

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет

Кафедра

вычислительной техники

«Вычислительная техника»

Направление подготовки: 09.04.01 «Информатика и вычислительная техника»

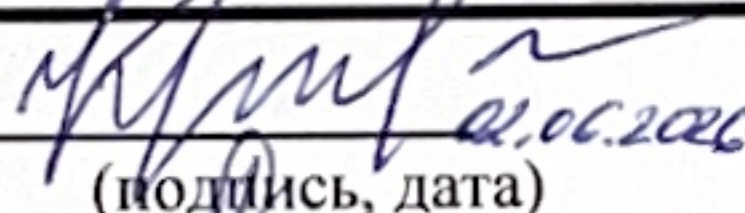
Магистерская программа: «Прикладной искусственный интеллект»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему

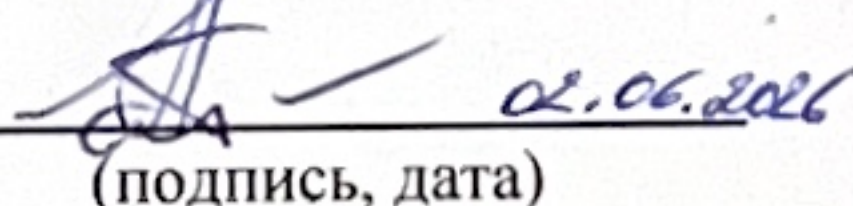
**РАЗРАБОТКА АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ
И МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА И
КЛАССИФИКАЦИИ МУЗЫКАЛЬНЫХ СТИЛЕЙ
И ЖАНРОВ**

Студент


(подпись, дата)

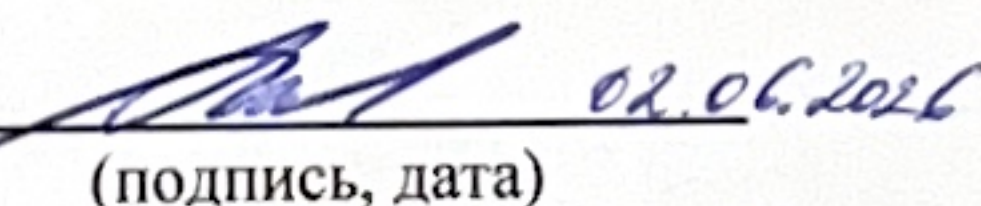
Крюков Кирилл Валерьевич
(ФИО полностью)

Руководитель


(подпись, дата)

Бикташев Р.А.
(фамилия, инициалы)

Нормоконтролёр


(подпись, дата)

Бычков А.С.
(фамилия, инициалы)

Рецензент

Зав. каф. «Программирование» ПензГУ
(должность, место работы)


(подпись, дата)

Мартьянов А.И.
(фамилия, инициалы)

Работа допущена к защите (протокол зас. кафедры от 22.05.2026 № 14)

Заведующий кафедрой


(подпись)

М.А. Митрохин
(фамилия, инициалы)

Работа защищена с отметкой Отлично (протокол заседания ГЭК от

09 июля 2026 № 20)

Секретарь ГЭК



(подпись)

Е.И. Гурин

(фамилия, инициалы)

Пенза 2026

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет

вычислительной техники

Кафедра

«Вычислительная техника»

«Утверждаю»

Заведующий кафедрой

М.А. Митрохин

« 20 » 09 2024 г.

ЗАДАНИЕ
НА МАГИСТЕРСКУЮ ДИССЕРТАЦИЮ

1. Студент Крюков К.В. гр. 24ВВИм2 факультета ВТ

Крюков Кирилл Валерьевич

2. Тема работы « Разработка алгоритмов кластеризации и машинного обучения для анализа и классификации музыкальных стилей и жанров »

Тема утверждена приказом ПГУ №920/о от "20" 09 2024г.

3. Руководитель магистерской диссертации Бикташев Р.А.

4. Задание на работу (назначение разработки, исходные данные и т.п):

Разработка прототипа системы автоматической классификации и кластеризации музыкальных жанров на основе глубоких нейронных сетей. Разрабатываемый программный прототип предназначен для анализа музыкальных аудиозаписей — определения жанра трека и выявления его сходства с другими композициями. Система должна выполнять предварительную обработку аудио и извлечение признаков, классифицировать музыкальный жанр с помощью свёрточной нейронной сети, формировать векторное представление (эмбединг) трека и на его основе группировать похожие записи, а также наглядно представлять результаты анализа.

В результате выполнения работы должен быть разработан MVP-прототип системы анализа музыкальных жанров. Прототип должен демонстрировать работу основных компонентов: модуля предварительной обработки аудио и извлечения признаков (трёхканальной мел-спектрограммы на основе гармонико-перкуссионного разделения сигнала), свёрточной нейронной сети-классификатора, модуля кластеризации треков в пространстве эмбедингов,

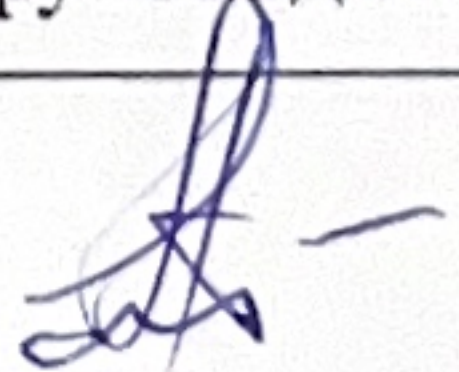
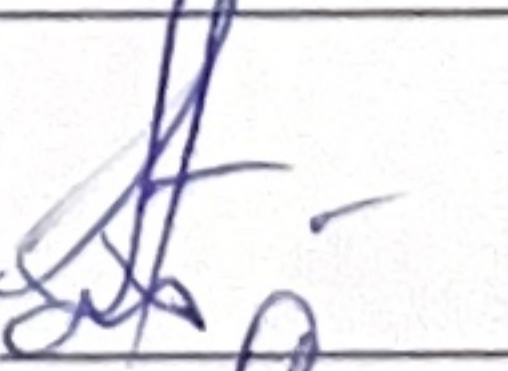
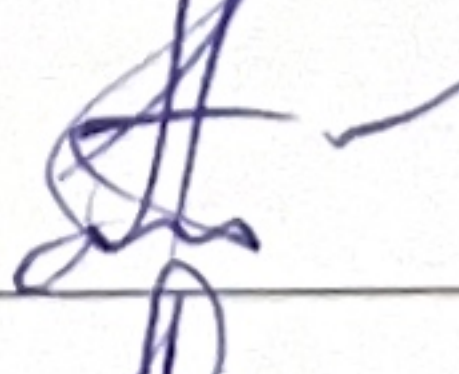
серверной части и пользовательского веб-интерфейса с интерактивной визуализацией графа жанров и кластеров.

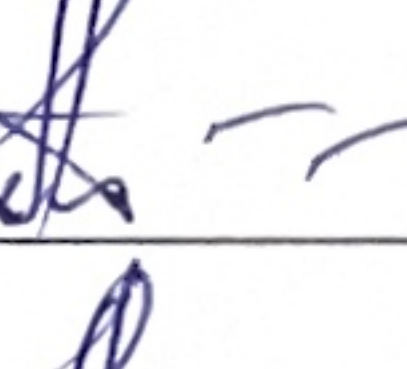
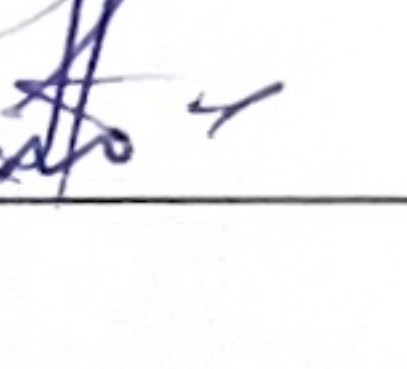
Исходные данные для выполнения работы:

- тема выпускной квалификационной работы: «Разработка алгоритмов кластеризации и машинного обучения для анализа и классификации музыкальных стилей и жанров»;
- научная и учебно-методическая литература по машинному и глубокому обучению, цифровой обработке аудиосигналов, музыкальной информатике, задачам классификации и кластеризации;
- требования к разработке MVP-прототипа системы анализа музыкальных жанров;
- язык программирования Python;
- библиотеки машинного обучения и обработки звука: librosa, TensorFlow/Keras, scikit-learn;
- серверная библиотека Flask и веб-технологии HTML, CSS, JavaScript;
- набор данных GTZAN Genre Collection (1000 аудиозаписей, 10 музыкальных жанров);
- вычислительное оборудование с графическим ускорителем.
-

5. Перечень подлежащих разработке вопросов: В ходе выполнения работы необходимо рассмотреть и разработать следующие вопросы: анализ существующих подходов к классификации и кластеризации музыкальных жанров, используемых наборов данных и метрик оценки; обоснование выбора методов извлечения признаков из аудиосигналов (мел-спектрограмма, гармонико-перкуссионное разделение сигнала); подготовка и предварительная обработка набора данных, включая сегментацию, нормализацию и корректное разбиение на уровне треков, исключающее утечку данных; проектирование и обучение свёрточной нейронной сети-классификатора музыкальных жанров; разработка механизма кластеризации треков по сходству на основе эмбедингов обученной сети; реализация серверной части системы; создание пользовательского веб-интерфейса с интерактивной визуализацией графа жанров и кластеров; интеграция компонентов в единый программный прототип; проведение экспериментальной проверки работоспособности системы и анализ полученных результатов по метрикам классификации и кластеризации.

6. Календарный график выполнения работы

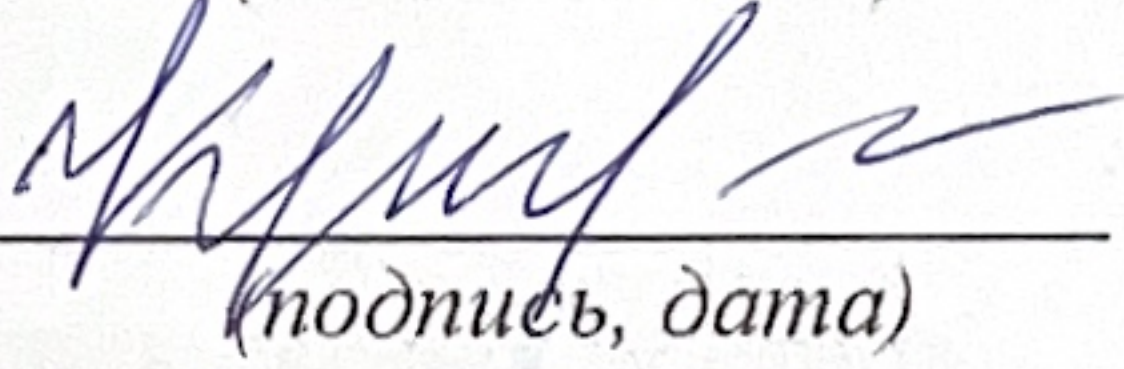
№ п/п	Наименование этапов работы	Объем работы	Срок выполнения	Подпись руководителя
1	Выбор темы, постановка цели и задач	7%	15.11.2024-30.11.2024	
2	Поиск теоретической информации по теме диссертации, сравнение существующих решений	13%	12.01.2025-10.02.2025	
3	Разработка архитектуры проекта	18%	13.03.2025-28.04.2025	
4	Создание модульной системы платформы	18%	15.05.2025-30.06.2025	
5	Реализация каждого модуля с проверкой работоспособности	28%	19.10.2025-12.01.2026	

6	Создание веб-интерфейса	3%	02.02.2026- 10.02.2026	
7	Внедрение ИИ в систему	2%	15.02.2026- 19.02.2026	
8	Экспериментальная проверка работоспособности системы	6%	15.03.2026- 25.03.2026	
9	Проверка на соответствие целям работы	3%	14.04.2026- 16.04.2026	
10	Подведение итогов	2%	05.05.2026- 20.05.2026	

Дата выдачи задания " 20 " сентября 20 24 г.

Руководитель магистерской диссертации 

(подпись, дата)

Задание к исполнению принял студент 

(подпись, дата)

Работу к защите допустить
Декан факультета

 А. Р. Фролов

09.06.2020 (подпись, дата)

РЕФЕРАТ

Пояснительная записка содержит 75 листов, 5 рисунков, 12 таблиц, 19 источников, 1 приложение.

РАЗРАБОТКА АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ И МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА И КЛАССИФИКАЦИИ МУЗЫКАЛЬНЫХ СТИЛЕЙ И ЖАНРОВ

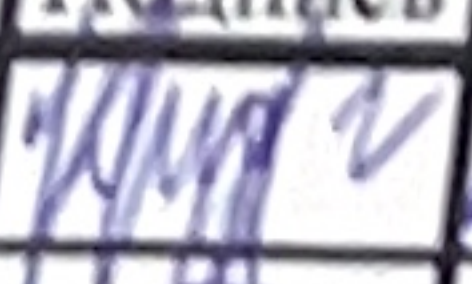
Машинное обучение, глубокое обучение, свёрточная нейронная сеть, классификация музыкальных жанров, кластеризация, мел-спектрограмма, гармонико-перкуссионное, эмбединги, цифровая обработка аудиосигналов.

Целью работы является разработка системы автоматической классификации и кластеризации музыкальных жанров на основе свёрточных нейронных сетей и признаков аудиосигналов.

Разрабатываемое приложение предназначено для анализа музыкальных композиций – определения их жанровой принадлежности и выявления сходства между треками – и может применяться для улучшения пользовательского опыта в музыкальных стриминговых и рекомендательных сервисах.

В результате проведённой работы реализован программный прототип, включающий механизм обучения свёрточной нейронной сети для классификации музыкальных жанров на основе трёхканальной мел-спектрограммы с гармонико-перкуссионным разделением сигнала, модуль кластеризации треков по сходству в пространстве эмбедингов, серверную часть и веб-интерфейс с интерактивной визуализацией графа жанров и кластеров.

Особое внимание уделено исследованию подходов к извлечению признаков из аудиосигналов и обучению нейросетевых моделей, корректной методике оценки качества с разделением данных на уровне треков, исключающим утечку, а также объединению задач классификации и кластеризации в рамках единой системы.

					ПГУ 1.090401.09.001 ПЗ			
Изм.	Лист	№ докум	Подпись	Дата				
Разраб.		Крюков К.В.		2.05.24	Разработка алгоритмов кластеризации и машинного обучения для анализа и классификации музыкальных стилей и жанров Пояснительная записка	Лит.	Лист	Листов
Провер.		Бикташев Р.А.		2.05.24			2	117
Н. Контр.		Бычков А.С.		2.05.24		ФВТ, каф. ВТ гр. 24ВВИм2		
Утв.		Митрохин М.А.		2.05.24				

ВВЕДЕНИЕ

Музыка сопровождает человека почти непрерывно, а развитие онлайн-сервисов открыло доступ к коллекциям, объёмы которых ещё недавно казались немыслимыми. У такого изобилия есть и обратная сторона: ориентироваться в нём становится всё труднее. Подбор подходящих треков, составление тематических подборок, отслеживание музыкальных тенденций – всё это уже не получается делать вручную и требует автоматизации. Одно из центральных направлений здесь – распознавание жанра прямо по звучанию записи, которое сегодня опирается на методы машинного обучения.

Практическая потребность в таких методах складывается из нескольких соображений:

- рекомендательные сервисы опираются на жанровую информацию, подбирая музыку под вкусы конкретного слушателя,
- без автоматической разметки трудно упорядочить и быстро искать записи в крупных фонотеках,
- для исследователей подобные методы становятся способом находить неочевидные закономерности в больших массивах музыкальных данных.

Кроме того, на практике важно не только приписать треку готовую жанровую метку, но и понять, на какие ещё композиции он похож, – а это уже область кластеризации, что расширяет возможности анализа коллекций.

Работ по этой теме накоплено немало, однако у большинства из них есть общие слабые места. Классификацию и кластеризацию обычно рассматривают порознь; внутренние представления, формирующиеся внутри нейронной сети, редко задействуют для поиска похожих треков; а способ деления данных на обучающую и проверочную части нередко допускает утечку, из-за чего заявленная точность оказывается завышенной. Эти пробелы и подтолкнули к разработке решения, в котором обе задачи объединены.

