УДК 004.934:004.8

СРАВНИТЕЛЬНЫЙ АНАЛИЗ XAI-ПОДХОДОВ ДЛЯ ОБНАРУЖЕНИЯ СИНТЕТИЧЕСКОЙ РЕЧИ

Ким К.К.

kamilla.k.kim@gmail.com

Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта,

100125, Узбекистан, г. Ташкент, м-в Буз-2, 17А

На сегодняшний день системы синтеза речи прогрессировали настолько, что сильно усложнили автоматическое выявление синтетической и дипфейк-речи. Большинство моделей-детекторов при этом остаются почти непрозрачными «чёрными ящиками». Из-за этого их сложно использовать в цифровой криминалистике, биометрии и других критически важных системах. В данной работе были сделаны обзор и сравнительный анализ методов объяснимого ИИ (XAI), которые применяются именно для этой задачи. Были рассмотрены градиентные, релевантностные и модель-независимые подходы: Integrated Gradients, Deep Taylor Decomposition, Layer-wise Relevance Propagation, а также LIME, SHAP и Grad-CAM. Основное внимание уделено принципам формирования объяснений и возможностям их интерпретации в контексте акустических характеристик речевого сигнала. Рассмотрены подходы к оценке качества объяснений, включая их достоверность, устойчивость и чувствительность к возмущениям, а также проанализированы ограничения опубликованных исследований. Проведённое сравнение позволило систематизировать сильные стороны и ограничения различных методов XAI, выявить проблемы сопоставимости существующих результатов и определить перспективные направления дальнейших исследований.

Ключевые слова: объяснимый искусственный интеллект, XAI, выявление синтетической речи, обнаружение аудио-дипфейков, интерпретируемость

Цитирование: *Ким К.К.* Сравнительный анализ XAI-подходов для обнаружения синтетической речи // Проблемы вычислительной и прикладной математики. – 2026. – № 5(76). – С. 72-94.

DOI: 10.71310/pcam.5_76.2026.06

1 Введение

Стремительное развитие нейросетевых технологий синтеза и преобразования речи привело к тому, что сейчас синтетическая и дипфейк-речь стала невероятно реалистичной, во многих случаях её практически невозможно отличить от настоящей на слух. Прогресс в text-to-speech (TTS) и voice conversion (VC) открыл массу применений: персонализированные голосовые ассистенты, клонирование голоса для видео и креатива, инструменты для изучения языков и повышения доступности [1].

Но при этом человеческое восприятие такой речи в качестве детектора работает довольно слабо. В одном из исследований средняя точность распознавания синтетической речи составила всего 73 % [2]. А в крупном эксперименте с более чем 1200 участниками люди пытались опираться на интуицию и лингвистические подсказки, но в сложных реалистичных условиях это помогало далеко не всегда. Более того,

даже современные модели машинного обучения не всегда уверенно превосходят человеческую интуицию. Это хорошо показывает, насколько на самом деле сложная задача: надёжно выявлять синтетическую речь [3].

В лабораторных условиях, особенно на стандартных датасетах вроде ASVspoof и FoR, современные детекторы синтетической речи демонстрируют очень неплохие результаты. Но как только их пытаются перенести в реальные и неконтролируемые сценарии, сразу появляются серьёзные проблемы: слабая обобщающая способность и почти полная непрозрачность решений. А это особенно опасно для ответственных приложений. К тому же огромное разнообразие условий записи, типов генеративных атак и самих моделей синтеза сильно подрывает их устойчивость в полевых условиях. Всё это ясно показывает, что нужны более глубокие исследования — не только для повышения точности классификации, но и для устойчивости к адверсариальным атакам и компрессии, а также для обеспечения интерпретируемости через методы объяснимого ИИ (XAI) [4].

В задачах обнаружения синтетической и дипфейк-речи современные нейросетевые детекторы обычно принимают на вход спектрограммы и другие временно-частотные представления. Даже при высоком уровне точности эти модели остаются практически «чёрными ящиками». Из-за этого очень сложно понять, как именно они принимают решения, и оценить их реальную надёжность и способность работать в полевых условиях. Именно поэтому методы объяснимого ИИ (XAI) сейчас рассматривают как важнейший инструмент повышения прозрачности. Методы XAI позволяют выявлять участки спектрограммы, вносящие вклад в решение модели, и анализировать их возможную связь с акустическими характеристиками речи — формантной структурой, распределением энергии и временной динамикой сигнала [5]. Кроме того, XAI хорошо помогает выявлять нежелательные или спуриозные корреляции, которые возникают из-за особенностей датасетов или артефактов предобработки. А это особенно важно для задач безопасности и голосовой биометрии [6].

Хотя методы XAI не избавляют полностью от проблемы «чёрного ящика» в глубоких моделях, их интеграция значительно повышает прозрачность, помогает лучше понимать логику решений и заметно увеличивает доверие к системам обнаружения синтетической речи. Благодаря этому такие детекторы становятся гораздо ближе к реальному практическому применению.

В данной работе представлен аналитический обзор, в котором дано сравнение и анализ самых распространённых методов XAI для выявления синтетической и дипфейк-речи. В отличие от большинства существующих обзоров по обнаружению синтетической речи, где объяснимость обычно разбирается либо в самом общем виде, либо применительно к другим модальностям данных [7–9], настоящая работа сознательно смещает акцент на специфику именно этой области и на реальную прикладную ценность получаемых объяснений.

В итоге удалось наглядно показать сильные и слабые стороны разных подходов, а также обозначить перспективные направления для дальнейших исследований. По сути, это сравнение — не просто обзор существующих решений, а инструмент, который помогает увидеть актуальные пробелы в области интерпретируемой детекции синтетической речи.

Вклад данного исследования

В отличие от обзоров, рассматривающих объяснимый искусственный интеллект в общем контексте или применительно к различным типам данных, настоящая рабо-

та сосредоточена на методах XAI для обнаружения синтетической и дипфейк-речи. Специфика этой задачи связана с временной и спектральной структурой речевого сигнала, а также с необходимостью соотносить объяснения модели с акустическими характеристиками речи. Основные вклады работы заключаются в следующем:

- Проведено структурированное сравнение градиентных методов, методов распространения релевантности и подходов, основанных на возмущении входных данных, с акцентом на принципы формирования объяснений и особенности применения к обнаружению аудиодипфейков.
- Систематизированы методы XAI по типу формируемых объяснений, уровню интерпретации, зависимости от архитектуры модели и вычислительным особенностям с учётом специфики аудиоданных.
- Проанализированы возможности интерпретации объяснений в терминах акустических характеристик речи, включая распределение энергии, временную динамику и частотные области, а также ограничения, связанные с подтверждением их акустической значимости.
- Обобщены опубликованные результаты оценки качества объяснений, включая их достоверность, устойчивость и чувствительность к возмущениям. Выявлены ограничения существующих исследований, связанные с различиями в данных, моделях и протоколах оценки, затрудняющими прямое количественное сопоставление методов.

