccades')
plt.legend()
plt.grid(True)
plt.show()

```

A6 Создание новых бинарных признаков для 1 наблюдения

```

df1['Saccade_L'] = (df1['SpeedL'] > saccade_threshold) &
(df1['SpeedL'] < sweep_threshold)
df1['Saccade_R'] = (df1['SpeedR'] > saccade_threshold) &
(df1['SpeedR'] < sweep_threshold)
df1['Sweep_L'] = df1['SpeedL'] > sweep_threshold
df1['Sweep_R'] = df1['SpeedR'] > sweep_threshold
df1['Fixation'] = (df1['SpeedL'] < saccade_threshold) &
(df1['SpeedR'] < saccade_threshold)
df1

```

A7 Сравнение графиков для наблюдений с нарушением и без нарушений

```

import matplotlib.pyplot as plt
import pandas as pd

plt.rcParams['font.size'] = 14
plt.rcParams['font.family'] = 'DejaVu Sans'
df1_limited = df1[df1['T']]

plt.figure(figsize=(10, 6))
plt.plot(df1_limited['T'], df1_limited['SpeedL'], color='blue',
label='Скорость левого глаза (df1)')

```

```

plt.plot(df1_limited['T'], df1_limited['SpeedR'], color='red',
label='Скорость правого глаза (df1)')
plt.xlabel('Время')
plt.ylabel('Скорость')
plt.title('Чтение с ограничениями (df1)')
plt.legend(loc='upper right')
plt.grid(True)
plt.show()

df2_limited = df2[df2['T']]

plt.figure(figsize=(10, 6))
plt.plot(df2_limited['T'], df2_limited['SpeedL'], color='green',
label='Скорость левого глаза (df2)')
plt.plot(df2_limited['T'], df2_limited['SpeedR'],
color='orange', label='Скорость правого глаза (df2)')
plt.xlabel('Время')
plt.ylabel('Скорость')
plt.title('Чтение без ограничений (df2)')
plt.legend(loc='upper right')
plt.grid(True)
plt.show()

```

A8 Временная индексации признаков движения глаз

```

saccade_threshold = 0.3
sweep_threshold = 2.5

for i in dir_names:
    lol = dfs_by_dir_name[i]
    lol['SpeedL'] = ((lol['LX'].diff() ** 2 + lol['LY'].diff() **
2) ** 0.5) / lol['T'].diff()
    lol['SpeedR'] = ((lol['RX'].diff() ** 2 + lol['RY'].diff() **
2) ** 0.5) / lol['T'].diff()
    lol['SpeedDiff'] = lol['SpeedL'] - lol['SpeedR']
    lol['Saccade_L'] = (lol['SpeedL'] > saccade_threshold) &
(lol['SpeedL'] < sweep_threshold)
    lol['Saccade_R'] = (lol['SpeedR'] > saccade_threshold) &
(lol['SpeedR'] < sweep_threshold)
    lol['Sweep_L'] = lol['SpeedL'] > sweep_threshold
    lol['Sweep_R'] = lol['SpeedR'] > sweep_threshold
    lol['Fixation'] = (lol['SpeedL'] < saccade_threshold) &
(lol['SpeedR'] < saccade_threshold)

dir_names = df_all['dir_name'].unique()
grouped = df_all.groupby('dir_name')

```

```

dfs_by_dir_name = {name: group for name, group in grouped}

columns_to_expand = ['LX', 'LY', 'RX', 'RY', 'SpeedL', 'SpeedR',
                    'SpeedDiff', 'Saccade_L', 'Saccade_R',
                    'Sweep_L', 'Sweep_R', 'Fixation']

merged_data = pd.DataFrame()