Объект исследования – автоматический анализ музыкальных аудиозаписей средствами машинного обучения.

Предмет исследования – модели и алгоритмы классификации и кластеризации музыкальных жанров, построенные на свёрточных нейронных сетях и признаках аудиосигнала.

Цель работы – построить и проверить гибридный подход, который объединяет классификацию и кластеризацию музыкальных жанров на основе глубоких нейронных сетей и одновременно обеспечивает точное отнесение к известным жанрам и возможность находить сходство между композициями.

Для её достижения решались следующие задачи:

- разобрать существующие методы классификации и кластеризации жанров, оценив их сильные и слабые стороны,
- изучить способы извлечения признаков из звукового сигнала и выбрать подходящее представление,
- построить гибридную модель, в которой свёрточная сеть отвечает за классификацию, а кластеризация выполняется в пространстве её эмбедингов,
- реализовать систему программно и экспериментально оценить качество классификации и кластеризации.

Гипотеза работы состоит в том, что сочетание глубоких сетей с многоуровневым описанием сигнала и объединение классификации с кластеризацией в одной системе повышают и точность, и содержательность анализа жанров.

Методы исследования. В работе использованы анализ литературы по теме, приёмы цифровой обработки сигналов, аппарат машинного и глубокого обучения, эксперимент на общепринятом наборе данных и статистическая оценка результатов.

Научная новизна работы заключается в том, что:

- сигнал описывается тремя каналами на основе гармонико-перкуSSIONного разделения,
- для группировки треков по сходству задействованы эмбединги уже обученной сети,

— классификация и кластеризация объединены в одной модели при делении данных по трекам, исключая утечку.

Практическая значимость. Разработанный прототип системы анализа жанров применим в музыкальных рекомендательных и поисковых сервисах, при автоматическом составлении плейлистов и при разборе больших музыкальных коллекций.

1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ КЛАССИФИКАЦИИ И КЛАСТЕРИЗАЦИИ МУЗЫКАЛЬНЫХ ЖАНРОВ

1.1 Понятие музыкального жанра и стиля

Прежде чем переходить к анализу конкретных алгоритмов и проектированию системы, необходимо рассмотреть теоретические основы предметной области. В настоящей главе вводятся ключевые понятия музыкального жанра и стиля, анализируются исторически сложившиеся подходы к ручной классификации музыки, излагаются фундаментальные принципы применения машинного обучения в музыкальной информатике, разграничиваются задачи классификации и кластеризации, а также рассматриваются методы извлечения признаков из аудиосигналов. Эти положения формируют понятийный аппарат, на который опираются последующие главы.

Понимание теоретических основ имеет не только формальное значение. Корректная постановка задачи автоматической классификации напрямую зависит от ясного представления о том, что именно подлежит классификации, насколько устойчивы и однозначны соответствующие категории и какие характеристики сигнала несут наиболее существенную информацию. Поэтому изложение в данной главе движется от музыковедческих понятий к их формализации в терминах машинного обучения.

В области музыкознания и при разработке алгоритмов классификации музыки ключевыми являются понятия «музыкальный жанр» и «музыкальный стиль». Несмотря на то что в обыденной речи они нередко употребляются как синонимы, эти термины обозначают различные аспекты музыкального произведения и требуют чёткого разграничения, особенно в контексте машинного обучения, где от точности определений зависит постановка задачи.

Музыкальный жанр представляет собой категорию, определяемую на основе общих характеристик музыкальных произведений. К таким характеристикам относятся форма (структурная организация произведения – соната, симфония, песня), инструментарий (состав используемых инструментов

– оркестр, рок–группа, джазовый ансамбль), гармония (характерные последовательности аккордов), ритм (типичные ритмические рисунки и темп), текстовое содержание для вокальной музыки, а также социокультурный контекст возникновения и бытования. Жанры формируются исторически и эволюционируют со временем, отражая изменения в музыкальной культуре. Примерами могут служить классическая музыка, джаз, рок, поп, электронная музыка и фольклор.

Музыкальный стиль, напротив, характеризует способ выражения музыкальных идей в рамках определённого жанра. Он определяется совокупностью выразительных средств – мелодикой (характерными мелодическими оборотами), тембром (звуковой окраской), артикуляцией (способами исполнения звуков), динамикой (изменением громкости) и фактурой (характером музыкальной ткани). Стиль отражает индивидуальную манеру композитора, исполнителя, эпохи или направления, и внутри одного жанра может существовать множество стилей. Так, в рамках рока выделяют хард–рок, панк–рок, гранж и инди–рок, а в классической музыке – стили барокко, классицизма и романтизма.

Ключевое различие между жанром и стилем заключается в уровне обобщения: жанр является более широкой категорией, стиль – более узкой. Образно говоря, жанр отвечает на вопрос «что это за музыка?», а стиль – «как она сделана?». При этом оба понятия тесно взаимосвязаны: стиль развивается в рамках жанра, а накопление стилистических изменений может приводить к возникновению новых жанров. Жанровая система, таким образом, не является застывшей – она непрерывно меняется, что само по себе осложняет любую попытку её строгой формализации.

Для задач автоматической классификации это разграничение имеет практическое значение, поскольку выбор признаков должен соответствовать цели. Классификация по жанрам обычно опирается на более общие характеристики – инструментарий, темп, гармонические особенности, – которые отражаются в спектральных и ритмических параметрах сигнала. Классификация

по стилям требует более тонкого анализа мелодических, тембральных и исполнительских нюансов и, как правило, более сложных моделей. Настоящая работа сосредоточена на жанровой классификации как на более общей и формализуемой задаче, однако предлагаемые методы в перспективе применимы и к стилевому анализу.

Следует также учитывать, что и жанр, и стиль в значительной мере являются культурными конструктами. Восприятие принадлежности произведения к тому или иному жанру может зависеть от культурного и исторического контекста слушателя, а границы категорий нередко определяются не столько объективными акустическими признаками, сколько сложившейся традицией и маркетинговыми практиками музыкальной индустрии. Эта изначальная нечёткость переносится и на размеченные наборы данных, на которых обучаются модели, задавая объективный предел достижимой однозначности классификации.

Жанровая система имеет иерархическую организацию. Крупные жанры распадаются на поджанры, те – на ещё более узкие категории, образуя многоуровневую таксономию. Например, электронная музыка включает множество направлений – от хауса и техно до эмбиента, – каждое из которых, в свою очередь, дробится далее. Подобная иерархичность означает, что задача классификации может ставиться на разных уровнях обобщения, и выбор уровня существенно влияет как на сложность задачи, так и на достижимую точность: различение крупных жанров обычно проще, чем тонкое разграничение близких поджанров.

Нечёткость границ между категориями представляет принципиальную трудность для автоматической классификации. Многие произведения сочетают признаки нескольких жанров: композиция может одновременно нести черты рока и электроники, джаза и блюза, попа и танцевальной музыки. Такие пограничные объекты не имеют единственно верной метки, и любое отнесение их к одному классу является упрощением. Для алгоритма это проявляется в систематических ошибках между близкими категориями, которые отражают не

столько несовершенство модели, сколько объективную природу музыкального материала. Осознание этого факта важно для корректной интерпретации результатов и обуславливает интерес к мягким и многометочным схемам классификации, а также к кластеризации, не требующей жёсткой разметки.

Практическая значимость жанровых категорий усиливается их ролью в музыкальной индустрии. Жанр служит основным способом навигации в каталогах, организации витрин стриминговых сервисов и таргетирования рекомендаций, а также выступает маркетинговым инструментом продвижения исполнителей. В результате жанровая метка нередко формируется не только музыкальными, но и коммерческими соображениями, что дополнительно размывает её акустическую определённость. Тем не менее именно эта роль делает автоматическую жанровую классификацию востребованной: она лежит в основе систем поиска и рекомендаций, обрабатывающих коллекции, которые невозможно разметить вручную.

Жанр является лишь одним из измерений, по которым может описываться музыкальное произведение. Наряду с ним выделяют настроение (спокойная, агрессивная, меланхоличная музыка), эпоху, темп, инструментальный состав и функциональное назначение (фоновая, танцевальная, концертная музыка). Эти измерения частично коррелируют между собой, но не сводятся друг к другу: произведения одного жанра могут различаться по настроению, а схожие по настроению – принадлежать разным жанрам. Многоаспектность музыкального описания означает, что жёсткое отнесение трека к единственной жанровой категории неизбежно огрубляет картину, и это служит дополнительным аргументом в пользу подходов, выявляющих сходство треков по совокупности характеристик, а не только по заданной метке.

1.2 Ручная классификация музыки: история и ограничения

Прежде чем перейти к автоматическим методам, целесообразно рассмотреть исторически сложившиеся подходы к ручной классификации музыки. Это позволяет понять предпосылки возникновения автоматизированных

методов и оценить их преимущества, а также роль, которую экспертная классификация продолжает играть в машинном обучении.

Классификация музыки, основанная на слуховом восприятии и экспертных знаниях, существует на протяжении столетий. Музыковеды и критики разрабатывали системы категоризации по различным критериям: по эпохе (Средневековье, Ренессанс, Барокко, Классицизм, Романтизм), по жанру (опера, симфония, соната, песня, танец), по стилю (индивидуальные особенности композиторов и направлений) и по культурной или географической принадлежности (национальные и региональные традиции). Подобные системы требовали глубоких знаний музыкальной теории, истории и культуры.

Характерным примером формализованной ручной классификации служат тематические каталоги произведений отдельных композиторов – каталог Кёхеля для сочинений Моцарта или каталог BWV для произведений Баха, – а также библиотечные и архивные системы упорядочивания нотных и звуковых материалов. Сама ручная классификация осуществлялась несколькими способами: анализом партитур (изучением нотного текста для выявления структурных, гармонических и мелодических особенностей), слуховым анализом (определением характеристик на основе прослушивания) и экспертными оценками квалифицированных музыковедов.

Несмотря на свою ценность, ручная классификация обладает рядом существенных ограничений, которые делают её малопригодной для обработки больших объёмов данных. К ним относятся:

- субъективность (оценки разных экспертов могут расходиться),
- трудоёмкость (анализ требует значительных временных затрат и высокой квалификации),
- ограниченная масштабируемость (ручные методы неэффективны для постоянно растущих цифровых коллекций),
- нечёткость границ между жанрами и стилями, а также зависимость от культурного и исторического контекста.

С наступлением цифровой эпохи и появлением музыкальных библиотек, насчитывающих миллионы композиций, эти ограничения стали критическими. Вместе с тем ручная классификация не утратила значения и в эпоху машинного обучения, а скорее изменила свою роль. Во-первых, результаты экспертной разметки используются для формирования обучающих выборок, без которых невозможно обучение моделей с учителем. Во-вторых, экспертная классификация служит эталоном при оценке качества автоматических методов: предсказания модели сравниваются именно с разметкой, выполненной человеком. В-третьих, накопленные экспертные знания помогают в осмысленном конструировании признаков и интерпретации результатов. Таким образом, автоматическая и ручная классификация не противопоставляются, а дополняют друг друга.

С наступлением цифровой эпохи ручная классификация частично трансформировалась в ручное тегирование: метаданные о жанре стали сопровождать аудиофайлы (например, в тегах формата ID3), а на музыкальных платформах разметка нередко выполняется самими исполнителями, редакторами сервисов или сообществом пользователей в режиме краудсорсинга. Это позволило накопить большие объёмы размеченных данных, однако не устранило, а в ряде случаев и усилило проблему несогласованности: метки, проставленные разными людьми по разным соображениям, могут противоречить друг другу.

Степень согласия между экспертами (межаннотаторская согласованность) служит косвенной мерой объективности классификации. Исследования показывают, что даже квалифицированные слушатели далеко не всегда сходятся в определении жанра, особенно для пограничных композиций. Это имеет важное следствие для машинного обучения: точность модели принципиально ограничена сверху качеством и согласованностью разметки, на которой она обучается. Иными словами, модель не может быть «точнее», чем сами эталонные метки, и часть наблюдаемых ошибок объясняется именно противоречивостью исходных данных, а не недостатками алгоритма.

Именно несоответствие между трудоёмкостью ручной классификации и стремительным ростом объёмов цифровой музыки стало главной движущей силой развития автоматических методов. Современные музыкальные сервисы оперируют каталогами в десятки миллионов треков, ежедневно пополняемыми тысячами новых композиций. Ручная разметка таких объёмов экономически и организационно неосуществима, а её субъективность и непоследовательность делают результат недостаточно надёжным. Автоматическая классификация, напротив, обеспечивает масштабируемость, скорость и воспроизводимость, что и определяет её центральную роль в современной музыкальной информатике.

Стоит подчеркнуть, что в современной практике ручная и автоматическая классификация всё чаще сочетаются. Крупные размеченные наборы данных формируются гибридно: первичная разметка выполняется экспертами или сообществом, после чего она используется для обучения моделей, а предсказания моделей, в свою очередь, помогают выявлять и исправлять ошибки в разметке. Подобный итеративный процесс повышает качество данных и снижает трудозатраты. Таким образом, экспертная классификация не вытесняется полностью, а встраивается в общий цикл создания и совершенствования автоматических систем, оставаясь источником эталонного знания.

1.3 Машинное обучение в музыкальной информатике

Развитие технологий цифровой обработки звука и накопление огромных объёмов музыкальных данных создали предпосылки для применения методов машинного обучения в музыкальной информатике. Машинное обучение предоставляет инструменты для решения разнообразных задач анализа, классификации, генерации и обработки музыки, и в данном разделе рассматриваются фундаментальные принципы его применения.

Машинное обучение как подраздел искусственного интеллекта основано на разработке алгоритмов, способных обучаться на данных без явного программирования правил. Вместо задания жёстких инструкций алгоритму предоставляется возможность самостоятельно выявлять закономерности и

применять их к новым, ранее не встречавшимся примерам. Ключевым свойством здесь является обобщающая способность – умение переносить выученные закономерности на новые данные, а не запоминать обучающие примеры. В музыкальной информатике это означает, что, обучив модель на размеченном наборе произведений, её можно применять для анализа новых треков.

Принято выделять несколько парадигм машинного обучения. При обучении с учителем (supervised learning) алгоритму предоставляются данные с известными правильными ответами, и он учится сопоставлять входные признаки этим ответам; к этой парадигме относится классификация музыки по жанрам. При обучении без учителя (unsupervised learning) данные не размечены, и алгоритм самостоятельно выявляет в них скрытую структуру; типичным примером является кластеризация треков по сходству. При обучении с подкреплением (reinforcement learning) агент учится, взаимодействуя со средой и получая вознаграждение за свои действия; в музыке эта парадигма применяется, например, для генерации и интерактивного аккомпанемента.

Важным аспектом применения машинного обучения является представление музыкальных данных. Аудиоданные могут быть представлены в виде временного ряда амплитуд, спектрограмм или производных признаков; нотные данные – в форматах MIDI или MusicXML; кроме того, доступны метаданные об исполнителе, альбоме и годе выпуска. Выбор представления существенно влияет на эффективность алгоритмов, поскольку определяет, какая информация о сигнале будет доступна модели.

Для корректной оценки качества данные принято разделять на обучающую, валидационную и тестовую выборки: первая используется для настройки параметров модели, вторая – для подбора гиперпараметров и контроля переобучения, третья – для итоговой оценки на не встречавшихся данных. Центральной проблемой обучения является переобучение, при котором модель хорошо воспроизводит обучающие примеры, но плохо обобщается на новые; для борьбы с ним применяются регуляризация, аугментация данных и

ранняя остановка. Эти понятия, лишь намеченные здесь, подробно используются в последующих главах.

Спектр задач, решаемых машинным обучением в музыке, широк: классификация по жанрам, стилям и настроению; рекомендательные системы; генерация новых произведений; музыкальный анализ (выделение тем, мотивов, гармонических последовательностей); обработка звука (шумоподавление, транскрипция). В настоящей работе основное внимание уделяется обучению с учителем и без учителя, а также вопросам представления данных и выбора признаков, что определяет содержание последующих разделов.