На основе проведённого анализа сформулированы практические рекомендации по выбору методов XAI и определены направления дальнейших исследований, необходимые для повышения достоверности, устойчивости и практической полезности объяснений в системах обнаружения синтетической речи.

2 Объяснимый ИИ для выявления синтетической и дипфейк-речи

2.1 Обнаружение синтетической и дипфейк-речи

Аудио-дипфейки — это высокоубедительные искусственные записи, которые очень точно копируют человеческий голос и тонкие особенности естественной речи. В литературе выделяют несколько основных типов синтетической речи. Самые изученные:

- text-to-speech (TTS) — синтез речи из текста с акцентом на естественность;
- voice conversion (VC) — преобразование голоса одного говорящего в голос другого при сохранении смысла;
- replay-атаки — воспроизведение заранее записанной реальной речи для обмана биометрических систем.

Есть и более сложные варианты: частичная подделка с встраиванием синтетических сегментов в реальное аудио, а также атаки с изменением эмоций и стиля речи. Всё это заметно усложняет различение настоящих и поддельных записей по временно-частотным признакам. Поэтому создание надёжных методов детекции синтетической речи стало важной задачей для безопасности в биометрии, телекоммуникациях и медиа [1, 10].

Современные методы обнаружения дипфейк-аудио основываются на анализе временно-частотных представлений, таких как спектрограммы, мел-частотные кепстральные коэффициенты (MFCC) и другие спектральные признаки, которые позво-

ляют выявлять аномалии, характерные для синтетических сигналов [11]. Такие признаки затем подаются на вход машинным моделям классификации, включая сверточные нейронные сети ResNeXt и гибридные архитектуры CNN-LSTM, которые обучаются различать bona fide и fake аудио. Оценка качества детекторов, как правило, проводится на специализированных наборах данных, таких как ASVspoof и его различные подмножества, которые содержат реальные записи и различные типы атак (например, TTS, VC, replay) [12, 13].

Несмотря на заметный прогресс и улучшение основных метрик — EER и min-tDCF, у нынешних методов обнаружения синтетической речи по-прежнему серьёзные проблемы с обобщаемостью и устойчивостью в реальных сценариях. Особенно сильно они дают сбой, когда появляются шумы, разные записывающие устройства или новые виды генеративных атак. Методы, которые полагаются в основном на спектральные признаки и обучены на классических датасетах вроде ASVspoof, отлично работают в лабораторных условиях. Но стоит перейти к реальным записям — и качество резко падает. Разница между результатами на ASVspoof и на in-the-wild выборках просто бросается в глаза. Это явно говорит о том, что обобщающая способность у текущих подходов пока довольно ограниченная, они плохо переносятся за пределы «тепличных» данных [14]. Поэтому, одной только гонки за точностью классификации уже недостаточно. Всё острее встаёт вопрос интерпретируемости моделей — почему они принимают то или иное решение. Именно этому уделяется внимание в дальнейшей части статьи.

2.2 Архитектуры и принципы организации объяснимого ИИ

Explainable Artificial Intelligence (XAI), объяснимый искусственный интеллект — это область искусственного интеллекта, направленная на обеспечение понимания и интерпретируемости моделей машинного обучения и глубокого обучения, особенно в случаях «чёрного ящика», когда даже разработчики не могут явно описать механизм принятия решения модели. XAI позволяет пользователям и исследователям понимать что и почему влияет на результаты модели, повышая доверие, выявляя ошибки и снижая риски несправедливых или неправильных выводов [15].

Прежде чем переходить к рассмотрению отдельных методов XAI, необходимо определить общие принципы построения объясняемых систем и типовую последовательность формирования интерпретаций. Такая постановка позволяет понять, каким образом информация о работе модели преобразуется в объяснение и какие характеристики этого объяснения необходимо учитывать. Эти принципы применимы, в частности, к задачам интерпретации моделей, используемых для обнаружения синтетической речи и аудио-дипфейков.

На рис. 1 представлена обобщенная схема XAI-пайплайна, центральным элементом которой является модуль формирования объяснений.

На его вход могут поступать различные компоненты информации о работе модели: исходные данные, полученные предсказания и, в зависимости от архитектуры и используемого метода, внутренние представления или промежуточные признаки. На основе этой информации выбранный XAI-метод, например SHAP или LIME, формирует интерпретацию принятого моделью решения. Результат может быть представлен в различных формах — от числовых оценок вклада отдельных признаков и визуализаций до локальных аппроксимаций поведения модели. Полученные объяснения могут использоваться не только для анализа отдельных решений, но и для последующей оценки и совершенствования модели.



Рис. 1 Ключевой компонент системы объяснимого ИИ [16].

Представленная схема также позволяет выделить несколько ключевых характеристик XAI-подходов: область действия объяснения (локальная или глобальная), уровень представления интерпретации, а также степень зависимости метода от исходных данных, конкретной архитектуры модели и особенностей предметной области [16].

В контексте обнаружения синтетической речи описанный подход может быть реализован с учетом специфики аудиосигналов. На первом этапе детектор получает акустическое представление входного сигнала в виде набора признаков или временно-частотной матрицы. В зависимости от используемой архитектуры такими представлениями могут выступать LFCC- или MFCC-признаки, а также спектрограммы. После обработки входных данных нейросетевая модель формирует итоговое бинарное решение о принадлежности сигнала к естественной или синтетически сгенерированной речи. XAI-компонент дополняет это решение интерпретацией, показывающей, какие участки входного представления оказали наибольшее влияние на классификацию. В частности, объяснение может указывать на наиболее информативные временные фрагменты, частотные области или локальные спектральные особенности сигнала. Это позволяет перейти от простого определения класса к анализу акустических характеристик, на которых основывается решение детектора, и тем самым выявить признаки, связанные с синтетическим происхождением речи.

На рис. 2 показана итеративная организация XAI-процесса, в которой получение объяснения рассматривается не как завершающий этап работы модели, а как часть последовательного цикла анализа и совершенствования.

В рамках такой схемы различные методы объяснения (A-Z) применяются на отдельных этапах для исследования поведения модели и выявления особенностей, влияющих на принимаемые ею решения. Полученные интерпретации могут использоваться при последующей корректировке модели, ее компонентов или параметров.