# Цикл по словарю с датафреймами
for dir_name, df in dfs_by_dir_name.items():
    new_columns = {}

    time_steps = df['T'].unique()

    for column in columns_to_expand:
        for step in time_steps:
            new_column_name = f'{column}{step}'
            new_columns[new_column_name] = df.loc[df['T'] ==
step, column].head(1).values

    new_data = pd.DataFrame(new_columns)

    for column in ['dir_name', 'status', 'sex']:
        new_data[column] = df[column].iloc[0]

    new_data['dir_name'] = dir_name

    merged_data = pd.concat([merged_data, new_data],
ignore_index=True)

merged_data

```

А9. Статистический анализ признаков по всем наблюдениям

```

df_all = pd.read_csv('/content/recorded_pandas.csv')

def distance_span(event):
    span_x_left = np.max(event['LX']) - np.min(event['LX'])
    span_y_left = np.max(event['LY']) - np.min(event['LY'])
    span_x_right = np.max(event['RX']) - np.min(event['RX'])
    span_y_right = np.max(event['RY']) - np.min(event['RY'])
    total_span = span_x_left + span_y_left + span_x_right +
span_y_right
    return total_span

def average_position(df):

```

```

    average_LX = df['LX'].mean()
    average_LY = df['LY'].mean()
    average_RX = df['RX'].mean()
    average_RY = df['RY'].mean()
    return average_LX, average_LY, average_RX, average_RY

def std_position(event):
    std_lx = np.std(event['LX'])
    std_ly = np.std(event['LY'])
    std_rx = np.std(event['RX'])
    std_ry = np.std(event['RY'])
    return std_lx, std_ly, std_rx, std_ry

def max_range(event):
    max_range_lx = np.max(event['LX']) - np.min(event['LX'])
    max_range_ly = np.max(event['LY']) - np.min(event['LY'])
    max_range_rx = np.max(event['RX']) - np.min(event['RX'])
    max_range_ry = np.max(event['RY']) - np.min(event['RY'])
    return max_range_lx, max_range_ly, max_range_rx,
max_range_ry

def accumulated_distance(event):
    acc_distance_lx = np.sum(np.abs(np.diff(event['LX'])))
    acc_distance_ly = np.sum(np.abs(np.diff(event['LY'])))
    acc_distance_rx = np.sum(np.abs(np.diff(event['RX'])))
    acc_distance_ry = np.sum(np.abs(np.diff(event['RY'])))
    return acc_distance_lx, acc_distance_ly, acc_distance_rx,
acc_distance_ry

dir_names = df_all['dir_name'].unique()
grouped = df_all.groupby('dir_name')
dfs_by_dir_name = {name: group for name, group in grouped}

all_fixation_dfs = []

for i in dir_names:
    df = dfs_by_dir_name[i]
    fixation_events = []
    start_time = None

    for index, row in df.iterrows():
        if row['Fixation'] == True and start_time is None:
            start_time = row['T']
        elif row['Fixation'] == False and start_time is not None:
            end_time = row['T']

```

```

        distance = distance_span(df[(df['T'] >= start_time) &
(df['T'] <= end_time)])
        average_LX, average_LY, average_RX, average_RY =
average_position(df[(df['T'] >= start_time) & (df['T'] <=
end_time)])
        std_lx, std_ly, std_rx, std_ry =
std_position(df[(df['T'] >= start_time) & (df['T'] <=
end_time)])
        max_range_lx, max_range_ly, max_range_rx, max_range_ry
= max_range(df[(df['T'] >= start_time) & (df['T'] <= end_time)])
        acc_distance_lx, acc_distance_ly, acc_distance_rx,
acc_distance_ry = accumulated_distance(df[(df['T'] >=
start_time) & (df['T'] <= end_time)])

        fixation_events.append({'Start_Time': start_time,
                                'End_Time': end_time,
                                'Duration': end_time -
start_time,
                                'Distance_Span': distance,
                                'Average_Position_LX':
average_LX,
                                'Average_Position_LY':
average_LY,
                                'Average_Position_RX':
average_RX,
                                'Average_Position_RY':
average_RY,
                                'Std_Position_LX': std_lx,
                                'Std_Position_LY': std_ly,
                                'Std_Position_RX': std_rx,
                                'Std_Position_RY': std_ry,
                                'Max_Range_LX': max_range_lx,
                                'Max_Range_LY': max_range_ly,
                                'Max_Range_RX': max_range_rx,
                                'Max_Range_RY': max_range_ry,
                                'Accumulated_Distance_LX':
acc_distance_lx,
                                'Accumulated_Distance_LY':
acc_distance_ly,
                                'Accumulated_Distance_RX':
acc_distance_rx,
                                'Accumulated_Distance_RY':
acc_distance_ry})
        start_time = None