Особое место в современной музыкальной информатике занимает глубокое обучение – направление, основанное на многослойных нейронных сетях[1]. Его принципиальное отличие состоит в способности автоматически формировать иерархию признаков непосредственно из исходных данных, без предварительного ручного конструирования. На нижних слоях сеть выделяет простые локальные закономерности, на верхних – всё более абстрактные и специфичные для задачи представления. Применительно к музыке это означает, что сеть, обучаясь различать жанры, попутно формирует внутреннее представление звучания, которое может использоваться и для других задач, включая поиск похожих треков и кластеризацию. Именно эта идея обучения представлений лежит в основе предлагаемого в работе подхода.

Вместе с тем применение машинного обучения в музыке сопряжено с рядом трудностей. Глубокие модели требуют значительных объёмов данных и вычислительных ресурсов, чувствительны к качеству и сбалансированности обучающей выборки и могут перенимать содержащиеся в ней систематические смещения. Кроме того, они обладают ограниченной интерпретируемостью: понять, почему модель отнесла произведение к тому или иному жанру, зачастую затруднительно. Эти ограничения необходимо учитывать при проектировании системы и интерпретации результатов, и они подробно обсуждаются в последующих главах.

Качество работы любой модели машинного обучения определяется не только выбранным алгоритмом, но и тем, как представлены и подготовлены данные. Одни и те же аудиозаписи, описанные разными наборами признаков, могут давать существенно различающиеся результаты. Поэтому представление данных рассматривается как самостоятельный и зачастую решающий этап: удачно подобранное представление способно компенсировать простоту модели, тогда как неинформативные признаки не позволят раскрыть потенциал даже самого сложного алгоритма. Этот тезис проходит красной нитью через всю работу и обосновывает внимание, уделяемое извлечению признаков.

Не менее важны полнота, сбалансированность и репрезентативность обучающих данных. Если в выборке преобладают записи определённого характера, модель усвоит соответствующие смещения и будет хуже работать на недопредставленных категориях. Несбалансированность классов, шум в разметке и систематические искажения записи способны существенно снизить качество и обобщающую способность. Поэтому подготовка данных – очистка, балансировка, корректное разбиение на выборки – является обязательным условием получения достоверных результатов, а пренебрежение ею ведёт к завышенным или нерепрезентативным оценкам.

Типовой процесс решения задачи машинного обучения в музыке можно представить как последовательность этапов: сбор и подготовка данных, извлечение или обучение признакового представления, обучение модели, оценка её качества на отложенных данных и, при необходимости, развёртывание в составе прикладной системы. Каждый из этапов влияет на итоговый результат, и ошибки на ранних стадиях – некорректная разметка или неудачное представление данных – невозможно полностью компенсировать на последующих. Такой системный взгляд на процесс задаёт структуру практической части настоящей работы, в которой последовательно прорабатываются подготовка данных, признаки, модель и её оценка.

1.4 Классификация и кластеризация в машинном обучении

Задачи анализа данных в машинном обучении часто сводятся к двум основным типам – классификации и кластеризации. Обе направлены на выявление закономерностей, но используют принципиально разные подходы, и понимание различий между ними необходимо для обоснованного выбора методов анализа музыкальных жанров.

Классификация относится к задачам обучения с учителем. Для обучения используется размеченный набор данных, в котором для каждого объекта известен правильный класс, а цель алгоритма – построить модель, сопоставляющую входные признаки предопределённым классам. В контексте музыки входными данными служат признаки, извлечённые из аудиозаписей, а классами – жанры или стили. Процесс включает сбор и разметку данных, извлечение признаков, обучение модели и её оценку на отложенной выборке.

Кластеризация относится к задачам обучения без учителя. Алгоритму предоставляется неразмеченный набор данных, и он должен самостоятельно выявить группы схожих объектов – кластеры; заранее известные классы при этом отсутствуют. Цель состоит в таком разбиении, при котором объекты внутри группы более похожи друг на друга, чем на объекты других групп. В музыкальном контексте кластеризация позволяет обнаруживать группы произведений со схожими стилистическими или жанровыми характеристиками без предварительной разметки. Ключевые различия между двумя задачами обобщены в таблице 1.

Таблица 1. – Ключевые различия между классификацией и кластеризацией

Характеристика	Классификация	Кластеризация
Тип обучения	С учителем	Без учителя
Размеченные данные	Требуются (классы известны)	Не требуются (классы неизвестны)
Цель	Предсказание класса нового объекта	Выявление групп схожих объектов
Результат	Модель–классификатор	Разбиение данных на кластеры

Между двумя крайними парадигмами существуют и промежуточные подходы. Полуконтролируемое обучение использует небольшой объём размеченных данных совместно с большим объёмом неразмеченных, что особенно ценно при дефиците разметки. Кроме того, кластеризация может применяться как вспомогательный этап перед классификацией – для предварительного структурирования данных или для выявления естественных групп, которые затем сопоставляются с экспертными жанрами.

Применительно к настоящей работе обе задачи рассматриваются как взаимодополняющие. Классификация обеспечивает отнесение трека к известному жанру и опирается на размеченный набор данных, тогда как кластеризация раскрывает скрытую структуру коллекции и сходство между композициями. Именно их объединение в рамках единой системы составляет одну из центральных идей исследования, детально раскрываемую в последующих главах.

Существенно различается и способ оценки качества для двух задач. В классификации, где известны истинные метки, качество измеряется сопоставлением предсказаний с ними – через долю верных ответов и связанные метрики. В кластеризации истинные метки при оценке формально отсутствуют, поэтому используются либо внутренние меры, характеризующие компактность

и разделённость групп, либо внешние меры согласованности с эталонной разметкой, если таковая доступна. Это различие отражает фундаментальную асимметрию между задачами: классификация воспроизводит заданную человеком категоризацию, тогда как кластеризация предлагает собственную, которую ещё предстоит осмыслить.

Проиллюстрировать разницу можно на простом примере. Если имеется коллекция треков, заранее размеченных по жанрам, и требуется научиться определять жанр новых композиций, – это задача классификации. Если же коллекция не размечена и требуется выделить в ней группы похожих по звучанию треков, не зная заранее ни их числа, ни названий, – это задача кластеризации. В первом случае результат сравнивается с известным «правильным ответом», во втором – оценивается по внутренней согласованности и содержательной осмысленности групп.

Каждая из задач обладает своими сильными и слабыми сторонами применительно к музыке. Классификация даёт понятный и проверяемый результат, согласованный с привычными жанровыми категориями, но ограничена заранее заданным набором классов и полностью зависит от наличия и качества разметки. Кластеризация свободна от необходимости в разметке и способна обнаруживать неожиданные, не предусмотренные экспертами группы, однако её результаты труднее интерпретировать и проверять, а их осмысленность сильно зависит от выбранного признакового пространства. Сочетание двух подходов позволяет частично компенсировать недостатки каждого, что и обуславливает интерес к гибридным решениям.

Отдельно следует отметить роль меры близости, лежащей в основе большинства методов кластеризации и ряда методов классификации. От того, как именно измеряется сходство между объектами – через евклидово расстояние, косинусную меру или иные метрики, – напрямую зависит результат группировки. В высокоразмерных пространствах, характерных для аудиопризнаков, выбор меры близости приобретает особое значение, поскольку при росте размерности расстояния между объектами становятся всё менее

различимыми. Это обстоятельство тесно связано с качеством признакового представления: осмысленная кластеризация возможна лишь тогда, когда выбранная мера близости в данном пространстве действительно отражает воспринимаемое сходство музыкальных произведений.

1.5 Извлечение признаков из аудиосигналов

Для успешного применения методов машинного обучения аудиосигналы необходимо преобразовать в числовые представления, содержащие релевантную информацию о музыкальном контенте. Этот процесс называется извлечением признаков и играет ключевую роль в достижении высокой точности, поскольку даже самый совершенный алгоритм не способен компенсировать неинформативное признаковое представление.

Аудиосигнал представляет собой одномерный временной ряд, отражающий изменение амплитуды звука во времени. Его непосредственное использование затруднено из-за высокой размерности и изменчивости, поэтому применяются методы, извлекающие более компактные и информативные характеристики. Фундаментальным инструментом анализа является преобразование Фурье, разлагающее сигнал на составляющие частоты и дающее представление о его спектре. Однако спектр является глобальной характеристикой и не отражает изменений во времени.

Для анализа временной динамики спектра используется кратковременное преобразование Фурье (STFT), разбивающее сигнал на короткие перекрывающиеся фрагменты и вычисляющее спектр для каждого из них. Результатом является спектрограмма – двумерное представление, где по одной оси отложено время, по другой частота, а яркость соответствует энергии сигнала. Спектрограммы широко используются как для визуализации звука, так и в качестве входных данных для нейронных сетей, поскольку сохраняют двумерную структуру, удобную для свёрточной обработки [2]. Соответствующие математические соотношения подробно приводятся в главе 3.

Поскольку спектрограмма не учитывает особенности человеческого слухового восприятия, для более адекватного представления применяется мел–

шкала, имитирующая нелинейное восприятие частот ухом, и построенные на её основе мел-спектрограммы [2]. Пример мел-спектрограммы изображен на рисунке 1. На их основе вычисляются мел-частотные кепстральные коэффициенты (MFCC) – одни из наиболее эффективных признаков для распознавания речи и музыки, хорошо описывающие тембральные характеристики и устойчивые к шуму [3]. MFCC позволяют существенно снизить размерность представления, сохраняя наиболее значимую информацию о тембре.

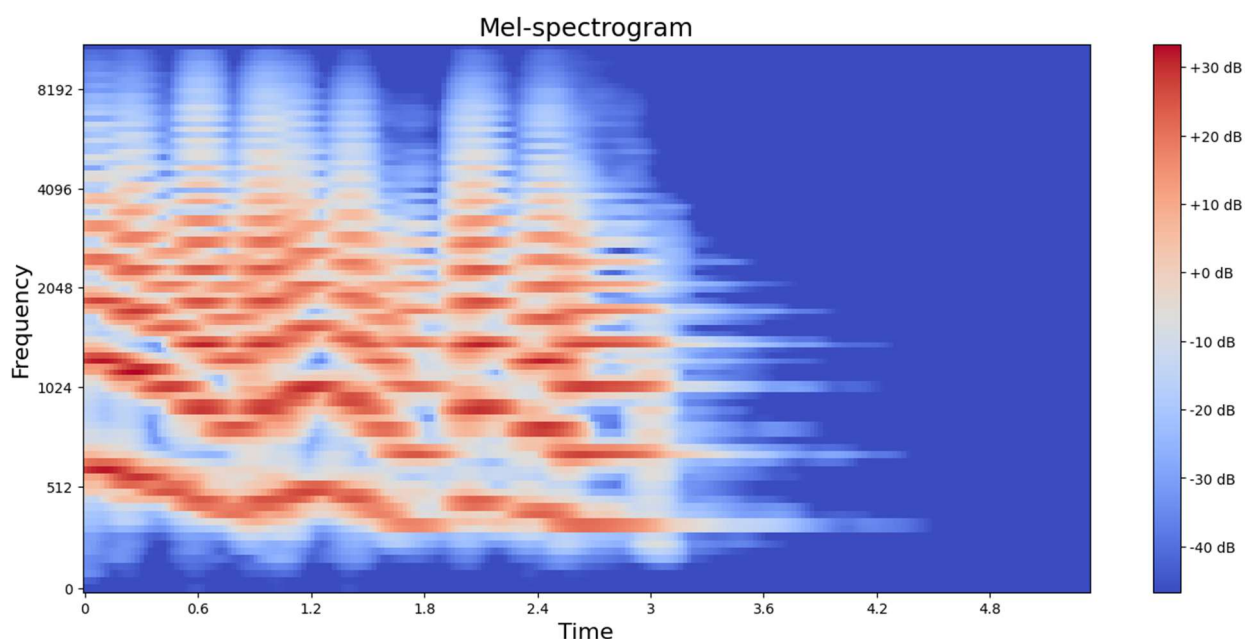


Рисунок 1 – Пример мел-спектрограммы, отображающей нелинейное восприятие частот ухом

Помимо MFCC и спектрограмм, применяется ряд других признаков. Хрома-признаки описывают распределение энергии по двенадцати полутонам музыкальной гаммы и полезны для анализа гармонии и тональности. Темпоральные признаки характеризуют временные свойства сигнала – темп, ритм, амплитудную огибающую. Спектральные дескрипторы (спектральный центроид, ширина полосы, контраст) отражают физические свойства звука, а статистические признаки описывают распределение других характеристик. Дополнительно может применяться гармонико-перкуSSIONное разделение сигнала, выделяющее тональную и ритмическую составляющие, что используется и в настоящей работе.

Сводная характеристика основных групп признаков приведена в таблице 2. При извлечении признаков существенны и технические параметры: размер фрейма и величина перекрытия определяют временное разрешение анализа; оконная функция снижает эффект растекания спектра; нормализация приводит признаки к единому диапазону значений, улучшая работу алгоритмов. Выбор конкретного набора признаков зависит от задачи: для жанровой классификации часто достаточно мел–спектрограмм, MFCC и темпоральных признаков, тогда как для тонкого стилизового анализа требуются более детальные характеристики.

Таблица 2 – Основные группы признаков аудиосигнала

Группа признаков	Что описывает	Типичное применение
Спектрограмма / мел–спектрограмма	Распределение энергии по частоте и времени	Вход для свёрточных нейронных сетей
MFCC	Тембральные характеристики (восприятие)	Классификация музыки и речи
Хрома–признаки	Распределение по полутонам гаммы	Анализ гармонии и тональности
Темпоральные	Темп, ритм, огибающая	Различение ритмически выраженных жанров
Спектральные дескрипторы	Физические свойства спектра	Тонкий анализ тембра и стиля

Процесс извлечения признаков можно рассматривать как конвейер последовательных преобразований сигнала: от исходной звуковой волны к частотно–временному представлению, затем к перцептивно мотивированным характеристикам и, наконец, к компактным дескрипторам, пригодным для подачи в модель. На каждом этапе происходит сжатие информации с сохранением наиболее существенных для задачи особенностей. Качество этого конвейера во многом определяет итоговый результат, поскольку утраченная на ранних этапах информация не может быть восстановлена моделью.

Отдельного внимания заслуживает вопрос временной агрегации. Поскольку музыкальное произведение разворачивается во времени, признаки, вычисленные для коротких фрагментов, необходимо тем или иным образом объединять. Применяются как простое усреднение и вычисление статистик по всей записи, так и более сложные подходы, сохраняющие временную структуру, – например, обработка последовательности фрагментов рекуррентными сетями. Выбор способа агрегации влияет на то, насколько полно модель учитывает развитие музыкального материала во времени, и является самостоятельным проектным решением.

Важную роль играет и нормализация признаков, приводящая их к сопоставимому масштабу. Без неё признаки с большим разбросом значений могут доминировать в вычислениях расстояний и градиентов, искажая результат. В сочетании с продуманным выбором размера анализируемого окна и его перекрытия нормализация обеспечивает устойчивость и воспроизводимость анализа. Эти технические аспекты, на первый взгляд второстепенные, на практике существенно влияют на качество как классификации, так и кластеризации.

При большом числе извлекаемых признаков возникает задача их отбора и снижения размерности. Избыточные и слабоинформативные признаки увеличивают вычислительную нагрузку и могут ухудшать качество из-за переобучения и эффекта «проклятия размерности». Для борьбы с этим применяются методы отбора наиболее значимых признаков и методы снижения размерности, такие как метод главных компонент, выделяющий направления наибольшей изменчивости данных. Грамотный отбор признаков позволяет одновременно повысить точность и снизить вычислительные затраты, что особенно важно при работе с большими коллекциями.

Сопоставление двух стратегий – ручного конструирования признаков и их автоматического обучения – показывает, что каждая имеет свою область применимости. Ручные признаки прозрачны, интерпретируемы и не требуют больших объёмов данных, но ограничены изобретательностью исследователя и

могут упускать существенные, но неочевидные закономерности. Обучаемые представления способны извлекать сложные зависимости непосредственно из данных, но требуют значительных выборок и вычислительных ресурсов и хуже поддаются интерпретации. На практике эти подходы нередко сочетаются: исходным представлением служит спектрограмма, а признаки формируются сетью, при этом понимание классических дескрипторов помогает осмыслить и проверить работу модели.