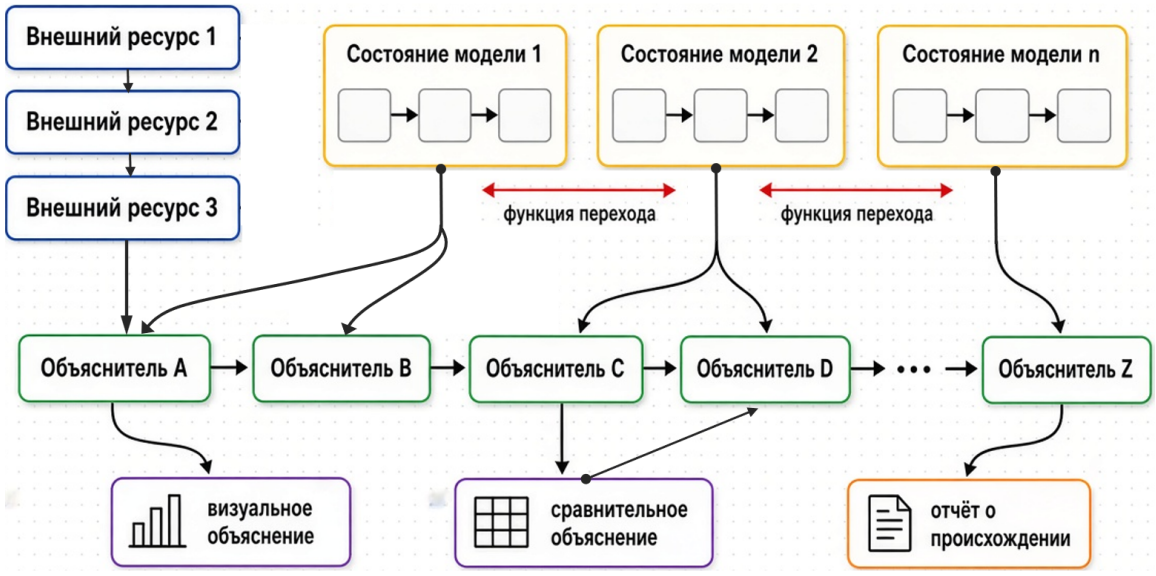


Рис. 2 Итеративный рабочий процесс XAI с использованием средств объяснения состояний модели и глобального мониторинга [16].

Помимо собственно процедур объяснения, схема предусматривает дополнительные механизмы контроля процесса, отслеживание происхождения и преобразования данных, оценку качества результатов и фиксацию изменений, внесенных в модель. Такой подход обеспечивает последовательное документирование процесса интерпретации и позволяет связывать результаты XAI-анализа с последующими этапами развития детектора [16].

В задачах детекции аудио-дипфейков такая итеративная схема отражает типичный цикл разработки и анализа: временно-частотное представление аудиосигнала поступает на гибридную модель (например, CNN-LSTM или ECAPA-TDNN), после чего XAI-методы, такие как Grad-CAM или IG, формируют тепловые карты значимости, локализующие спектральные и временные артефакты синтеза. Полученные объяснения помогают не только понять, почему модель приняла то или иное решение, но и значительно улучшить её: оптимизировать архитектуру, отобрать самые важные признаки и сделать систему устойчивее к новым генеративным атакам.

Современные детекторы синтетической и дипфейк-речи почти всегда строятся на глубоких нейронных сетях и показывают очень хорошие результаты на стандартных датасетах. Однако из-за сложности этих моделей и нелинейного способа обработки признаков зачастую остаётся загадкой, какие именно факторы на самом деле влияют на финальный вердикт. Вот здесь методы XAI и оказываются крайне полезными, они позволяют глубже разобраться в поведении модели. Это особенно важно для задач обнаружения аудио-дипфейков, где надёжность и доверие к результатам выходят на первый план.

В данной работе разбираются наиболее распространённые и показательные методы XAI, которые уже используются или могут быть адаптированы для анализа синтетической речи: IG, DTD, LRP, LIME, SHAP и Grad-CAM. Выбор пал именно на них, потому что они заметно отличаются по своим теоретическим основам, уровням интерпретации и тому, насколько часто их применяют в исследованиях по объяснимости глубоких моделей.

2.3 Методологическая основа для сравнительного анализа XAI

В работе использовался структурированный подход для сравнительного анализа методов объяснимого искусственного интеллекта (XAI), применяемых при обнаружении синтетической и дипфейк-речи. В отличие от традиционных систематических обзоров, задача состояла не в том, чтобы охватить максимально возможное число публикаций или провести библиометрический анализ. Основной интерес представляло сравнение самих XAI-подходов, их особенностей, ограничений и возможностей применения к речевым данным.

Для анализа были выбраны рецензируемые статьи из ведущих научных журналов и материалов конференций, индексируемых в Scopus и Web of Science преимущественно за период с 2020 по 2026 год. При этом внимание уделялось прежде всего качеству и содержанию исследований, а не количеству рассмотренных работ. Первоначальный поиск выявил 60 публикаций. После удаления 10 дубликатов и проверки названий и аннотаций для полнотекстового анализа было отобрано 50 работ. По результатам полнотекстовой проверки в итоговый корпус вошли 33 публикации, соответствующих установленным критериям включения. Критериями включения являлись:

– публикация в рецензируемом научном журнале или материалах научной конференции;
– непосредственная связь исследования с обнаружением синтетической речи, аудио-дипфейков или применением методов XAI к данным задачам; для отдельных обзорных работ — значимость представленного материала для анализа общих подходов к объяснимости моделей в смежных предметных областях;
– наличие описания используемого метода XAI и его применения или оценки.

Исключались работы без непосредственной связи с тематикой обзора. Отбор проводился на основании анализа названий, аннотаций и, при необходимости, полных текстов публикаций. После проверки соответствия критериям сформирован итоговый корпус работ, использованных для сравнительного анализа. Сравнение методов проводилось по нескольким критериям, среди которых принцип формирования объяснений, зависимость от типа модели, уровень интерпретации и вычислительные особенности.

Отдельно рассматривались характеристики, важные именно для задач обнаружения аудио-дипфейков. Анализ включал оценку того, насколько получаемые объяснения связаны с информативными акустическими признаками речи, сохраняют ли они устойчивость при изменении входных данных и как реагируют на шумы или спектральные артефакты. Это позволило оценить методы с учётом специфики рассматриваемой задачи, а не только с точки зрения общей интерпретируемости. Такой способ организации анализа помогает более наглядно увидеть сильные и слабые стороны существующих подходов и определить вопросы, которые всё ещё требуют дальнейшего изучения.

На рис. 3 показана предложенная схема сравнительного анализа XAI-методов для задач обнаружения синтетической и дипфейк-речи. Она отражает всю последовательность этапов: от определения области и выбора методов до их оценки и обобщения результатов.

Предложенный подход можно рассматривать как удобную отправную точку для сравнения методов объяснимого ИИ (XAI), применимых к задаче выявления синтетической и дипфейк-речи.