    fixation_events_df = pd.DataFrame(fixation_events)

```

```

fixation_df = pd.DataFrame()

for column in fixation_events_df.columns[2:]:
    fixation_df[column + '_mean_fixation'] =
[fixation_events_df[column].mean()]
    fixation_df[column + '_std_fixation'] =
[fixation_events_df[column].std()]
    fixation_df[column + '_median_fixation'] =
[fixation_events_df[column].median()]

for column in ['dir_name', 'status', 'sex']:
    fixation_df[column] = df[column].iloc[0]
all_fixation_dfs.append(fixation_df)

```

A10 Модель логистической регрессии с лучшими результатами

```

merged_df =
pd.read_csv('/content/drive/MyDrive/URFU/Диплом/данные/data_for_
models.csv')
merged_df = merged_df.set_index('dir_name')
merged_df.drop(columns=['dir_name'], inplace=True)

features_to_scale = merged_df.drop(columns=['status',
'sex']).columns
merged_df[features_to_scale] =
scaler.fit_transform(merged_df[features_to_scale])

data = merged_df

# Выбор признаков с наибольшей корреляцией
selected_features = [
    'Duration_median_fixation',
    'Duration_mean_fixation',
    'Accumulated_Distance_LX_median_fixation',
    'Accumulated_Distance_LX_mean_fixation',
    'Duration_std_fixation',
    'Accumulated_Distance_LY_median_fixation',
    'Accumulated_Distance_LY_mean_fixation',
    'Max_Range_LX_median_fixation',
    'Max_Range_LX_mean_fixation',
    'Accumulated_Distance_RY_median_fixation',
    'Accumulated_Distance_LY_std_fixation',
    'Accumulated_Distance_LX_std_fixation',
    'Max_Range_LY_median_fixation',
    'Std_Position_LX_median_fixation',

```

```

        'Accumulated_Distance_RY_mean_fixation',
        'Distance_Span_mean_saccade_R',
        'Max_Range_LX_median_saccade_R',
        'Accumulated_Distance_RX_mean_saccade_R',
        'Max_Range_LX_mean_saccade_R',
        'Accumulated_Distance_LX_median_saccade_R'
    ]

# Выделение признаков и целевой переменной
X = data[selected_features]
y = data['status']

# Масштабирование данных с сохранением DataFrame
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_scaled_df = pd.DataFrame(X_scaled, columns=X.columns)

# трейн и тест
X_train, X_test, y_train, y_test = train_test_split(X_scaled_df,
y, test_size=0.3, random_state=42)

# Параметры для GridSearch
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag',
'saga'],
    'penalty': ['l1', 'l2', 'none']
}

# GridSearch
grid_search =
GridSearchCV(estimator=LogisticRegression(max_iter=10000,
random_state=42),

                param_grid=param_grid,
                cv=5,
                scoring='accuracy',
                verbose=1,
                n_jobs=-1)
grid_search.fit(X_train, y_train)

# Лучшие параметры, лучшая оценка и лучшая модель
best_params = grid_search.best_params_
best_score = grid_search.best_score_
best_model = grid_search.best_estimator_

# Предикт

```

```
y_pred_best = best_model.predict(X_test)

# Оценка модели
accuracy_best = accuracy_score(y_test, y_pred_best)
conf_matrix_best = confusion_matrix(y_test, y_pred_best)
class_report_best = classification_report(y_test, y_pred_best)

print("Best Model Parameters:", best_params)
print("Best Model Score:", best_score)
print("Accuracy of the best model:", accuracy_best)
print("Confusion Matrix:\n", conf_matrix_best)
print("Classification Report:\n", class_report_best)
```