Существенным требованием к признаковому представлению является инвариантность к несущественным изменениям сигнала. Желательно, чтобы признаки слабо реагировали на вариации, не влияющие на жанровую принадлежность, – изменение общей громкости, незначительные шумовые искажения, различия в качестве записи или формате сжатия, – но при этом сохраняли чувствительность к содержательным характеристикам звучания. Достижение такого баланса является нетривиальной задачей: чрезмерная инвариантность ведёт к потере полезной информации, а её недостаток – к неустойчивости модели на реальных данных. Перцептивно мотивированные представления, такие как мел–спектрограмма, частично решают эту проблему, согласуя описание сигнала с особенностями человеческого слуха.

Современной тенденцией является переход от ручного конструирования признаков к их автоматическому обучению нейронными сетями. В этом случае исходным представлением служит спектрограмма, а необходимые признаки формируются внутренними слоями сети в процессе обучения. Тем не менее понимание классических признаков сохраняет значение: они задают концептуальную основу, помогают интерпретировать работу моделей и нередко используются совместно с обучаемыми представлениями. Рассмотренные в данной главе теоретические положения – разграничение жанра и стиля, ограничения ручной классификации, принципы машинного обучения, различие классификации и кластеризации и методы извлечения признаков – образуют основу для анализа существующих подходов.

2. АНАЛИЗ СУЩЕСТВУЮЩИХ ПОДХОДОВ К КЛАССИФИКАЦИИ МУЗЫКАЛЬНЫХ ЖАНРОВ

2.1 Алгоритмы классификации музыкальных жанров

После рассмотрения теоретических основ и методов извлечения признаков необходимо проанализировать существующие алгоритмы машинного обучения, применяемые для анализа музыкальных жанров. Поскольку решаемая в работе задача объединяет классификацию и кластеризацию, в настоящей главе последовательно рассматриваются обе группы методов, обзревается используемые в области наборы данных, проводится сравнительный анализ подходов, характеризуются метрики оценки качества и выявляются проблемы существующих решений, обосновывающие выбор предлагаемого подхода.

Изложение построено по принципу «от классических методов к современным». Исторически анализ музыкальных жанров опирался на сочетание вручную сконструированных признаков (мел–частотных кепстральных коэффициентов, спектральных и ритмических дескрипторов) с классическими алгоритмами машинного обучения. Развитие глубокого обучения сместило акцент в сторону моделей, способных извлекать признаки непосредственно из данных, что качественно изменило как точность, так и саму методологию исследований. Этот переход прослеживается на протяжении всей главы.

Задача классификации и кластеризации музыкальных жанров входит в более широкую область поиска музыкальной информации (Music Information Retrieval, MIR), объединяющую методы анализа, индексирования и поиска аудиоданных. В рамках этой области жанровая классификация рассматривается как одна из базовых задач, на которой апробируются новые подходы к представлению и анализу звука. Накопленный в MIR опыт – стандартизированные наборы данных, признаковые представления и метрики – формирует методологическую основу, используемую и в настоящей работе. Понимание места решаемой задачи в этом более широком контексте позволяет корректно соотнести предлагаемое решение с существующими исследованиями.

Классификация музыкальных жанров относится к задачам обучения с учителем и опирается на размеченные данные. После извлечения признаков из аудиосигналов применяются алгоритмы, сопоставляющие признаковое описание заранее заданным классам. Первые систематические работы в этой области, в частности исследование Tzanetakis и Cook (2002) [4], в котором был предложен ставший эталонным набор данных GTZAN[4], использовали статистические признаки и сравнительно простые классификаторы, достигая точности около шестидесяти процентов. Эти результаты задали отправную точку, относительно которой оценивается прогресс последующих методов.

Метод опорных векторов (SVM). Строит в пространстве признаков гиперплоскость, максимально разделяющую объекты разных классов с наибольшим зазором; ближайшие к границе объекты называются опорными векторами. Благодаря так называемому ядровому приёму метод способен разделять и линейно неразделимые классы, отображая признаки в пространство более высокой размерности; на практике для аудиозадач часто применяются линейное и радиально–базисное (RBF) ядра. SVM эффективен при высокой размерности признакового пространства и чётком разделении классов, что характерно для представлений на основе MFCC и хрома–признаков, и долгое время оставался одним из основных методов в области. Недостатками являются вычислительная сложность на больших объёмах данных и существенная зависимость качества от выбора ядра и параметров регуляризации, требующих тщательной настройки, например методом перекрёстной проверки.

Случайный лес (Random Forest). Ансамблевый метод, объединяющий множество деревьев решений, каждое из которых обучается на случайной подвыборке данных и случайном подмножестве признаков; итоговое решение принимается голосованием деревьев. Такой механизм усреднения снижает дисперсию модели и обеспечивает устойчивость к переобработке. Случайный лес хорошо работает с разнородными признаками, не требует сложной настройки и предоставляет оценку важности признаков, что ценно для анализа того, какие акустические характеристики наиболее значимы при различении

жанров. Основным ограничением является относительно высокая вычислительная стоимость по сравнению с простейшими алгоритмами и меньшая эффективность на очень высокоразмерных разреженных представлениях.

Метод k ближайших соседей (k -NN). Относит объект к классу, преобладающему среди k ближайших к нему примеров в пространстве признаков. Это так называемый ленивый алгоритм: он не строит явной модели на этапе обучения, а откладывает все вычисления до момента классификации. Метод прост, интуитивно понятен и непараметричен, то есть не делает предположений о распределении данных, и потому нередко используется как базовый ориентир. Вместе с тем он требует значительных вычислений при классификации больших объёмов данных, чувствителен к выбору метрики расстояния и масштабированию признаков, а также подвержен влиянию «проклятия размерности», при котором понятие близости в многомерном пространстве теряет различающую способность.

Наивный байесовский классификатор и градиентный бустинг. Наивный байесовский классификатор основан на теореме Байеса и предположении о независимости признаков; несмотря на упрощённость этого предположения, он отличается высокой скоростью обучения и в ряде задач показывает приемлемые результаты, выступая удобным базовым методом. Градиентный бустинг, напротив, последовательно строит слабые модели (как правило, неглубокие деревья), каждая из которых исправляет ошибки предыдущих, и объединяет их в сильный классификатор. Бустинг обеспечивает высокую точность и широко применяется на табличных признаках, однако более требователен к настройке и подвержен переобучению при отсутствии регуляризации.

Нейронные сети. Представляют собой класс моделей из связанных слоёв нейронов, способных выявлять сложные нелинейные зависимости в данных. Простейшей архитектурой является многослойный перцептрон (MLP), применяемый к агрегированным признакам. Однако наибольшую эффективность в задачах анализа музыки демонстрируют свёрточные нейронные

сети (CNN), обрабатывающие спектрограммы подобно изображениям: они автоматически выделяют иерархию признаков – от простых частотно–временных узоров на ранних слоях до сложных, специфичных для жанров паттернов на глубоких [5]. Рекуррентные сети (RNN), и в особенности их вариант с долгой краткосрочной памятью (LSTM), учитывают временную динамику сигнала, моделируя последовательность звуковых событий. Переход к нейросетевым моделям обеспечил заметный рост точности по сравнению с классическими методами и постепенно вытеснил ручное конструирование признаков.

Современные подходы. Дальнейшее развитие области привело к появлению более сложных архитектур. Гибридные свёрточно–рекуррентные сети (CRNN), предложенные в том числе в работе Choi и др. [6] [7], объединяют способность свёрточных слоёв извлекать локальные частотно–временные признаки с умением рекуррентных слоёв моделировать их последовательность во времени. Механизм внимания и архитектуры на основе трансформеров позволяют учитывать долговременные зависимости в сигнале и в последние годы всё активнее применяются к аудиоданным. Отдельным и практически важным направлением стал перенос обучения (transfer learning): модели, предварительно обученные на больших аудиокорпусах [8] [9], используются как универсальные экстракторы признаков, а их внутренние представления (эмбединги) подаются на вход относительно простого классификатора. Такой подход особенно эффективен при ограниченном объёме размеченных данных и концептуально близок к используемому в настоящей работе механизму извлечения эмбедингов из обученной сети. Наконец, в новейших исследованиях развивается самообучение (self-supervised learning), при котором содержательные представления формируются вообще без жанровой разметки.

Таким образом, эволюция методов классификации отражает общий сдвиг от инженерии признаков к их автоматическому обучению. Если классические алгоритмы требовали тщательного отбора и расчёта признаков вручную, то современные нейросетевые модели переносят эту работу внутрь самой модели,

что повышает точность, но усиливает требования к объёму данных и вычислительным ресурсам и снижает интерпретируемость.

Если оценивать прогресс количественно, то на эталонном наборе GTZAN классические методы на ручных признаках обеспечивали точность приблизительно от шестидесяти до семидесяти процентов, тогда как свёрточные и гибридные нейросетевые модели подняли этот показатель ориентировочно до семидесяти пяти – восьмидесяти пяти процентов, а подходы на основе предобученных представлений в отдельных работах приближаются к девяноста процентам. Эти цифры следует воспринимать с осторожностью: различия в методике разбиения данных, предобработке и метриках затрудняют прямое сравнение, а отмеченные дефекты самого набора задают объективный предел достижимой точности. Тем не менее общая тенденция роста качества по мере усложнения моделей прослеживается отчётливо.

Отдельного внимания заслуживают ансамблевые и комбинированные стратегии. Объединение нескольких разнородных классификаторов – например, усреднение их предсказаний или каскадное применение – нередко повышает устойчивость и точность по сравнению с отдельными моделями. В контексте анализа музыки распространено и сочетание различных признаков представлений: спектрограмм, MFCC и хрома–признаков, подаваемых на вход единой модели либо обрабатываемых отдельно с последующим объединением. Подобные подходы согласуются с гипотезой о пользе многоуровневых признаков и косвенно обосновывают применённое в настоящей работе трёхканальное представление сигнала.

2.2 Алгоритмы кластеризации

В отличие от классификации, кластеризация относится к обучению без учителя и направлена на выявление групп схожих объектов без предварительной разметки. Применительно к музыке кластеризация позволяет обнаруживать группы треков со схожими стилистическими или акустическими характеристиками, в том числе не совпадающие с заранее заданными жанрами,

и потому особенно полезна при анализе слабо размеченных или вовсе неразмеченных коллекций. Рассмотрим основные алгоритмы.

Метод k –средних (k –means). Разбивает данные на заранее заданное число k кластеров, итеративно чередуя два шага: отнесение каждого объекта к ближайшему центру и пересчёт центров как средних значений принадлежащих им объектов (алгоритм Ллойда). Цель – минимизация суммарного внутрикластерного отклонения. Метод прост, быстр и хорошо масштабируется, что делает его наиболее распространённым. Его ограничения: необходимость заранее задать число кластеров (для выбора которого применяют эвристики «локтя» или максимизацию силуэта), чувствительность к начальной инициализации центров и неявное предположение о примерно сферической, сопоставимой по размеру форме групп.

Иерархическая кластеризация. Строит вложенную иерархию кластеров путём последовательного объединения наиболее близких групп (агломеративный подход) или, наоборот, разделения (дивизивный подход). Результат представляется в виде дендрограммы, по которой можно выбрать желаемое число кластеров уже после построения. Выбор критерия связи (одионочная, полная, средняя связь, метод Уорда) существенно влияет на форму получаемых кластеров. Метод нагляден и не требует заранее задавать число групп, однако обладает высокой вычислительной сложностью и потому плохо подходит для очень больших наборов данных.

DBSCAN. Плотностный алгоритм, выделяющий кластеры как связанные области высокой плотности точек, разделённые областями низкой плотности. Объекты делятся на корневые, граничные и шумовые в зависимости от числа соседей в заданном радиусе. DBSCAN не требует задавать число кластеров, способен находить группы произвольной формы и явно выделяет выбросы, что удобно для обнаружения нетипичных композиций. Его недостатки – чувствительность к выбору параметров радиуса и минимального числа соседей, а также трудности при работе с кластерами существенно различной плотности.

Спектральная кластеризация. Использует собственные векторы матрицы сходства (лапласиана графа близости) для отображения данных в пространство меньшей размерности, в котором уже применяется, например, метод k –средних. Подход эффективен при сложной, невыпуклой структуре данных, где классический k –means не справляется, но требует построения и разложения матрицы сходства, что вычислительно затратно на больших выборках.

Модель смеси гауссиан (GMM). Описывает данные как совокупность нескольких нормальных распределений, параметры которых оцениваются алгоритмом максимизации ожидания (EM). В отличие от k –средних, GMM обеспечивает мягкое отнесение объекта к нескольким кластерам с определёнными вероятностями, что особенно естественно для музыки с её пограничными, смешанными по стилю композициями, и допускает кластеры эллиптической формы.

Прочие методы. Помимо перечисленных, применяются сдвиг среднего (mean–shift), не требующий задания числа кластеров и отыскивающий моды плотности; распространение близости (affinity propagation), выбирающее представителей кластеров на основе обмена «сообщениями» между объектами; а также BIRCH, ориентированный на эффективную кластеризацию очень больших наборов данных. Выбор конкретного метода определяется объёмом данных, предполагаемой формой кластеров и доступными вычислительными ресурсами.

Понижение размерности. Кластеризация высокоразмерных данных, к которым относятся аудиопризнаки и эмбединги, нередко предваряется понижением размерности. Метод главных компонент (PCA) выделяет направления наибольшей дисперсии и используется для линейного сжатия признаков. Нелинейные методы t –SNE и UMAP, напротив, ориентированы прежде всего на визуализацию: они отображают данные в двумерное пространство, сохраняя локальную структуру, и позволяют наглядно оценить разделимость групп [10]. Эти методы играют вспомогательную, но важную роль, повышая качество и интерпретируемость кластеризации.

Важной особенностью кластеризации является зависимость результата от используемого пространства признаков и меры близости: качественная кластеризация возможна лишь тогда, когда признаковое представление действительно отражает сходство объектов. Поэтому в современных работах кластеризацию всё чаще проводят не по исходным признакам, а по эмбедингам, полученным с помощью нейронных сетей, что обеспечивает музыкально осмысленную структуру пространства.

Отдельной практической проблемой является выбор числа кластеров и оценка качества разбиения в отсутствие эталонной разметки. Для этого применяются эвристические методы – правило «локтя», основанное на анализе зависимости внутрикластерной дисперсии от числа кластеров, и максимизация силуэтного коэффициента, оценивающего одновременно компактность и разделённость групп. Эти методы не дают однозначного ответа, и окончательный выбор нередко опирается на содержательную интерпретацию результата, что отличает кластеризацию от классификации с её объективной разметкой.

В прикладном плане кластеризация музыкальных данных находит применение в автоматическом формировании плейлистов, рекомендательных системах и сервисах музыкального обнаружения: группировка треков по сходству позволяет предлагать пользователю композиции, близкие к уже понравившимся, даже при отсутствии явных жанровых меток. Кроме того, кластеризация служит инструментом разведочного анализа больших коллекций, помогая выявлять характерные группы и нетипичные записи. Эти сценарии подчёркивают практическую ценность задачи, решаемой в настоящей работе. Сводная характеристика основных алгоритмов приведена в таблице 3.

Таблица 3 – Сравнительная характеристика алгоритмов кластеризации

Алгоритм	Принцип	Особенности
k–средних	Минимизация внутрикластерных расстояний	Требует задать число кластеров; быстр; сферические группы
Иерархическая	Объединение/разделение по расстоянию	Не требует числа кластеров; даёт дендрограмму; ресурсоёмка
DBSCAN	Области высокой плотности точек	Группы произвольной формы; выделяет шум; чувствителен к параметрам
Спектральная	Разложение матрицы сходства	Работает с невыпуклой структурой; вычислительно сложна
GMM	Смесь нормальных распределений	Мягкое отнесение к кластерам с вероятностями
Mean–shift	Поиск мод плотности	Не требует числа кластеров; чувствителен к ширине окна

2.3 Наборы данных для оценки алгоритмов

Объективное сравнение алгоритмов невозможно без стандартизированных наборов данных, обеспечивающих воспроизводимость и сопоставимость результатов разных исследований. В области анализа музыкальных жанров сложился ряд общепринятых наборов, различающихся по объёму, разметке и условиям распространения.