Рис. 3 Методологическая схема сравнительного анализа методов XAI для задач обнаружения синтетической и дипфейк-речи

2.4 Методы XAI для выявления синтетической речи

Градиентные методы атрибуции.

К первой группе относятся градиентные методы объяснимого ИИ, оценивающие вклад входных признаков в предсказание модели на основе аксиом чувствительности (ненулевые атрибуции для признаков, влияющих на предсказание) и инвариантности реализации (идентичность объяснений для функционально эквивалентных моделей).

Метод Integrated Gradients (IG) позволяет обойти ограничение, связанное с локально нулевыми градиентами в областях плоского ландшафта предсказаний, за счет интегрирования градиентов выходной функции f вдоль непрерывного пути от базовой точки x' (baseline) к исследуемому входу x:

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_0^1 \frac{\partial f\big(x' + \alpha(x - x')\big)}{\partial x_i}\, d\alpha, \tag{1}$$

где i — индекс входного признака, $\mathrm{IG}_i(x)$ — атрибуция вклада признака i в предсказание, $\alpha \in [0, 1]$ — параметр линейной интерполяции между x' и x, $\partial f / \partial x_i$ — чувствительность выхода модели к признаку [17, 18].

Вычисление интеграла по $\alpha \in [0, 1]$ заменяется его численной аппроксимацией в виде конечной суммы градиентов, вычисленных в последовательности промежуточных точек. Такой подход позволяет учитывать накопленный вклад градиента вдоль всего пути и снижает чувствительность атрибуций к локальным вариациям и проблеме насыщения градиентов, характерной для глубоких нелинейных моделей. В задачах анализа аудиосигналов IG может применяться к временно-частотным представлениям, в частности к спектрограммам, позволяя определить временные и частотные области, вносящие наибольший вклад в различение естественной (bona fide) и синтетической речи.

Grad-CAM (Gradient-Weighted Class Activation Mapping) формирует карты значимости на основе градиентов выходного сигнала целевого класса относительно активаций выбранного сверточного слоя. Полученные градиенты используются для

оценки вклада отдельных каналов активационных карт и последующего построения пространственно-временного распределения важности. Для целевого класса c и набора активационных карт A^k выбранного сверточного слоя вычисляются веса значимости:

$$L_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{2}$$

где y^c — выход модели для класса c, Z — нормирующий множитель, а градиенты отражают вклад соответствующих активаций в итоговое решение. Полученные веса используются для линейной комбинации карт признаков с последующим применением функции ReLU:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\Big(\sum_k L_c^k A^k\Big). \tag{3}$$

Операция ReLU применяется к полученной карте активации и позволяет выделить области с положительным вкладом в предсказание целевого класса, подавляя отрицательные значения. В результате формируется тепловая карта, локализующая участки входного представления, наиболее связанные с решением модели [19]. В задачах обнаружения синтетической речи и аудио-дипфейков Grad-CAM может применяться к временно-частотным представлениям, таким как спектрограммы и кепстральные карты, обеспечивая визуализацию временных и частотных областей, используемых детектором при классификации [21]. Это позволяет исследовать, связаны ли наиболее значимые области с локальными спектральными особенностями, временными несоответствиями или другими акустическими характеристиками, потенциально обусловленными процессом синтеза.

Методы распространения релевантности.

Данную группу составляют методы распространения релевантности, ориентированные на декомпозицию выходного решения модели.

Layer-wise Relevance Propagation (LRP) представляет собой метод XAI, который объясняет предсказание нейронной сети путём последовательного распространения релевантности от выходного слоя к входным признакам. Метод не предполагает единственного универсального правила перераспределения: конкретный механизм зависит от типа слоя и выбранной схемы распространения релевантности. В общем случае вклад нейронов последующего слоя распределяется между нейронами предыдущего слоя с учётом их активаций и весов связей. Повторяя эту процедуру от слоя к слою, можно оценить вклад отдельных входных признаков в итоговое решение модели [22].

Важным свойством LRP является сохранение суммарной релевантности при её распространении по слоям сети. В общем виде данное условие выражается равенством:

$$\sum_i R_i^{(0)} = \sum_j R_j^{(l)} = \sum_k R_k^{(l+1)} = \ldots = f(x). \tag{4}$$

Здесь $R_i^{(0)}$ — релевантность, приписанная i-му входному элементу, $R_j^{(l)}$ — релевантность, присвоенная j-му нейрону слоя l. Приведённое равенство выражает свойство сохранения релевантности, согласно которому её суммарная величина остаётся неизменной при распространении между слоями нейронной сети и соответствует выходному значению модели $f(x)$. Это равенство характеризует сохранение общей реле-

вантности, тогда как конкретный способ её распределения определяется выбранным правилом LRP. При выполнении соответствующих условий положительные значения релевантности указывают на признаки, поддерживающие предсказание, а отрицательные — на признаки, свидетельствующие против него [20, 21]. В задачах обнаружения синтетической речи LRP может применяться для исследования вклада временных и спектральных характеристик аудиосигнала в решение классификатора.

Deep Taylor Decomposition (DTD) представляет собой формализованное обобщение LRP, основанное на локальном разложении Тейлора выходной функции нейронной сети. DTD обеспечивает теоретически обоснованную декомпозицию решения модели и часто рассматривается как один из наиболее строгих методов анализа атрибуций. DTD лежит в основе теоретического обоснования LRP и моделирует релевантность нейрона k на слое l как простую функцию активаций нижележащего слоя, что особенно полезно для интерпретации моделей детекции синтетической речи в задачах XAI. В DTD релевантность выражается как:

$$R_k(a) = \max\Big(0, \sum_i a_i w_{ik}\Big) c_k, \tag{5}$$

где $a = [a_1 \dots a_d]$ — активации слоя $l-1$, w_{ik} — веса связи между нейронами i слоя $l-1$ и k слоя l, а c_k — константа (на выходе равна предсказанию модели). Разложение в ряд Тейлора вокруг корневой точки $\widetilde{a}$ даёт:

$$R_k(a) = \underbrace{R_k(\widetilde{a})}_{0} + \underbrace{\sum_i (a_i - \widetilde{a}_i) \cdot \big[\nabla R_k(a)\big]_i}_{\text{перераспределяемая релевантность}} + \underbrace{\varepsilon}_{0}, \tag{6}$$

где первый член исчезает по выбору $\widetilde{a}$, второй определяет правило перераспределения релевантности вниз по слоям [20, 22]. Данный подход позволяет анализировать вклад отдельных компонентов входного признакового представления в решение классификатора и локализовать области, наиболее значимые для обнаружения синтетической речи. При использовании MFCC или амплитудных спектрограмм получаемые объяснения могут указывать на спектральные аномалии и другие особенности признаков, связанные с решением модели.

Модель-независимые методы интерпретации моделей.