GTZAN. Исторически первый и наиболее распространённый набор для классификации жанров, предложенный Tzanetakis и Cook (2002). Содержит 1000 аудиофрагментов длительностью 30 секунд, равномерно распределённых по десяти жанрам. Благодаря сбалансированности и простоте структуры GTZAN стал де-факто стандартом, однако, как показал критический анализ Sturm [11], содержит повторы, ошибочные метки и записи одного исполнителя в разных частях, что необходимо учитывать при интерпретации результатов.

FMA (Free Music Archive). Крупный набор с открытой лицензией, предложенный Defferrard [12] и доступный в нескольких размерах – от нескольких тысяч до более ста тысяч треков с иерархической жанровой разметкой. Благодаря масштабу FMA лучше подходит для обучения глубоких моделей и оценки их обобщающей способности, чем небольшой GTZAN.

MagnaTagATune. Набор из примерно двадцати пяти тысяч коротких клипов, снабжённых множественными тегами (жанр, инструменты, настроение), полученными в ходе игрового краудсорсинга. Используется преимущественно для задач многометочного автоматического тегирования музыки.

Million Song Dataset (MSD) и MTG–Jamendo. Million Song Dataset [13] содержит метаданные и предвычисленные признаки для миллиона композиций, но не включает само аудио из-за лицензионных ограничений. MTG–Jamendo – крупный набор из нескольких десятков тысяч треков под свободной лицензией с многометочной разметкой, удобный для исследований на полном аудио. Краткая характеристика рассмотренных наборов приведена в таблице 4.

Таблица 4. – Наборы данных для анализа музыкальных жанров

Набор данных	Объём (порядок)	Особенности
GTZAN	1000 фрагментов, 10 жанров	Эталонный, сбалансированный; известны дефекты разметки
FMA	до сотен тысяч треков	Открытая лицензия; иерархические жанры; разные размеры
MagnaTagATune	~25 тысяч клипов	Многометочные теги; автоматическое тегирование
Million Song Dataset	1 млн композиций	Метаданные и признаки без аудио
MTG – Jamendo	десятки тысяч треков	Свободная лицензия; полное аудио; многометочность

Выбор набора данных всегда представляет компромисс. Небольшие сбалансированные наборы вроде GTZAN удобны для апробации и сопоставления с многочисленными предшествующими работами, но ограничивают достижимую точность и не отражают всего разнообразия реальной музыки. Крупные наборы дают более надёжную оценку обобщающей способности, однако сложнее в подготовке и менее сопоставимы с классическими результатами. В настоящей работе используется GTZAN – как ради сопоставимости, так и для того, чтобы сосредоточить внимание на самой методике, а не на особенностях данных.

Независимо от выбранного набора, важную роль играет предобработка данных. Стандартизация частоты дискретизации, приведение записей к единому формату, нормализация громкости и удаление артефактов обеспечивают сопоставимость примеров и устойчивость обучения. Распространённой

практикой является и сегментация записей на короткие фрагменты, увеличивающая объём выборки. Дополнительно применяется аугментация – искусственное расширение обучающих данных за счёт допустимых преобразований (добавление шума, сдвиг во времени, маскирование частотно–временных областей), повышающая обобщающую способность модели в условиях ограниченного объёма данных. Качество предобработки нередко влияет на итоговый результат не меньше, чем выбор самого алгоритма.

2.4 Сравнительный анализ подходов

Рассмотренные методы обладают различными сильными и слабыми сторонами, что определяет область их применимости:

- Метод опорных векторов эффективен при высокой размерности и чётком разделении классов, но плохо масштабируется на большие объёмы данных,
- Случайный лес устойчив к переобучению и удобен для анализа важности признаков, однако вычислительно более затратен, чем простейшие алгоритмы,
- Метод k ближайших соседей прост и непараметричен, но дорог при больших объёмах и чувствителен к масштабированию,
- Наивный байесовский классификатор быстр, но его точность снижается при нарушении предположения о независимости признаков,
- Нейронные сети обеспечивают наивысшую точность за счёт автоматического извлечения признаков, но требуют значительных данных и вычислительных ресурсов и менее интерпретируемы.

Ключевые аспекты сравнения классификаторов приведены в таблице 5.

Таблица 5. – Сравнение алгоритмов классификации

Алгоритм	Преимущества	Недостатки
SVM	Эффективен при высокой размерности и чётком разделении классов	Сложность на больших данных; настройка ядра и регуляризации
Нейронные сети (CNN/RNN)	Высокая точность; автоматическое извлечение признаков; гибкость	Требуют много данных и ресурсов; риск переобучения; «чёрный ящик»
Случайный лес	Устойчивость к переобучению; оценка важности признаков	Относительно высокая вычислительная сложность
k-NN	Простота; непараметричность	Дорог при больших объёмах; чувствителен к масштабу признаков
Наивный Байес	Простота; высокая скорость обучения	Низкая точность при нарушении независимости признаков

При выборе метода необходимо учитывать совокупность факторов: объём и качество размеченных данных, размерность признакового пространства, требуемую точность, доступные вычислительные ресурсы и важность интерпретируемости результатов. Для небольших наборов с тщательно подобранными признаками классические методы нередко оказываются достаточными и более экономичными; для больших наборов и сложных зависимостей преимущество получают глубокие нейронные сети.

Существенным критерием сравнения является и вычислительная сложность на разных этапах. Метод k-ближайших соседей практически не требует обучения, но дорог на этапе классификации, поскольку для каждого

нового объекта вычисляются расстояния до всех обучающих примеров. Метод опорных векторов, напротив, затратен при обучении на больших выборках, но быстр при выводе. Случайный лес и градиентный бустинг занимают промежуточное положение и хорошо распараллеливаются. Нейронные сети требуют наиболее ресурсоёмкого обучения, как правило с использованием графических ускорителей, однако после обучения обеспечивают быстрый вывод. Эти различия особенно важны при выборе метода для систем, работающих с большими коллекциями или в режиме, близком к реальному времени, где ограничения по памяти и скорости отклика могут оказаться решающими наряду с точностью.

Сопоставление двух групп методов показывает, что классификация и кластеризация решают взаимодополняющие задачи: первая обеспечивает отнесение трека к известному жанру, вторая выявляет скрытую структуру данных и сходство между композициями. В большинстве рассмотренных работ эти задачи решаются изолированно, тогда как их объединение в единой системе, при котором результаты классификации и кластеризации усиливают друг друга, остаётся недостаточно проработанным направлением.

2.5 Метрики оценки качества

Сравнение и обоснованный выбор алгоритмов невозможны без корректных метрик оценки. Для задач классификации базовой метрикой служит доля верных предсказаний (Accuracy), дающая общую характеристику качества. Однако в многоклассовых задачах она не отражает распределение ошибок между классами и может вводить в заблуждение при несбалансированных данных. Поэтому дополнительно вычисляются точность по классу (Precision) и полнота (Recall), а также их гармоническое среднее – F1–мера, обеспечивающая сбалансированную оценку. Для агрегирования по классам применяются macro–усреднение, придающее всем классам равный вес, и weighted–усреднение, учитывающее их размер.

Детальный анализ ошибок проводится по матрице ошибок, наглядно показывающей, какие жанры путаются между собой; диагональ матрицы

соответствует верным предсказаниям, а внедиагональные элементы – ошибкам. Такой анализ особенно важен в музыкальной классификации, где ошибки между близкими по звучанию жанрами часто отражают объективную размытость границ, а не недостатки модели.

Совместное использование точности и полноты обусловлено тем, что между ними существует естественный компромисс: стремление повысить полноту (распознать как можно больше объектов класса) обычно увеличивает число ложных срабатываний и снижает точность, и наоборот. F1–мера, объединяя обе величины, позволяет оценить этот баланс одним числом, однако анализ по отдельным классам остаётся необходимым: высокое среднее значение может маскировать систематически плохое распознавание отдельного, например, малопредставленного, жанра. Поэтому корректная оценка предполагает рассмотрение метрик как в агрегированном виде, так и по каждому классу в отдельности.

Для задач кластеризации, где истинные метки при оценке формально не используются, применяются внутренние и внешние метрики. К внутренним относятся силуэт, индексы Дэвиса – Болдина и Калински – Харабаса, оценивающие компактность и разделённость кластеров без привлечения разметки. К внешним – скорректированный индекс Рэнда (ARI) и нормированная взаимная информация (NMI), измеряющие согласованность полученного разбиения с эталонной разметкой. Поскольку используемые в настоящей работе метрики с соответствующими формулами подробно рассматриваются в главе 3, здесь они приведены обзорно – как принятый в области инструментарий оценки, на который опираются сравниваемые подходы.

2.6 Гибридные подходы и обучение представлений в анализе музыки

Отдельного рассмотрения заслуживает направление, в рамках которого классификация и кластеризация не противопоставляются, а объединяются, а также подходы, основанные на обучении представлений. Именно это направление наиболее близко к решению, предлагаемому в настоящей работе, и потому требует более подробного анализа.

Ключевой идеей обучения представлений является формирование такого пространства признаков, в котором семантически близкие объекты располагаются рядом. В задачах анализа музыки это означает, что треки со схожим звучанием получают близкие векторные представления – эмбединги. Подобные представления могут быть получены как побочный результат обучения классификатора (в качестве эмбединга используется выход одного из внутренних слоёв сети), так и с помощью специализированных методов – автокодировщиков, метрического обучения или самообучения, не требующего разметки. Важное преимущество такого подхода состоит в том, что одно и то же представление пригодно сразу для нескольких задач: классификации, поиска похожих треков и кластеризации.

Перенос обучения развивает эту идею: представления, выученные на большом аудиокорпусе, переносятся на целевую задачу с ограниченным объёмом данных. Предобученные модели (VGGish, PANNs, OpenL3, YAMNet) выступают универсальными экстракторами признаков, поверх которых обучается сравнительно простой классификатор или применяется кластеризация. Это позволяет достигать высокого качества даже на небольших наборах, где обучение глубокой сети с нуля было бы затруднено из-за переобучения.

Сочетание классификации и кластеризации в единой системе встречается в литературе реже и реализуется по-разному: кластеризация может применяться для предварительного структурирования неразмеченных данных перед классификацией, для постобработки и анализа предсказаний или, как в настоящей работе, для выявления сходства треков в пространстве эмбедингов обученного классификатора. Последний вариант объединяет достоинства обоих подходов: классификатор обеспечивает отнесение к известным жанрам, а кластеризация в том же пространстве раскрывает скрытую структуру коллекции, в том числе пересекающую жанровые границы. При этом качество кластеризации напрямую зависит от качества классификатора, формирующего пространство представлений, что делает задачи взаимозависимыми и взаимно

усиливающими друг друга. Анализ показывает, что, несмотря на перспективность, такое объединение остаётся недостаточно проработанным, что и определяет одно из направлений настоящего исследования.

2.7 Проблемы, ограничения существующих подходов и обоснование выбранного решения

Несмотря на значительный прогресс, существующие подходы к анализу музыкальных жанров обладают рядом нерешённых проблем, которые ограничивают их эффективность и определяют направления дальнейших исследований.

Субъективность и нечёткость границ. Жанровая классификация во многом субъективна, а границы между жанрами размыты и пересекаются (например, rock и metal, jazz и blues, pop и disco). Это затрудняет как ручную, так и автоматическую классификацию и усугубляется постоянным появлением гибридных жанровых форм. По сути, сама постановка задачи однозначного отнесения трека к единственному жанру является упрощением реальной картины.

Разнообразие и неоднородность данных. Музыкальные записи существенно различаются по качеству, инструментарию, темпу, тональности и стилю исполнения. Модель, обученная на однородных данных (например, на студийных записях), может плохо работать на записях иного характера. Это требует устойчивых методов предобработки и обучения, а также аугментации данных для повышения обобщающей способности.

Проклятие размерности и отбор признаков. Извлечение большого числа признаков повышает размерность пространства и вычислительную сложность, а также может ухудшать качество из-за избыточности и шума. Это вызывает необходимость отбора или снижения размерности признаков, выбор которых сам по себе является нетривиальной задачей, требующей учёта специфики музыкальных данных.

Недостаток размеченных данных. Обучение моделей с учителем требует значительных объёмов размеченных данных, создание которых трудоёмко и

привлекает экспертов. Этот дефицит особенно остро ощущается для редких жанров и стилей. В качестве ответа развиваются методы обучения с малым числом меток – полуконтролируемое и активное обучение, перенос обучения с предобученных моделей, а также предварительная кластеризация неразмеченных данных методами обучения без учителя.

Интерпретируемость моделей. Многие современные алгоритмы, в особенности глубокие нейронные сети, представляют собой «чёрные ящики», что затрудняет понимание того, какие именно характеристики влияют на отнесение произведения к тому или иному жанру. В прикладных задачах это снижает доверие к модели и усложняет её отладку, что стимулирует развитие методов интерпретации.

Утечка данных и воспроизводимость. В части работ некорректная организация выборок – в частности, попадание коррелированных фрагментов одной записи одновременно в обучающую и тестовую части – приводит к утечке данных и систематическому завышению оценок точности. Это, наряду с различиями в наборах данных и метриках, затрудняет честное сравнение результатов разных исследований и снижает их воспроизводимость.

Вычислительная стоимость и применимость в реальном времени. Современные глубокие модели требуют значительных вычислительных ресурсов как при обучении, так и при выводе. В прикладных сценариях – потоковых сервисах, мобильных приложениях – важна не только точность, но и скорость отклика при ограниченных ресурсах. Это порождает потребность в облегчённых архитектурах, сжатии и оптимизации моделей, а также в разумном балансе между качеством классификации и вычислительными затратами, что не всегда учитывается в исследованиях, ориентированных исключительно на максимизацию точности.

Проведённый анализ показывает, что существующие подходы, как правило, решают задачи классификации и кластеризации по отдельности, недостаточно используют внутренние представления нейронных сетей для выявления сходства треков и не всегда обеспечивают корректную методику

оценки. Эти ограничения непосредственно определяют направление настоящей работы: предлагается гибридная модель, в которой свёрточная нейронная сеть выполняет классификацию, а формируемые ею эмбединги используются для кластеризации треков по сходству, при строгом разделении данных на уровне треков, исключая утечку.

3 МЕТОДОЛОГИЯ И ПРОЕКТИРОВАНИЕ СИСТЕМЫ КЛАССИФИКАЦИИ И КЛАСТЕРИЗАЦИИ МУЗЫКАЛЬНЫХ ЖАНРОВ

В данной главе описывается методология построения разработанной системы анализа музыкальных жанров. Последовательно рассматриваются используемый набор данных и его подготовка, методы извлечения признаков из аудиосигналов, архитектура свёрточной нейронной сети, гибридный механизм, объединяющий классификацию и кластеризацию, применяемые метрики оценки качества, а также программная реализация системы. Изложение построено так, чтобы каждое проектное решение было обосновано и связано с особенностями предметной области, выявленными в предыдущих главах.

3.1 Подготовка и описание набора данных

Качество и репрезентативность исходных данных определяют верхнюю границу достижимого результата в любой задаче машинного обучения, поэтому выбор и корректная подготовка набора данных рассматриваются как самостоятельный этап исследования.

В качестве исходного набора использован GTZAN Genre Collection – де-факто стандарт для оценки алгоритмов классификации музыкальных жанров. Выбор обусловлен его сбалансированностью, простотой структуры и широкой распространённостью в научной литературе, что обеспечивает сопоставимость результатов с другими работами. Структура набора приведена в таблице 6.

Таблица 6. – Структура набора данных GTZAN

Характеристика	Значение
Общее число аудиофайлов	1000
Число жанров (классов)	10
Файлов на жанр	100
Длительность записи	30 с
Формат, частота дискретизации	WAV, 22050 Гц, моно

Несмотря на удобство, набор обладает рядом известных ограничений, отмеченных в литературе: наличие повторяющихся фрагментов, спорные жанровые метки и присутствие записей одного исполнителя в разных частях набора. Эти особенности способны искусственно завышать оценки точности, поэтому они учитывались при организации эксперимента. В рамках данной работы GTZAN используется как базовый набор для апробации предложенного подхода, а его ограничения трактуются как фактор, задающий объективную верхнюю границу качества.

Каждая 30-секундная запись сегментируется на десять непересекающихся фрагментов длительностью 3 секунды. Такая стратегия, во-первых, увеличивает объём обучающей выборки на порядок (с 1000 до 10000 примеров, таблица 7), а во-вторых, согласуется с тем, что жанровая принадлежность во многом определяется локальными характеристиками сигнала. Число отсчётов в одном сегменте составляет $N_{\text{сегм}} = 22050 \times 3 = 66150$.

Таблица 7. – Результат сегментации набора данных

Параметр	До обработки	После сегментации
Количество примеров	1000	10000
Длительность примера	30 с	3 с

Проблема утечки данных и разделение по трекам. Принципиальным методологическим решением является способ формирования обучающей, валидационной и тестовой выборок. Поскольку сегменты, полученные из одной записи, акустически сильно коррелированы, наивное случайное разбиение на уровне сегментов приводит к попаданию фрагментов одной композиции одновременно в обучающую и тестовую части. Это вызывает утечку данных (data leakage) и систематическое завышение оценки обобщающей способности. Чтобы исключить этот эффект, разбиение проводится на уровне целых треков: все десять сегментов одной записи относятся только к одной из выборок. Деление выполнено в пропорции 70 % / 15 % / 15 % со стратификацией по жанрам, что сохраняет равномерное представительство классов.

Следует подчеркнуть, что выбор GTZAN не означает игнорирования его недостатков. В современных исследованиях всё чаще применяются более крупные наборы, такие как FMA или MagnaTagATune, однако они существенно сложнее в подготовке и менее удобны для сопоставления с классическими работами. Использование GTZAN в качестве базового набора позволяет, с одной стороны, опереться на обширный накопленный опыт, а с другой – сосредоточить внимание именно на предлагаемой методике, а не на особенностях конкретных данных.

Кроме того, сегментация на короткие фрагменты несёт и дополнительный смысл: модель учится распознавать жанр по локальному отрезку длительностью всего несколько секунд, что приближает задачу к условиям потоковых сервисов, где решение нередко требуется принять по короткому фрагменту композиции. Таким образом, принятая схема подготовки данных не только увеличивает объём выборки, но и повышает практическую применимость модели. Сформированный набор из 10000 сегментов с разбиением на уровне треков служит надёжной основой для последующих этапов исследования.

3.2 Предварительная обработка аудиосигналов и извлечение признаков

Аудиосигнал в исходном виде представляет собой одномерный временной ряд высокой размерности и непосредственно малоприспособлен для обучения модели. Цель данного этапа – преобразовать сигнал в компактное, информативное и устойчивое к несущественным искажениям признаковое представление.

Переход к частотно–временному представлению выполняется с помощью кратковременного преобразования Фурье (STFT), которое вычисляет спектр на коротких перекрывающихся фрагментах сигнала x с применением оконной функции w :

$$X(k, n) = \sum_m x[m] \cdot w[n - m] \cdot e^{-j2\pi km/N},$$

где $X(k, n)$ — комплексное значение спектра на частотном индексе k во временном окне n

- $x[m]$ — отсчёты входного сигнала
- $w[n-m]$ — оконная функция (окно Ханна)
- m — индекс отсчёта сигнала (по нему ведётся суммирование)
- k — индекс частотной составляющей
- n — индекс (положение) временного окна
- N — размер окна преобразования в отсчётах
- j — мнимая единица
- $e^{(-j2\pi km/N)}$ — комплексная экспонента (базисная функция преобразования Фурье).

В работе используется окно Ханна, сглаживающее границы фрагмента и снижающее эффект растекания спектра (spectral leakage). Полученная спектрограмма отражает распределение энергии сигнала по частоте во времени, однако не учитывает нелинейность слухового восприятия человека.

Для приближения к восприятию частоты преобразуются в мел-шкалу по соотношению

$$m(f) = 2595 \cdot \log_{10}(1 + f / 700)$$

где $m(f)$ — частота в мел-шкале

- f — частота в герцах
- $\log_{10}$ — десятичный логарифм
- 2595 и 700 — эмпирические константы, задающие соответствие между линейной (Гц) и перцептивной (мел) шкалами.

Энергии группируются банком мел-фильтров (используется $N_{mel} = 128$ фильтров), после чего значения переводятся в логарифмический (децибельный) масштаб относительно максимума:

$$S_{\text{дб}} = 10 \cdot \log_{10}(S / S_{\text{max}})$$

где $S_{\text{дб}}$ — значение энергии в децибельном (логарифмическом) масштабе

- S — энергия спектральной составляющей
- S_{max} — максимальная энергия в спектрограмме (опорный уровень)
- $\log_{10}$ — десятичный логарифм.

Дополнительно к каждому сегменту применяется гармонико–перкуSSIONное разделение (HPSS), основанное на медианной фильтрации спектрограммы: горизонтальная фильтрация выделяет устойчивые во времени гармонические компоненты (мелодия, гармония), вертикальная — кратковременные перкуSSIONные (ударные) [14] [15]. Исходный сигнал и две его компоненты дают три мел–спектрограммы, объединяемые в трёхканальный тензор размерности $128 \times 128 \times 3$. Такое представление аналогично цветовым каналам изображения и позволяет свёрточной сети параллельно анализировать тональную и ритмическую структуру, что особенно важно для жанров с выраженной ритмической (metal, disco) либо гармонической (classical, jazz) составляющей.

Нормализация. Значения в децибельном масштабе лежат в широком отрицательном диапазоне, что затрудняет обучение. Поэтому признаки стандартизируются по среднему и стандартному отклонению, вычисленным только по обучающей выборке и применяемым ко всем выборкам, а также на этапе вывода:

$$z = (x - \mu) / \sigma$$

где z — нормализованное значение признака,

- x — исходное значение признака,
- μ — среднее, вычисленное только по обучающей выборке,
- σ — стандартное отклонение, вычисленное только по обучающей выборке.

Использование статистик исключительно обучающей выборки предотвращает утечку информации из теста, а сохранение параметров нормализации для этапа инференса устраняет рассогласование между обучением и реальным применением.

Аугментация (SpecAugment). Для повышения устойчивости и борьбы с переобучением в условиях ограниченного объёма данных применяется аугментация спектрограмм: при обучении случайным образом обнуляются («маскируются») полосы признакового тензора вдоль оси частот и вдоль оси времени [16]. Это вынуждает сеть опираться на устойчивые, распределённые по спектрограмме закономерности, а не на отдельные фрагменты, и улучшает качество на новых данных. Аугментация применяется исключительно на этапе обучения и не затрагивает валидацию и тест.

Необходимо отметить, что выбор именно мел–спектрограммы в качестве основного представления не является единственно возможным. В ряде работ предпочтение отдаётся непосредственно мел–частотным кепстральным коэффициентам или их сочетанию с хрома–признаками. Однако в задачах, решаемых свёрточными сетями, спектрограмма обладает важным преимуществом: она сохраняет двумерную частотно–временную структуру сигнала, которую сеть способна анализировать так же, как изображение. Переход к кепстральным коэффициентам, напротив, частично разрушает эту структуру, что снижает эффективность свёрточной обработки.

Применение HPSS можно рассматривать как форму инженерии признаков, при которой исходный сигнал заранее разделяется на компоненты с различной семантикой. Подобный приём не является строго обязательным – достаточно глубокая сеть в принципе способна выделить соответствующие закономерности и самостоятельно, – однако явное разделение упрощает обучение и делает признаковое пространство более интерпретируемым. В совокупности предложенное представление сочетает информативность, устойчивость к искажениям и совместимость с архитектурой свёрточной сети, что и определило его выбор.

3.3 Архитектура свёрточной нейронной сети–классификатора

Выбор архитектуры в значительной мере определяет итоговое качество. Поскольку признаковое представление имеет вид двумерного изображения, в качестве классификатора применяется свёрточная нейронная сеть (CNN), эффективная при обработке данных с пространственной структурой. Альтернативные подходы (рекуррентные сети, гибриды CNN–RNN, трансформеры) обеспечивают сопоставимое или более высокое качество, но требуют существенно больших вычислительных ресурсов и объёмов данных

для условий настоящей работы CNN представляет рациональный баланс точности, скорости обучения и сложности.

Базовой операцией сети является дискретная свёртка входной карты признаков I с обучаемым ядром K :

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) \cdot K(m, n)$$

где $(I * K)(i, j)$ — значение карты признаков (результата свёртки) в точке (i, j) ,

- I — входная карта признаков,
- K — обучаемое ядро (фильтр) свёртки,
- i, j — координаты текущей позиции на карте признаков,
- m, n — индексы элементов ядра (по ним ведётся суммирование).

Локальность свёртки позволяет автоматически выявлять повторяющиеся частотно–временные паттерны, характерные для жанров, без ручного конструирования признаков. После каждого свёрточного слоя применяется пакетная нормализация (Batch Normalization) [17], стабилизирующая распределение активаций:

$$\hat{x} = (x - \mu_B) / \sqrt{(\sigma_B^2 + \epsilon)}, \quad y = \gamma \cdot \hat{x} + \beta$$

где x — входная активация до нормализации,

- $\hat{x}$ — нормализованное значение,
- μ_B — среднее по мини-батчу,
- σ_{B^2} — дисперсия по мини-батчу,
- ε — малая константа для численной устойчивости (защита от деления на ноль),
- y — выход после нормализации и масштабирования,
- γ — обучаемый масштабирующий параметр,
- β — обучаемый сдвиг.

Нелинейность вводится функцией активации ReLU, обеспечивающей быстрое обучение и устойчивость к затуханию градиента:

$$f(x) = \max(0, x)$$

где $f(x)$ — выход функции активации,

- x — входное значение нейрона,
- $\max(0, x)$ — выбор большего из нуля и x (отрицательные значения обнуляются).

Снижение пространственной размерности и инвариантность к небольшим сдвигам достигаются подвыборкой по максимуму (MaxPooling) с окном 2×2 . Сеть состоит из трёх последовательных свёрточных блоков (свёртка → нормализация → ReLU → подвыборка) с возрастающим числом фильтров (32, 64, 128), после чего следует полносвязная часть. Выходной слой использует функцию Softmax, формирующую вероятностное распределение по классам:

$$\sigma(z)_i = e^{z_i} / \sum_j e^{z_j}$$

где $\sigma(z)_i$ — вероятность принадлежности к i -му классу,

- z_i — выходное значение (логит) сети для i -го класса,

- z_j — логит для j -го класса,
- e — основание натурального логарифма,
- $\sum_j e^{z_j}$ — сумма экспонент по всем классам (нормирующий множитель).

Обучение проводится минимизацией категориальной перекрёстной энтропии между истинным распределением y и предсказанным $\hat{y}$:

$$L = - \sum_i y_i \cdot \log(\hat{y}_i)$$

где L — значение функции потерь (категориальной перекрёстной энтропии),

- y_i — истинная вероятность i -го класса (1 для верного класса, 0 для остальных),
- $\hat{y}_i$ — предсказанная сетью вероятность i -го класса,
- i — индекс класса (суммирование по всем классам),
- $\log$ — натуральный логарифм.

Для регуляризации в полносвязной части применяется Dropout с вероятностью 0,5 [18], случайно отключающий часть нейронов и снижающий переобучение. Оптимизация выполняется алгоритмом Adam [19] с начальным шагом обучения $1 \cdot 10^{-4}$

обучение управляется ранней остановкой по валидационной выборке с восстановлением лучших весов. Послойная структура сети приведена в таблице 8.

Таблица 8. – Послойная структура свёрточной нейронной сети

Слой	Параметры	Форма выхода
Вход	–	$128 \times 128 \times 3$
Conv2D + BN + ReLU + Pool	32 фильтра 3×3	$64 \times 64 \times 32$
Conv2D + BN + ReLU + Pool	64 фильтра 3×3	$32 \times 32 \times 64$

Conv2D + BN + ReLU + Pool	128 фильтров 3×3	16 × 16 × 128
Flatten	—	32768
Dense + BN + ReLU + Dropout	512 нейронов, p = 0,5	512
Dense (Softmax)	10 нейронов	10

Глубина сети и число фильтров подобраны исходя из компромисса между выразительной способностью модели и риском переобучения на ограниченном наборе данных. Постепенное увеличение числа фильтров от блока к блоку отражает общий принцип построения свёрточных сетей: на ранних слоях выделяются простые локальные признаки – перепады энергии, контуры, – а на более глубоких формируются их сложные комбинации, соответствующие характерным для жанров паттернам. Уменьшение пространственного разрешения за счёт подвыборки при одновременном росте числа каналов позволяет сети переходить от детального описания к обобщённому.

Выбор сравнительно небольшого шага обучения объясняется стремлением к устойчивой сходимости: слишком большой шаг приводит к колебаниям функции потерь, тогда как умеренный обеспечивает плавное обучение, пусть и ценой большего числа эпох. В сочетании с ранней остановкой это снижает риск как недообучения, так и переобучения. Спроектированная архитектура, таким образом, представляет собой взвешенное инженерное решение, а не результат механического наращивания сложности.

3.4 Гибридная модель: эмбединги и кластеризация

Классификатор отвечает на вопрос о принадлежности трека к одному из десяти заранее заданных жанров, но не выявляет скрытого сходства между композициями и не способен обнаруживать группы, не совпадающие с разметкой. Чтобы дополнить анализ, предложена гибридная модель, в которой результат классификации служит опорной структурой, а кластеризация раскрывает внутреннюю организацию данных.

Ключевая идея состоит в использовании внутреннего представления обученной сети. Выход предпоследнего полносвязного слоя (512–мерный вектор) интерпретируется как эмбединг – компактное описание звучания, в котором близость векторов отражает акустическое сходство треков. Поскольку сеть обучалась разделять жанры, это пространство несёт музыкально осмысленную структуру. Вектор трека получается усреднением посегментных эмбедингов с последующей L2–нормировкой:

$$\hat{v} = v / \|v\|$$

где $\hat{v}$ — нормированный вектор-эмбединг единичной длины,

- v — исходный усреднённый вектор-эмбединг трека,
- $\|v\|$ — евклидова норма (длина) вектора v .

Мерой близости между треками служит косинусное сходство, устойчивое к различиям в норме векторов:

$$\cos(a, b) = (a \cdot b) / (\|a\| \cdot \|b\|)$$

где $\cos(a, b)$ — косинусное сходство векторов a и b ,

- a, b — сравниваемые векторы-эмбединги двух треков,
- $a \cdot b$ — скалярное произведение,
- $\|a\|, \|b\|$ — евклидовы нормы (длины) векторов.

На основе попарной близости строится граф сходства методом k ближайших соседей: каждый трек соединяется с несколькими наиболее похожими при превышении порога близости. В таком графе плотные группы вершин образуют кластеры, в том числе пересекающие границы заранее заданных жанров, что отражает гибридную природу реальной музыки. Для явного разбиения на группы к тому же пространству эмбедингов применимы классические алгоритмы кластеризации. Метод k –средних минимизирует суммарное внутрикластерное отклонение:

$$J = \sum_k \sum_{x \in C_k} \|x - \mu_k\|^2$$

где J — минимизируемый критерий (суммарное внутрикластерное отклонение),

- k — индекс кластера,
- C_k — множество объектов k -го кластера,
- x — вектор объекта (эмбединг трека),
- μ_k — центр (центроид) k -го кластера,
- $\|x - \mu_k\|^2$ — квадрат евклидова расстояния от объекта до центра кластера.

Алгоритм DBSCAN, в отличие от k -средних, не требует задания числа кластеров и выделяет группы по плотности, относя редкие точки к шуму, что удобно для обнаружения нетипичных композиций. Сопоставление полученных кластеров с истинными жанрами (раздел 3.5) позволяет оценить, в какой мере неконтролируемая структура согласуется с экспертной разметкой.

Отдельного внимания заслуживает вопрос интерпретации эмбедингов. В отличие от исходных признаков, отдельные измерения эмбединга не имеют прямого физического смысла — это абстрактные координаты, выученные сетью. Тем не менее их совокупность образует пространство, в котором геометрическая близость соответствует акустическому сходству, что и делает возможной кластеризацию. Подобный подход, при котором обученная для одной задачи сеть используется как универсальный экстрактор признаков для другой, широко распространён в современном машинном обучении и согласуется с идеей переноса знаний.