Отдельную категорию образуют модель-агностичные методы, которые не зависят от внутренней структуры классификатора.

LIME (Local Interpretable Model-agnostic Explanations) аппроксимирует поведение сложной модели в локальной окрестности конкретного примера с помощью интерпретируемого суррогатного классификатора. LIME генерирует возмущенные версии исходного аудиосэмпла x_i в его локальной окрестности N_{x_i} (например, маскируя частотные бины, временные фреймы или компоненты признакового вектора), пропускает их через целевую модель f_M и аппроксимирует поведение f_M простой интерпретируемой моделью f_S (обычно линейной регрессией):

$$f_M(x) \approx f_S(x), \qquad x \in N_{x_i}, \tag{7}$$

где коэффициенты f_S интерпретируются как вклады признаков в решение о синтетичности речи [21, 22]. Положительные и отрицательные веса указывают соответственно на вклад признаков в сторону объясняемого класса и противоположного класса.

SHAP (SHapley Additive exPlanations) представляет собой постфактум-метод интерпретации, основанный на концепции значений Шепли из теории кооперативных игр. В рамках данного подхода признаки рассматриваются как «игроки», совместно формирующие итоговое предсказание модели, а значение Шепли характеризует вклад каждого признака в это предсказание с учетом его взаимодействия с другими признаками. Для признака i его важность определяется на основе изменения выходной функции модели при включении данного признака в различные подмножества признаков S. Итоговый вклад $\varphi(i)$ рассчитывается как взвешенное усреднение его маржинального вклада по всем возможным подмножествам S:

$$\varphi(i) = \sum_{S \subseteq \{1,\dots,p\} \setminus \{i\}} \frac{|S|!\,(p-1-|S|)!}{p!} \big(e_{S\cup\{i\}} - e_S\big), \tag{8}$$

где p — общее количество признаков (размерность пространства признаков), S — произвольное подмножество признаков, не включающее признак i, e_S — условное предсказание модели, когда известны только признаки из S (остальные усреднены по фону).

В результате применения SHAP для каждого аудио-примера мы получаем вектор значений $\varphi_1, \varphi_2, \dots, \varphi_p$:

$$f(x^*) = e_\varnothing + \sum_{i=1}^{p} \varphi(i), \tag{9}$$

где $e_\varnothing$ — это базовое (среднее) предсказание модели по всей выборке (обычно близко к 0.5 при сбалансированных классах real/fake) [21, 22]. В задачах обнаружения синтетической и дипфейк-речи SHAP позволяет определить, какие акустические признаки вносят наибольший вклад в решение модели о принадлежности аудиофрагмента к естественной или синтетически сгенерированной речи [23]. Полученные значения атрибуции дают возможность исследовать акустические характеристики, на которые опирается детектор при классификации, а также сопоставить их с особенностями синтетического сигнала. Такой анализ может использоваться для выявления информативных признаков и определения направлений дальнейшего совершенствования методов обнаружения синтетической речи.

Таким образом, современные XAI-подходы в задачах выявления синтетической речи и аудио-дипфейков формируют многоуровневую систему интерпретации, объединяющую архитектурные принципы объясняемых моделей, итеративные рабочие схемы и специализированные методы атрибуции. Рассмотренные техники — от градиентных и разложенческих методов до модельно-независимых подходов — позволяют локализовать дискриминативные временно-частотные области, выявлять характерные артефакты синтеза и анализировать вклад акустических признаков в итоговое решение детектора. Данный теоретический базис создает основу для последующего сравнительного анализа XAI-методов, направленного на оценку их интерпретируемости, вычислительной эффективности и практической применимости в задачах детекции синтетической и дипфейк-речи.

3 Сравнительный анализ методов XAI для обнаружения синтезированной речи и аудио-дипфейков

3.1 Критерии оценки и параметры сравнения

В настоящем разделе проводится систематический сравнительный анализ методов объяснимого искусственного интеллекта, применяемых в задачах выявления синтетической и дипфейк-речи. Для обеспечения последовательности и сопоставимости результатов сравнение организовано по набору унифицированных критериев, отражающих как методологические особенности XAI-подходов, так и их интерпретируемость и практическую применимость в аудиодомене.

В качестве базовых измерений используются технические и методологические характеристики методов, включая тип объяснимости, уровень интерпретации, требования к модели, степень архитектурной зависимости и вычислительную сложность. Данные параметры позволяют оценить универсальность методов, их интегрируемость в различные архитектуры детекторов и масштабируемость при анализе аудиоданных. Эти характеристики обобщены в сравнительной таблице технических и методологических свойств XAI-подходов.

Дополнительно вводится группа критериев, ориентированных на интерпретируемость и доменно-специфические свойства объяснений в задачах синтетической речи. К ним относятся тип интерпретируемых элементов (временные фрагменты, частотные области, признаки или скрытые представления), семантическое соответствие акустически значимым характеристикам речи, устойчивость объяснений, чувствительность к шумам и пертурбациям, а также достоверность и согласованность с поведением модели. Эти аспекты позволяют оценить, в какой степени объяснения отражают реальные дискриминативные артефакты синтеза и пригодны для форензик-анализа и валидации детекторов.

Таким образом, совокупность технических, интерпретативных и количественных измерений формирует многоосевую рамку сравнительного анализа XAI-подходов, используемую далее для систематического сопоставления методов, выявления их преимуществ, ограничений и условий применимости в задачах детекции синтетической и дипфейк-речи.

3.2 Методологический анализ и оценка интерпретируемости

В данном подразделе проводится сравнительный анализ методов объяснимого искусственного интеллекта с точки зрения их методологических характеристик и интерпретируемости в задачах обнаружения синтетической и дипфейк-речи. Рассматриваются как технические особенности используемых XAI-подходов, так и характер формируемых ими объяснений, включая уровень их представления, зависимость от модели и применимость к аудиоданным. Отдельное внимание уделяется тому, насколько получаемые интерпретации позволяют связать решение детектора с конкретными временными, спектральными и акустическими характеристиками речевого сигнала.

Технические и методологические характеристики рассматриваемых XAI-подходов представлены в табл. 1. Методы существенно различаются по типу объяснимости, требованиям к модели и степени зависимости от архитектуры модели. Эти различия определяют возможности применения XAI-подходов к нейросетевым и признаковым детекторам синтетической речи и учитываются при последующем анализе их интерпретируемости.

IG определяет вклад входных признаков в предсказание нейросети путём интегрирования градиентов вдоль пути от базового входа к исследуемому примеру. Метод не требует изменения архитектуры, но нуждается в доступе к градиентам. Выбор базового значения влияет на объяснение и должен быть обоснован [17, 18, 20]. В задаче обнаружения синтетической речи IG применяют для анализа временных и частотных областей аудиосигнала или спектрограммы, связанных с решением о подлинности.

DTD разлагает выход нейросети на вклады отдельных признаков путём последовательного распространения релевантности. Метод позволяет выявить компоненты, поддерживающие решение классификатора, однако зависит от архитектуры сети и правил разложения [20, 22]. В анализе синтетической речи DTD локализует спектрально-временные компоненты, повлиявшие на классификацию, но карта релевантности сама по себе не доказывает их физическую или акустическую обоснованность.

LRP распределяет релевантность от выхода классификатора к входным признакам через слои сети с помощью специальных правил распространения [20, 22]. В отличие от IG, метод не опирается на интегрирование градиентов. В задачах аудиодипфейков LRP помогает определить участки спектрограммы, поддерживающие решение о подлинности или синтетичности речи. Интерпретация зависит от выбранных правил, а визуально выделенные области не всегда соответствуют причинно-значимым признакам.

LIME строит локальную интерпретируемую аппроксимацию модели, анализируя изменение предсказаний при возмущении входа. Главное преимущество — возможность объяснять модели без доступа к внутренней архитектуре [21, 22]. В обнаружении синтетической речи LIME выявляет временные интервалы или частотные компоненты, влияющие на решение детектора. Качество объяснений зависит от сегментации, характера возмущений и числа примеров, поэтому результаты следует проверять на устойчивость.

SHAP оценивает вклад признаков на основе значений Шепли. Метод позволяет анализировать как отдельные примеры, так и общие закономерности поведения классификатора. В работе Ge et al. [23] показано, что объяснения помогают исследовать спектрально-временные области, выявлять различия между моделями и случаи опоры на участки без речи или определённые частотные диапазоны. Вычислительная сложность и зависимость от модели определяются выбранной реализацией.

Grad-CAM формирует карту локализации, взвешивая активации выбранного слоя градиентами целевого класса. Метод визуализирует области, связанные с решением, но требует подходящего слоя с картами активаций и доступа к градиентам [19, 21]. В анализе синтетической речи его применяют к спектрограммам для исследования значимых временных и частотных областей. Разрешение карты определяется выбранным слоем, а визуальная выразительность не гарантирует точности или причинной обоснованности объяснения.

Технические характеристики методов, рассмотренные в табл. 1, определяют способы формирования объяснений, однако сами по себе не характеризуют их содержательную интерпретируемость для аудиоданных. В табл. 2 методы сопоставляются с точки зрения того, какие элементы аудиопредставления могут быть проанализированы с их помощью, насколько получаемые объяснения допускают связь с акустическими характеристиками речевого сигнала и какие ограничения необходимо учитывать при такой интерпретации.

Таблица 1 Технические и методологические характеристики методов XAI

Метод XAI	Ключевой принцип	Тип объяснимости	Требования к модели	Зависимость от модели	Вычислительные затраты
IG [17, 20]	Интегрирование градиентов вдоль пути от базового входа к анализируемому	Градиентная атрибуция признаков	Дифференцируемая модель с доступом к градиентам	Высокая: необходим доступ к градиентам	Умеренные, чаще высокие: зависят от числа шагов интегрирования
DTD [20, 22]	Разложение выходного сигнала на вклады признаков на основе ряда Тейлора	Разложение релевантности	Нейронная сеть с поддерживаемыми правилами разложения	Высокая: зависит от архитектуры и правил распространения	Умеренные: зависят от числа слоёв, правил и реализации
LRP [20, 22]	Послойное распределение релевантности от выхода к входу	Распространение релевантности	Нейронная сеть с поддерживаемыми правилами распространения	Высокая: необходим доступ к внутренним вычислениям	Обычно умеренные; зависят от архитектуры и правил
LIME [21, 22]	Локальная аппроксимация предсказаний интерпретируемой моделью на основе возмущений входа	Модель-независимые	Доступ к функции предсказания, возможность формировать возмущённые входы	Низкая: не требуется доступ к внутренней архитектуре	Зависят от количества возмущений и стоимости одного предсказания
SHAP [23]	Оценка вклада признаков на основе значений Шепли из теории кооперативных игр	Аддитивная атрибуция признаков	Зависят от алгоритма: доступ к предсказаниям либо к поддерживаемым внутренним операциям	От низкой до высокой в зависимости от реализации	Могут быть высокими: зависят от алгоритма, числа признаков и фоновой выборки
Grad-CAM [19, 21]	Взвешивание карт активаций градиентами целевого класса	Градиентная локализация признаков	Доступ к градиентам и подходящему слою с картами активаций	Высокая: зависит от архитектуры и выбранного слоя	Обычно умеренные: дополнительные затраты на градиенты и построение карты

Анализ опубликованных исследований показывает, что интерпретируемость XAI-методов при обнаружении синтетической речи в значительной степени определяется используемым представлением аудиосигнала и архитектурой детектора. Для градиентных и релевантностных методов это особенно заметно в работе Lim et al. [20], где DTD, IG и LRP применялись к CNN- и CNN-LSTM-моделям на наборах данных ASVspoof 2019 и LJSpeech. Для CNN-модели все три метода выделяли области, соответствующие формантам F1-F3, тогда как при добавлении LSTM характер распределения атрибуций изменялся: DTD и LRP в меньшей степени зависели от формантной структуры, а карты LRP становились менее распознаваемыми с точки зрения человеческого восприятия. Это показывает, что связь атрибуций с акусти-

Таблица 2 Особенности интерпретации решений детекторов синтетической речи с помощью методов XAI

Метод XAI	Интерпретируемые элементы	Связь с акустическими характеристиками	Локализация признаков синтетической речи
IG [17, 20]	Временные и частотные области входного представления	Позволяет сопоставлять атрибуции с распределением энергии и спектральными структурами	Зависит от метода оценки и исследуемых данных, количественная валидация локализации ограничена
DTD [21, 22]	Области входного представления, имеющие релевантность для решения	Может использоваться для анализа спектральных и временных особенностей, связанных с классификацией	Возможна, но качество интерпретации зависит от получаемых карт релевантности
LRP [20]	Входные признаки и промежуточные представления	Позволяет анализировать распределение вклада по временно-частотным компонентам	Ограниченная интерпретируемость для аудио
LIME [21, 22]	Выбранные признаки или сегменты входных данных	Связь определяется тем, какие акустические признаки или сегменты используются при построении объяснения	Зависит от способа сегментации и представления аудио
SHAP [21, 23]	Отдельные признаки или группы признаков	Может связывать вклад признаков с акустическими характеристиками при соответствующем представлении данных	Зависит от определения признаков и способа их группировки
Grad-CAM [19, 24]	Области внутренних свёрточных представлений	Позволяет визуально сопоставлять значимые области с временно-частотной структурой спектрограммы	Обеспечивает визуализацию значимых областей; точность локализации требует отдельной количественной проверки

ческими характеристиками не является неизменным свойством метода и зависит от архитектуры анализируемой модели [20].