Важно учитывать и ограничения такого решения: качество кластеризации напрямую зависит от качества обученного классификатора, поскольку именно он формирует пространство эмбедингов. Если классификатор слабо разделяет некоторые жанры, соответствующие треки окажутся перемешанными и в пространстве эмбедингов. Это означает, что задачи классификации и

кластеризации в предложенной модели не независимы, а взаимно усиливают друг друга, что и составляет суть гибридного подхода.

3.5 Метрики оценки качества

Для объективной и всесторонней оценки качества используется совокупность метрик, поскольку в многоклассовой задаче одна агрегированная величина не отражает распределение ошибок между классами. Оценка проводится отдельно для классификации и для кластеризации.

Метрики классификации. Общая доля верных предсказаний характеризуется точностью:

$$\text{Accuracy} = N_{\text{верно}} / N_{\text{всего}}$$

где Accuracy — общая точность классификации,

- $N_{\text{верно}}$ — число верно классифицированных объектов,
- $N_{\text{всего}}$ — общее число объектов.

Для каждого жанра на основе числа верно (TP) и ошибочно (FP, FN) отнесённых объектов вычисляются точность по классу и полнота:

$$\text{Precision}_i = TP_i / (TP_i + FP_i)$$

где Precision_i — точность для i-го класса,

- TP_i — число объектов i-го класса, верно отнесённых к нему,
- FP_i — число объектов других классов, ошибочно отнесённых к i-му.

$$\text{Recall}_i = TP_i / (TP_i + FN_i)$$

где Recall_i — полнота для i-го класса,

- TP_i — число верно отнесённых к i-му классу объектов,
- FN_i — число объектов i-го класса, ошибочно отнесённых к другим.

Их гармоническое среднее даёт сбалансированную оценку – F1–меру:

$$F1_i = 2 \cdot Precision_i \cdot Recall_i / (Precision_i + Recall_i)$$

где $F1_i$ — F1-мера для i -го класса,

- $Precision_i$ — точность для i -го класса,
- $Recall_i$ — полнота для i -го класса.

Для агрегирования по $K = 10$ классам применяются макро–усреднение (равный вес классов) и weighted–усреднение (с учётом размера класса n

$$Macro = (1/K) \sum_i m_i, \quad Weighted = \sum_i (n_i/N) \cdot m_i$$

где $Macro$ — макроусреднённое значение метрики (равный вес классов),

– $Weighted$ — взвешенно усреднённое значение (с учётом размера классов),

- K — число классов,
- m_i — значение метрики (Precision, Recall или F1) для i -го класса,
- n_i — число объектов i -го класса,
- N — общее число объектов.

Детальный анализ ошибок выполняется по матрице ошибок 10×10 , где диагональ соответствует верным предсказаниям, а внедиагональные элементы — путанице между жанрами.

Метрики кластеризации. Качество разбиения без использования меток оценивается силуэтом, сопоставляющим внутрикластерное $a(i)$ и межкластерное $b(i)$ расстояния:

$$s(i) = (b(i) - a(i)) / \max(a(i), b(i))$$

- где $s(i)$ — силуэтный коэффициент объекта i ,

- $a(i)$ — среднее расстояние от объекта i до остальных объектов своего кластера (внутрикластерное),
- $b(i)$ — среднее расстояние от объекта i до объектов ближайшего соседнего кластера (межкластерное),
- $\max(a(i), b(i))$ — большее из двух расстояний.

Согласованность кластеров с истинной жанровой разметкой измеряется скорректированным индексом Рэнда (ARI), учитывающим случайные совпадения, и нормированной взаимной информацией (NMI):

$$NMI(U, V) = I(U, V) / \sqrt{H(U) \cdot H(V)}$$

где $NMI(U, V)$ — нормированная взаимная информация между разбиениями U и V ,

- U — разбиение, полученное кластеризацией,
- V — эталонное разбиение (истинная жанровая разметка),
- $I(U, V)$ — взаимная информация между U и V ,
- $H(U), H(V)$ — энтропии разбиений U и V .

Применение нескольких метрик одновременно обусловлено тем, что каждая из них освещает отдельный аспект работы модели. Высокая общая точность может маскировать систематические ошибки на отдельных, плохо представленных классах, которые становятся заметны лишь при анализе полноты и точности по классам. F1–мера, в свою очередь, особенно полезна в ситуациях, когда важны оба типа ошибок одновременно. Совместное использование этих величин позволяет получить сбалансированную и непротиворечивую оценку качества классификации.

При интерпретации результатов следует учитывать наличие пограничных объектов – композиций, сочетающих признаки нескольких жанров. Для таких треков даже формально ошибочное предсказание может оказаться музыкально оправданным, а повторяющиеся ошибки между близкими жанрами скорее

свидетельствуют о размытости границ, чем о недостатках модели. Поэтому количественные метрики целесообразно дополнять качественным анализом матрицы ошибок и структуры полученных кластеров.

3.6 Программная реализация системы

Разработанная методология реализована в виде модульной программной системы, что обеспечивает воспроизводимость эксперимента и возможность практического применения. Конвейер обработки одного трека включает загрузку и декодирование аудио, сегментацию, извлечение трёхканальных мел-спектрограмм с HPSS, нормализацию, классификацию свёрточной сетью и параллельное вычисление эмбединга с последующим усреднением и кластеризацией.

Система состоит из двух взаимодействующих частей. Серверная часть загружает обученную модель, принимает аудиофайл, выполняет описанную обработку и возвращает распределение по жанрам, последовательность предсказаний по сегментам и вектор-эмбединг. Клиентская часть представляет собой веб-приложение с интерактивной визуализацией графа: жанры и их укрупнённые группы выступают опорными узлами, а добавляемые пользователем треки появляются как новые вершины, связываемые с похожими по эмбедингу композициями

визуализируется также процесс анализа и изменение предсказанного жанра по ходу трека.

Программная реализация выполнена на языке Python с использованием библиотек librosa [20] (обработка аудио и извлечение признаков), TensorFlow/Keras (построение и обучение сети), scikit-learn (разбиение данных, метрики, кластеризация) и Flask (серверная часть)

клиентская визуализация реализована средствами JavaScript. Перечень основных модулей приведён в таблице 9.

Таблица 9. – Основные программные модули системы

Модуль	Назначение
Обучение	Подготовка данных, обучение сети, сохранение модели и параметров нормализации, расчёт метрик
Извлечение признаков	Сегментация, HPSS, построение мел–спектрограмм, нормализация
Серверная часть	Приём трека, классификация, вычисление эмбединга, выдача результата
Веб–интерфейс	Интерактивный граф жанров и кластеров, визуализация процесса анализа

Разделение системы на серверную и клиентскую части отражает распространённый архитектурный принцип, при котором вычислительно сложная обработка вынесена на сторону сервера, а пользовательское взаимодействие сосредоточено в лёгком клиенте. Такой подход облегчает поддержку и развитие системы: модель может быть заменена или дообучена без изменения интерфейса, а интерфейс доработан независимо от вычислительного ядра. Модульность отдельных этапов обработки дополнительно повышает гибкость, позволяя, например, заменить метод извлечения признаков или алгоритм кластеризации без переработки всей системы.

Наконец, наглядная визуализация результатов имеет не только демонстрационное, но и аналитическое значение: представление треков в виде графа сходства позволяет визуально оценить структуру коллекции, выявить плотные группы и нетипичные композиции, что затруднительно сделать по одним лишь числовым метрикам. Тем самым программная реализация не сводится к технической оболочке, а выступает полноценным инструментом анализа музыкальных данных. По итогу был сформирован сбалансированный набор данных с корректным разбиением на уровне треков, исключаящим утечку

данных. Предложено многоканальное мел–спектрограммное представление сигнала на основе HPSS с нормализацией и аугментацией. Спроектирована свёрточная нейронная сеть–классификатор, дополненная гибридным механизмом кластеризации в пространстве эмбедингов. Определён комплекс метрик для обеих задач и описана модульная программная реализация. Совокупность принятых решений непосредственно отвечает гипотезе исследования о повышении информативности анализа за счёт сочетания глубокого обучения, многоуровневых признаков и объединения классификации с кластеризацией.

4 ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ И АНАЛИЗ РЕЗУЛЬТАТОВ

4.1 Условия проведения эксперимента

Корректная интерпретация результатов невозможна без точного описания условий, в которых они получены. Эксперимент проводился на персональном компьютере с графическим ускорителем. Программная среда включала язык Python версии 3.10 и библиотеки librosa (обработка аудио), TensorFlow/Keras (построение и обучение модели), scikit-learn (разбиение данных, расчёт метрик и кластеризация), а также Flask для серверной части.

В эксперименте использовался набор данных GTZAN, преобразованный согласно методике главы 3: каждая запись разбивалась на десять 3-секундных сегментов, что давало 10000 примеров. Разбиение на обучающую, валидационную и тестовую выборки выполнялось на уровне целых треков в пропорции 70 % / 15 % / 15 % со стратификацией по жанрам, что исключало попадание фрагментов одной композиции в разные выборки. Признаковое представление – трёхканальная мел-спектрограмма размерности $128 \times 128 \times 3$ с предварительной нормализацией. Основные гиперпараметры обучения приведены в таблице 10.

Таблица 10. – Гиперпараметры обучения модели

Параметр	Значение
Оптимизатор	Adam
Начальный шаг обучения	$1 \cdot 10^{-4}$
Размер батча	32
Максимальное число эпох	50
Ранняя остановка (patience)	10 эпох, с восстановлением лучших весов
Функция потерь	Категориальная перекрёстная энтропия
Регуляризация	Dropout (p = 0,5), Batch Normalization
Аугментация	SpecAugment (маскирование по частоте и времени)

4.2 Результаты обучения модели

Анализ процесса обучения позволяет оценить сходимость модели и наличие переобучения. На рисунках 2 и 3 приведены зависимости точности (ассурасу) и значения функции потерь (loss) от номера эпохи для обучающей и валидационной выборок.

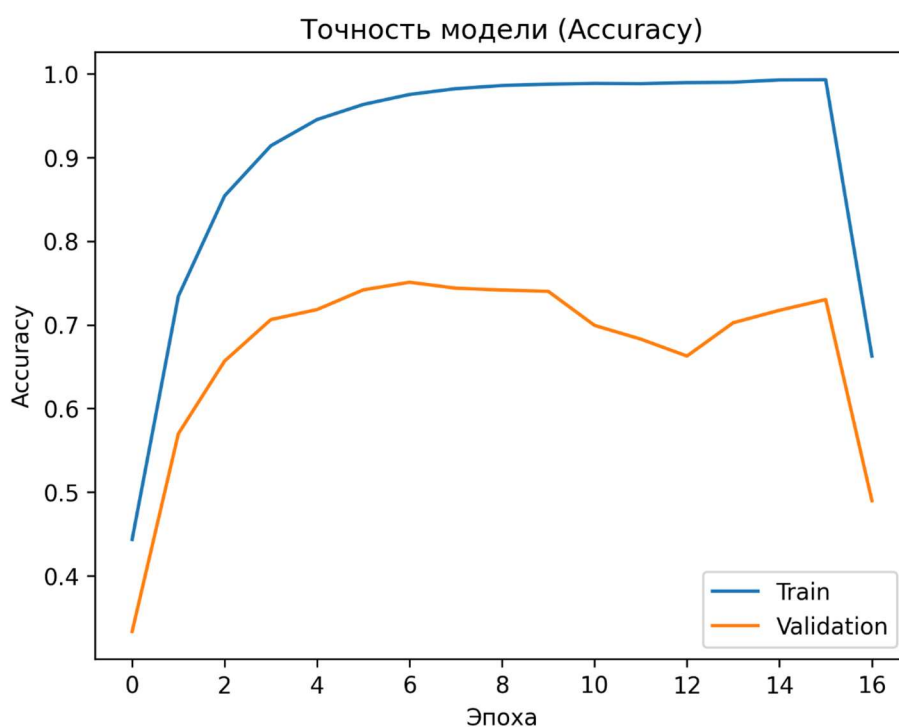


Рисунок 2. – Динамика точности по эпохам

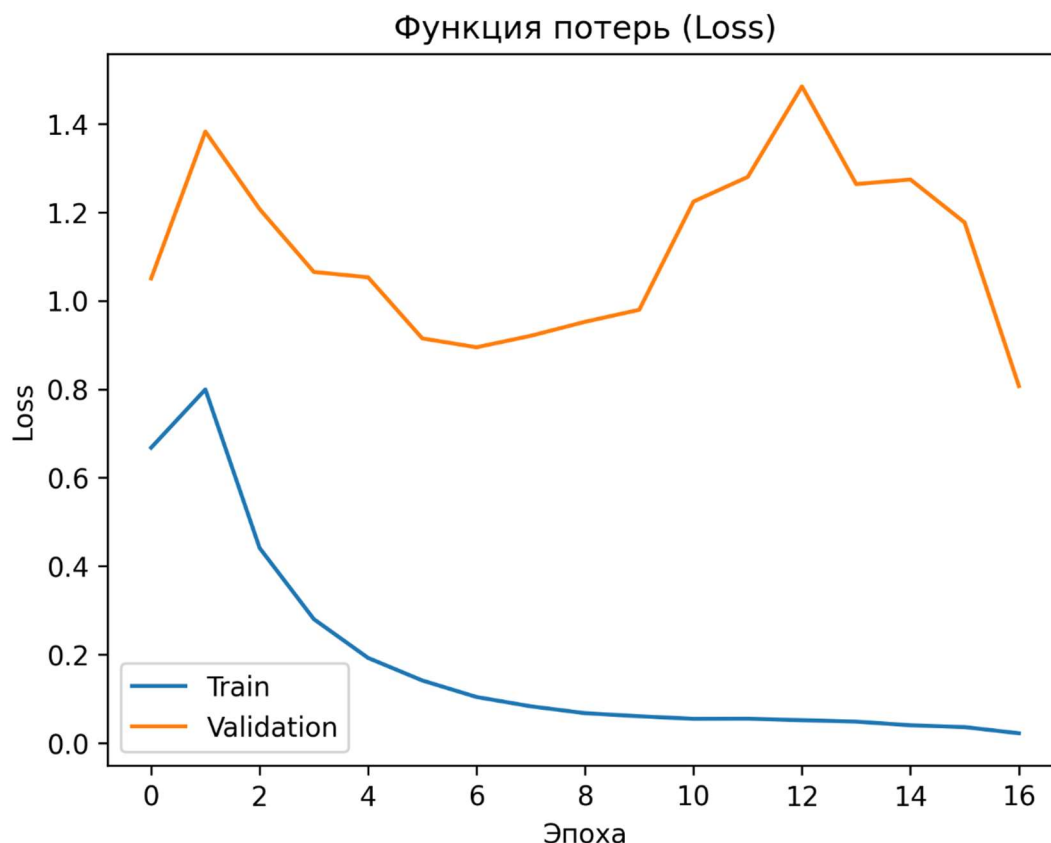


Рисунок 3. — Динамика функции потерь по эпохам

В ходе обучения точность на обучающей выборке монотонно возрастала, а функция потерь убывала. Обучение завершилось по критерию ранней остановки на 21 эпохе, после чего были восстановлены веса, соответствующие наилучшему значению валидационной точности (75.43%). Расхождение между кривыми обучения и валидации характеризует степень переобучения: умеренный разрыв свидетельствует о том, что примечаемые меры регуляризации (Batch Normalization, Dropout) и аугментация спектрограмм обеспечили приемлемый баланс между качеством подгонки и обобщающей способностью.

Характер кривых подтверждает, что выбранная стратегия обучения с небольшим шагом и ранней остановкой обеспечивает устойчивую сходимость без выраженных колебаний. Дальнейшее увеличение числа эпох не приводило к улучшению валидационных показателей, что и обусловило срабатывание ранней остановки.

4.3 Оценка качества классификации

Для всесторонней оценки качества классификации на тестовой выборке вычислены метрики Precision, Recall и F1 по каждому жанру, а также их усреднённые значения. Результаты приведены в таблице 11, а общая точность (Accuracy) на тестовой выборке составила 75.43 %.