Для LIME интерпретация также зависит от способа представления и возмущения входных данных: изменение отдельных сегментов или признаков позволяет оценивать их влияние на локальное решение модели, однако результат определяется выбранной схемой сегментации и процедурой формирования возмущений. Поэтому получаемые атрибуции целесообразно рассматривать как локальное приближение поведения конкретной модели, а не как прямое отображение акустически значимых свойств речевого сигнала [21, 22].

В исследованиях использования SHAP имеются непосредственные свидетельства его применения к задачам обнаружения речевого спуфинга и синтетической речи. Ge et al. [23] использовали SHAP для анализа решений детекторов, работающих с речевыми сигналами и спектрограммными представлениями, и показали возможность выявления артефактов, характерных для отдельных атак, а также различий между синтетической речью и voice conversion. Таким образом, SHAP может использоваться не только для оценки вклада отдельных признаков, но и для анализа закономерностей, связанных с конкретными типами атак [21, 22].

Применение Grad-CAM также позволяет связывать решения сверточных детекторов с отдельными областями спектрального представления. В работе Kwak et al. [24] Grad-CAM использовался для визуализации областей спектрограммы, оказывающих существенное влияние на обнаружение голосового спуфинга. При этом получаемая локализация должна рассматриваться прежде всего как средство визуального анализа поведения модели, а не как самостоятельное доказательство того, что выделенная область представляет причинно-значимый акустический признак синтетической речи [19, 24].

В целом, рассмотренные исследования подтверждают возможность использования XAI для локализации и анализа признаков, связанных с решениями детекторов синтетической речи, однако не позволяют установить универсальное преимущество одного метода над другими. Характер получаемых объяснений зависит от архитектуры детектора, представления аудиосигнала и процедуры формирования атрибуций. Кроме того, визуальная или признаковая интерпретация сама по себе не гарантирует достоверность объяснения, его устойчивость или соответствие реальным акустическим механизмам синтетической речи. Поэтому данные характеристики требуют отдельной эмпирической оценки с использованием сопоставимых критериев.

3.3 Надёжность и практическая применимость объяснений XAI

Оценка надёжности объяснений XAI-методов в задаче обнаружения синтетической речи остаётся недостаточно разработанной. Большинство существующих исследований сосредоточено на демонстрации самих карт атрибуции и качественном описании выделяемых областей, в то время как вопросы устойчивости объяснений, их связи с акустически осмысленными признаками синтеза и практической применимости в реальных условиях изучены фрагментарно. В связи с неоднородностью экспериментальных постановок, архитектур детекторов и используемых датасетов проведение прямого количественного сравнения методов в настоящее время невозможно. Ниже обобщены эмпирические наблюдения, доступные в рассмотренных работах, и сформулированы ограничения, которые определяют текущий уровень практической применимости XAI в данной области. В табл. 3 обобщены эмпирические наблюдения, доступные в рассмотренных работах, и сформулированы ограничения, которые определяют текущий уровень практической применимости XAI в данной области.

Анализ таблицы показывает, что наиболее проработанным аспектом на данный момент является зависимость характера объяснений от архитектуры детектора. Результаты свидетельствуют, что даже при использовании одних и тех же XAI-методов (IG, DTD, LRP) переход от простой свёрточной сети к архитектуре с LSTM существенно меняет распределение атрибуций. Это указывает на ограниченную переносимость интерпретаций между различными детекторами.

В части связи с акустическими признаками имеющиеся данные позволяют говорить о способности методов выделять формантные области и спектрально-временные зоны, используемые классификатором [20, 23]. Однако сама по себе атрибуция не устанавливает причинно-следственную связь этих зон с артефактами синтеза. Вопрос о том, отражают ли выделенные области действительно информативные признаки синтетической речи или побочные корреляции в данных, остаётся открытым.

Наиболее существенным пробелом является отсутствие исследований устойчивости объяснений к шуму, компрессии и изменению условий записи. Без таких данных невозможно оценить пригодность XAI-методов для реальных аудиофорензик-сценариев. Аналогично, количественные метрики надёжности объяснений в рассмот-

Таблица 3 Эмпирические данные и ограничения оценки надёжности объяснений XAI

Критерий оценки	Результаты исследований	Аналитическая интерпретация	Ограничения имеющихся данных
Устойчивость к шуму и изменению условий записи	В рассмотренных работах нет достаточной сопоставимой оценки объяснений при контролируемом добавлении шума, изменении SNR	Устойчивость объяснений остаётся открытым вопросом для дальнейших исследований	Нельзя ранжировать методы по шумоустойчивости на основании имеющихся результатов
Практическая локализация значимых областей	Методы IG, DTD, LRP [20] и Grad-CAM [20] применяются для выделения значимых участков спектрограммы	Карты локализации помогают анализировать, на какие области опирается детектор	Количественная оценка точности локализации в рассмотренных работах практически отсутствует
Количественная оценка надёжности	Работы [20, 23, 24] демонстрируют применение XAI для анализа детекторов, но используют разные модели и экспериментальные постановки.	Имеющиеся исследования полезны для качественного анализа и постановки дальнейших экспериментов	Нет единой сопоставимой оценки достоверности, стабильности карт атрибуций и их устойчивости к акустическим искажениям

ренных работах практически не применяются, что не позволяет провести обоснованное ранжирование методов.

С практической точки зрения наиболее перспективными выглядят подходы, сочетающие относительно стабильную локализацию акустически осмысленных областей с приемлемой вычислительной сложностью. Вместе с тем даже для этих методов требуется дополнительная экспериментальная проверка устойчивости объяснений в условиях, приближенных к реальным.

Проведённый сравнительный анализ показал, что существующие XAI-методы позволяют получать качественные представления о том, на какие спектрально-временные области опираются детекторы синтетической речи. Вместе с тем уровень их надёжности и практической применимости в настоящее время остаётся ограниченным. Основные ограничения связаны с сильной зависимостью объяснений от архитектуры модели, отсутствием доказанной связи атрибуций с причинными признаками синтеза, а также практически полным отсутствием данных об устойчивости объяснений к акустическим искажениям.

4 Обсуждение результатов сравнительного анализа

4.1 Ключевые выводы и методологические аспекты

Сравнительный анализ показывает, что методы объяснимого искусственного интеллекта обладают взаимодополняющими возможностями при исследовании решений детекторов синтетической речи. Они позволяют оценивать вклад отдельных признаков, выявлять значимые спектральные и временные области и анализировать закономерности, используемые классификаторами. При этом характер объяснений зависит от архитектуры модели, способа представления аудиосигнала и выбранного метода XAI.