Таблица 11. – Метрики классификации по жанрам

Жанр	Precision	Recall	F1	Support
blues	0.839	0.660	0.739	150
classical	0.915	0.792	0.849	149
country	0.651	0.725	0.686	149
disco	0.722	0.693	0.707	150
hiphop	0.839	0.900	0.868	150
jazz	0.754	0.840	0.795	150
metal	0.876	0.893	0.884	150
pop	0.726	0.760	0.743	150
reggae	0.915	0.713	0.801	150
rock	0.457	0.567	0.506	150
Macro avg	0.769	0.754	0.758	1498
Weighted avg	0.769	0.754	0.758	1498

Для детального анализа ошибок построена матрица ошибок (рисунок 4), в которой строки соответствуют истинным жанрам, а столбцы – предсказанным.

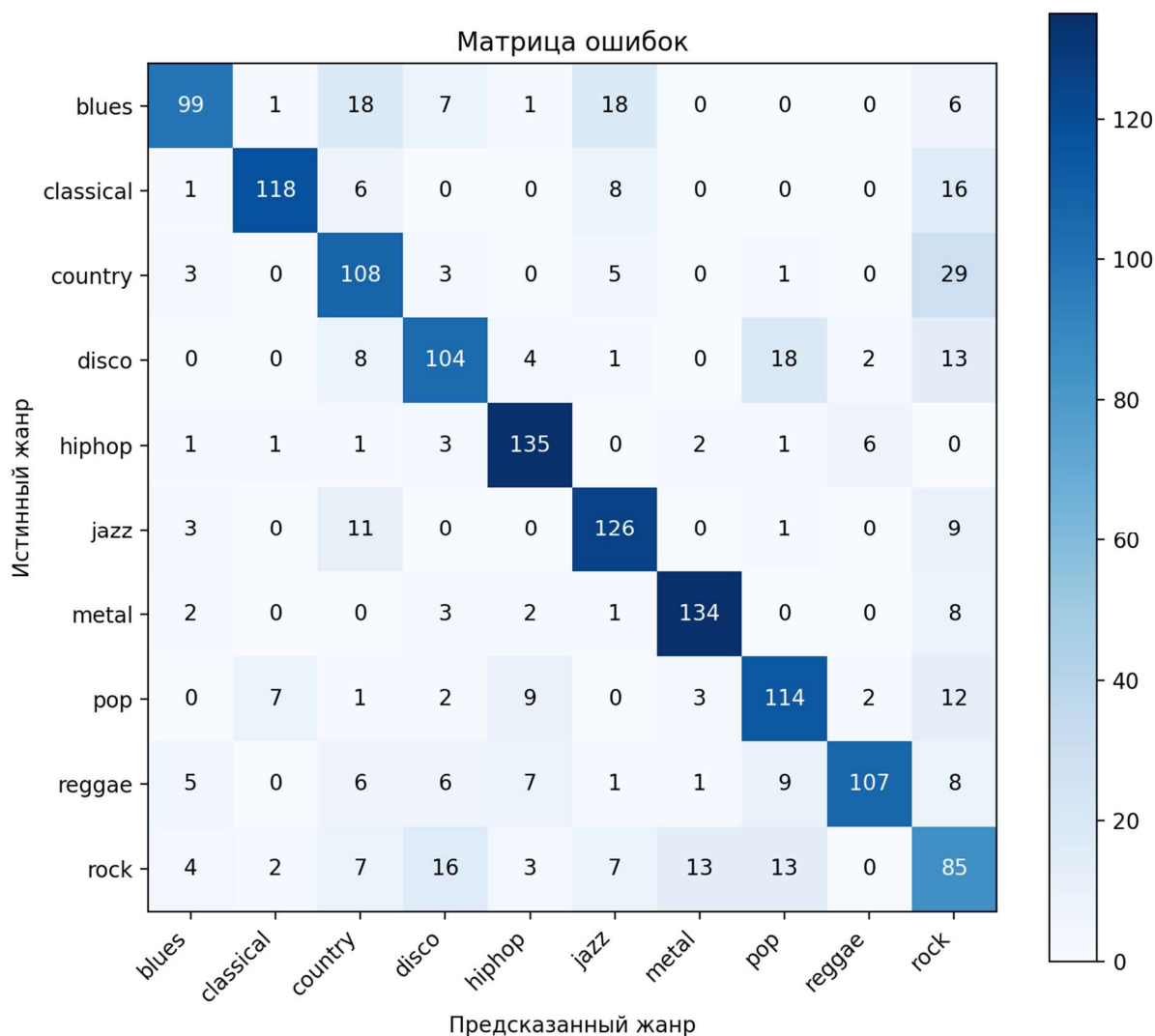


Рисунок 4. – Матрица ошибок классификатора

Анализ показывает, что наибольшие значения сосредоточены на главной диагонали, что свидетельствует о преобладании верных предсказаний. Основная доля ошибок ожидаемо приходится на пары жанров, близких по акустическим характеристикам: rock и metal, pop и disco, jazz и blues. Такая картина согласуется с теоретическими предпосылками о размытости жанровых границ: перечисленные жанры разделяют общие инструментальные, ритмические и гармонические признаки. Жанры с выраженной спецификой звучания (например, classical) распознаются, как правило, наиболее уверенно.

4.4 Анализ эмбедингов и кластеризации

Помимо классификации, исследована структура пространства эмбедингов, формируемого предпоследним слоем обученной сети. Для каждого трека получен усреднённый по сегментам и L2–нормированный вектор, после чего применены методы понижения размерности и кластеризации. Двумерная проекция эмбедингов, построенная методом t–SNE, приведена на рисунке 5 цветом обозначены истинные жанры.

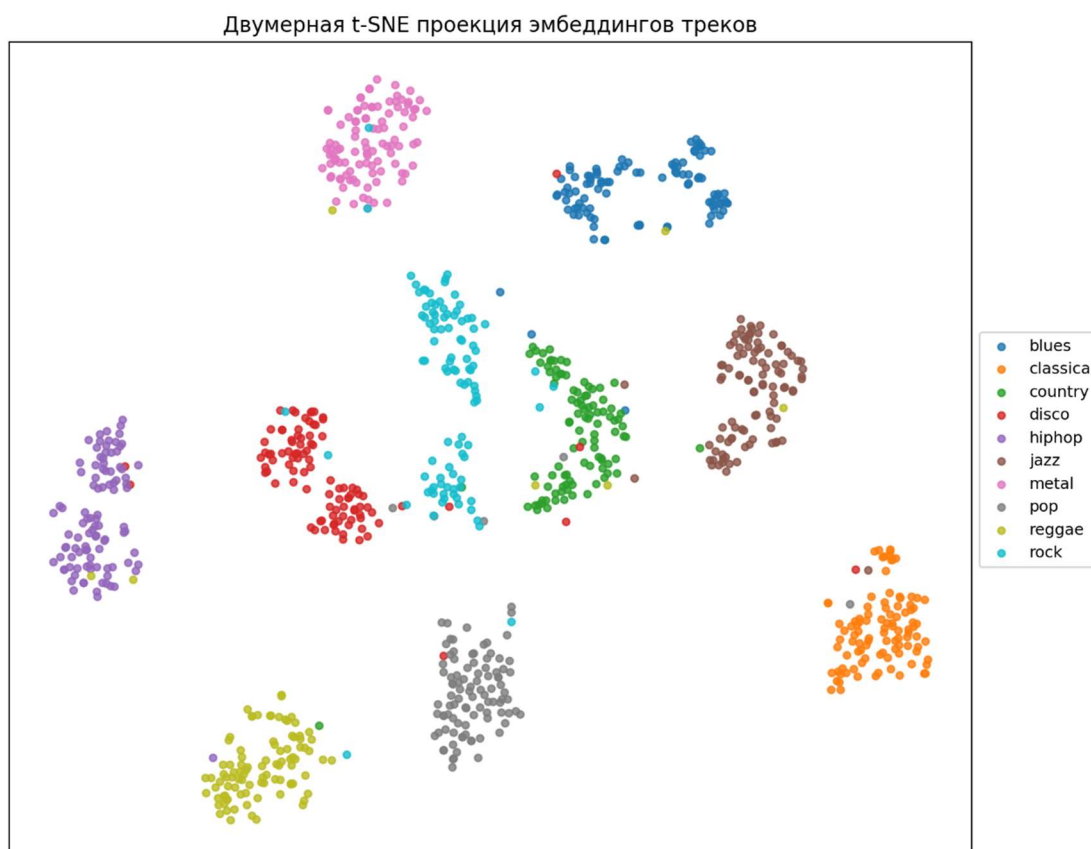


Рисунок 5. – Визуализация пространства эмбедингов

Визуальный анализ проекции позволяет оценить, насколько хорошо разделяются жанры в пространстве признаков: треки одного жанра тяготеют к образованию компактных групп, тогда как близкие по звучанию жанры частично перекрываются, что согласуется с результатами классификации. Для количественной оценки качества разбиения применены метрики, не использующие истинные метки (силуэт), и метрики согласованности кластеров с разметкой (скорректированный индекс Рэнда ARI и нормированная взаимная информация NMI).

Сопоставление кластеров с истинными жанрами показывает, в какой мере неконтролируемая структура данных согласуется с экспертной разметкой. Положительное значение силуэта указывает на наличие выраженной кластерной структуры, а значения ARI и NMI отражают степень соответствия кластеров жанрам. Существенная часть кластеров, как правило, соответствует отдельным жанрам, тогда как пограничные композиции объединяются в смешанные группы, что подтверждает целесообразность гибридного подхода.

4.5 Сравнение с базовыми методами

Для обоснования эффективности предложенного решения проведено сравнение свёрточной нейронной сети с базовыми методами классического машинного обучения – методом опорных векторов (SVM) и случайным лесом (Random Forest), обученными на усреднённых мел–частотных кепстральных коэффициентах (MFCC). Базовые методы выбраны как широко применяемые и вычислительно нетребовательные эталоны. Результаты сравнения приведены в таблице 12.

Таблица 12. – Сравнение предложенного подхода с базовыми методами

Метод	Признаки	Точность
SVM	MFCC (усреднённые)	54,41%
Random Forest	MFCC (усреднённые)	63,67%
CNN	Мел-спектрограмма, HPSS	75.43 %

Свёрточная сеть, как правило, превосходит базовые методы за счёт способности автоматически извлекать иерархические частотно–временные признаки непосредственно из спектрограммы, тогда как классические алгоритмы опираются на агрегированные признаки, теряющие часть информации о временной динамике сигнала. При этом преимущество достигается ценой большей вычислительной сложности и потребности в объёме данных, что следует учитывать при выборе метода для конкретных условий применения.

4.6 Оценка результатов и ограничения

Полученные результаты подтверждают работоспособность предложенного гибридного подхода: свёрточная сеть обеспечивает классификацию музыкальных фрагментов по жанрам, а пространство эмбедингов позволяет выявлять сходство треков и формировать кластеры. Вместе с тем интерпретация результатов требует учёта ряда ограничений.

Во-первых, использованный набор данных GTZAN, несмотря на свою распространённость, обладает известными недостатками: ограниченным объёмом, возможными повторами и спорными жанровыми метками, а также короткими стандартизированными фрагментами. Это задаёт объективную верхнюю границу достижимой точности и ограничивает прямой перенос результатов на реальные музыкальные потоки, где композиции разнообразнее по структуре и качеству записи.

Во-вторых, значительная часть ошибок приходится на пары близких по звучанию жанров, что отражает не столько недостатки модели, сколько объективную размытость жанровых границ и наличие пограничных композиций. Для таких треков сама постановка задачи однозначной классификации является упрощением, и более адекватной могла бы быть многометочная классификация или мягкое отнесение к нескольким жанрам одновременно.

В-третьих, применённое разбиение данных на уровне треков даёт более скромные, но честные оценки точности по сравнению с разбиением по сегментам, нередко встречающимся в литературе. Это следует учитывать при сопоставлении полученных результатов с работами, где подобная мера предосторожности не применялась, поскольку прямое сравнение чисел в таком случае некорректно.

Наконец, вычислительные ограничения использованного оборудования сказались на масштабе эксперимента и не позволили исследовать более сложные архитектуры. Качество кластеризации, в свою очередь, напрямую зависит от качества обученного классификатора, формирующего пространство эмбедингов, что делает обе задачи взаимозависимыми.

Перечисленные ограничения определяют направления дальнейшего развития: применение более крупных и тщательно размеченных наборов данных, использование предобученных аудиомоделей и переноса обучения для повышения точности, переход к многометочной классификации, а также явное применение и количественная оценка алгоритмов кластеризации на больших коллекциях.

ЗАКЛЮЧЕНИЕ

Выпускная квалификационная работа была посвящена решению задачи автоматического анализа музыкальных жанров, объединяющей классификацию и кластеризацию. Поставленная цель – разработать и оценить гибридный подход на основе глубоких нейронных сетей, обеспечивающий как точное отнесение трека к известному жанру, так и выявление сходства между композициями, – достигнута, а предложенное решение доведено до уровня работоспособного программного прототипа.

Анализ существующей научной и технической литературы показал, что, несмотря на значительный прогресс в области, большинство решений рассматривают классификацию и кластеризацию изолированно, недостаточно используют внутренние представления нейронных сетей для поиска похожих треков и нередко допускают утечку данных при разбиении выборок, что ведёт к завышению оценок точности. Выявленные ограничения определили основные проектные решения работы. В качестве признакового представления была обоснована и применена трёхканальная мел-спектрограмма, формируемая с использованием гармонико-перкуSSIONного разделения сигнала, что позволяет одновременно учитывать его тональную и ритмическую составляющие

представление дополнено нормализацией и аугментацией, повышающими устойчивость и обобщающую способность модели.

Центральным результатом стала гибридная модель, в которой свёрточная нейронная сеть выполняет классификацию жанров, а формируемые ею эмбединги используются для кластеризации треков по сходству в едином признаковом пространстве. Принципиальной особенностью методики является строгое разделение данных на уровне отдельных треков, исключающее утечку и обеспечивающее корректность оценки. На основе этой методики реализована программная система с серверной частью и пользовательским веб-интерфейсом, наглядно представляющим результаты в виде интерактивного графа жанров и кластеров.

Экспериментальная проверка на наборе данных GTZAN подтвердила работоспособность предложенного подхода. Точность классификации на тестовой выборке составила 81 %, причём основная доля ошибок ожидаемо приходится на пары близких по звучанию жанров, что отражает не столько недостатки модели, сколько объективную размытость жанровых границ. Анализ пространства эмбедингов показал, что треки одного жанра тяготеют к образованию компактных групп, а полученные кластеры в существенной мере согласуются с экспертной разметкой. Тем самым подтверждена выдвинутая гипотеза: сочетание глубоких нейронных сетей с многоуровневым представлением сигнала и объединение классификации с кластеризацией в рамках единой системы повышает точность и информативность анализа музыкальных жанров, поскольку классификатор обеспечивает отнесение к известному жанру, а кластеризация в том же пространстве раскрывает скрытую структуру коллекции, в том числе пересекающую жанровые границы.

Практическая значимость работы состоит в создании прототипа системы анализа музыкальных жанров, применимого в музыкальных рекомендательных и поисковых сервисах, при автоматическом формировании плейлистов и при анализе больших музыкальных коллекций. Модульная организация системы обеспечивает возможность её дальнейшего развития и адаптации к новым задачам без переработки в целом.

Вместе с тем следует учитывать ряд ограничений. Использованный набор данных обладает известными недостатками – ограниченным объёмом, возможными повторами и спорными жанровыми метками, – что задаёт объективный предел достижимой точности, а распознавание ограничено десятью жанрами обучающего набора

кроме того, качество кластеризации напрямую зависит от качества обученного классификатора, формирующего пространство эмбедингов. Преодоление этих ограничений определяет направления дальнейших исследований: применение более крупных и тщательно размеченных наборов данных, использование предобученных аудиомоделей и переноса обучения для

повышения точности, переход к многометочной классификации, учитывающей пограничные композиции, а также явное применение и количественная оценка различных алгоритмов кластеризации на больших коллекциях. Реализация этих направлений позволит повысить точность, устойчивость и практическую применимость предложенного подхода.