Модельно-агностические подходы, включая LIME и отдельные реализации SHAP, позволяют анализировать предсказания без обязательного доступа к внутренней

структуре детектора, однако вычислительные затраты зависят от алгоритма и количества оцениваемых возмущений. Методы, использующие градиенты, распространение значимости или активации нейронной сети, предоставляют другие возможности анализа, но также ограничены особенностями конкретной архитектуры.

Несмотря на потенциал XAI для исследования акустических признаков синтетической речи, визуализация значимых областей сама по себе не подтверждает их специфичность для синтеза или достоверность объяснения решения модели. Кроме того, рассмотренные исследования не обеспечивают достаточной сопоставимой количественной оценки устойчивости объяснений к шуму и изменению условий записи.

Таким образом, имеющихся данных недостаточно для универсального ранжирования рассмотренных методов по надёжности и практической эффективности. Дальнейшие исследования целесообразно направить на разработку единого протокола оценки, учитывающего соответствие объяснений поведению модели, акустическую обоснованность выделенных признаков, стабильность при изменении входных данных и вычислительные затраты.

4.2 Ограничения современных XAI-подходов в задачах обнаружения аудио-дипфейков

Несмотря на возможности XAI при анализе решений детекторов синтетической речи, существующие подходы характеризуются рядом существенных ограничений.

Во-первых, отсутствует общепринятый эталон интерпретаций. В отличие от компьютерного зрения, где локализацию можно сопоставить с разметкой объектов, в аудиодомене нет достоверной эталонной разметки дискриминативных артефактов синтеза [25]. Это затрудняет объективную оценку корректности и полноты объяснений.

Во-вторых, большинство методов обладает выраженной архитектурной зависимостью. Градиентные и релевантностные подходы чувствительны к структуре сети, функциям активации и глубине модели. В результате характер атрибуций может существенно меняться при переходе от одной архитектуры детектора к другой, что снижает переносимость и воспроизводимость интерпретаций.

В-третьих, недостаточно изучена устойчивость объяснений к реальным акустическим условиям. Шум, изменение отношения сигнал/шум, сжатие и искажения записи способны влиять как на предсказания детектора, так и на формируемые attribution-карты. Рассмотренные исследования не предоставляют сопоставимой оценки устойчивости XAI-методов в таких условиях.

В-четвёртых, отсутствует унифицированная система оценки надёжности объяснений. Различия в моделях, датасетах и протоколах экспериментов затрудняют сравнение результатов. Качество классификации (например, EER) не является прямой мерой достоверности объяснения. Для более объективной оценки необходимо учитывать соответствие объяснения поведению модели, достоверность, стабильность атрибуций и вычислительные затраты.

Указанные ограничения снижают практическую применимость существующих XAI-подходов в задачах аудиофорензики и определяют направления дальнейших исследований.

4.3 Открытые проблемы и перспективные направления исследований

Результаты обзора выявляют ряд ключевых открытых задач, определяющих направления дальнейших исследований в области объяснимой детекции аудиодипфейков. Одной из центральных проблем остается разработка стандартизированных про-

токолов оценки интерпретируемости, включающих формализованные метрики стабильности, достоверности и локализационной точности объяснений. Формирование таких протоколов является необходимым условием для сопоставимого и воспроизводимого анализа XAI-методов.

Важным направлением является исследование устойчивости объяснений к реальным акустическим условиям. Существующие работы почти не оценивают поведение карт атрибуций при добавлении шума, изменении отношения сигнал/шум, компрессии и других искажениях, характерных для практических сценариев. Разработка робастных XAI-методов, сохраняющих интерпретируемость в таких условиях, представляется необходимой для применения в аудиофорензике.

Перспективным направлением является развитие временно-ориентированных XAI-подходов, способных учитывать динамическую структуру речи и выявлять последовательные паттерны синтеза [26]. Интеграция механизмов временного внимания и последовательных моделей интерпретации может существенно повысить чувствительность к долгосрочным аномалиям и переходным эффектам, характерным для нейросинтеза [27].

Остаётся актуальной проблема связи выделяемых областей с акустически осмысленными признаками. Атрибуция сама по себе не доказывает, что локализованные зоны соответствуют причинным артефактам синтеза, а не побочным закономерностям данных. В условиях активного противодействия со стороны генеративных моделей надёжность объяснений приобретает особое значение для анализа аудио-дипфейков и судебной экспертизы [28, 29]. Интеграция человеко-ориентированных подходов [30], учитывающих перцептивные и когнитивные особенности восприятия речи, является приоритетным вектором исследований. Такие подходы открывают возможность создания интерпретаций, пригодных не только для исследователей, но и для экспертов-лингвистов и аналитиков безопасности [30–33].

5 Заключение

В данной работе представлен систематизированный сравнительный анализ методов объяснимого искусственного интеллекта, применяемых в задачах выявления синтетической и дипфейк-речи. На основе унифицированной методологической рамки выполнено сопоставление архитектурных принципов, интерпретируемости, вычислительных характеристик и практической применимости ключевых XAI-подходов в аудиодомене.

Для применения методов объяснимого искусственного интеллекта в аудиофорензических сценариях существенное значение имеют возможность локализации дискриминативных областей во временно-частотном представлении аудиосигнала и согласованность получаемых объяснений с акустическими характеристиками речи. Вместе с тем имеющиеся эмпирические данные не позволяют установить универсальное преимущество отдельных методов XAI или однозначно ранжировать их по эффективности. Поэтому выбор метода должен учитывать особенности анализируемой модели, характер решаемой задачи и требования к достоверности интерпретации. Вместе с тем анализ выявил ряд существенных ограничений современных методов XAI, включая зависимость объяснений от архитектуры модели, нестабильность результатов интерпретации и отсутствие стандартизированных протоколов оценки их качества.

Представленный анализ формирует целостное представление о текущем состоянии области и определяет ключевые направления дальнейших исследований, включая разработку робастных временно-ориентированных XAI-методов и формализа-

цию критериев надежности объяснений. Полученные выводы могут служить ориентиром для разработки прозрачных и надежных систем детекции синтетической и дипфейк-речи в прикладных системах.

Декларации

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Доступность данных. Работа является обзорной; новые данные не создавались.

Литература

[1] *Zhang B., Cui H., Nguyen V. [et al.]* Audio deepfake detection: what has been achieved and what lies ahead // Sensors. — 2025. — Vol. 25, №7. — Art. 1989. — doi: https://doi.org/10.3390/s25071989.

[2] *Mai K. T., Bray S., Davies T. [et al.]* Warning: humans cannot reliably detect speech deepfakes // PLoS ONE. — 2023. — Vol. 18, №8. — Art. e0285333. — doi: https://doi.org/10.1371/journal.pone.0285333.

[3] *Warren K., Tucker T., Crowder A. [et al.]* «Better be computer or I'm dumb»: a large-scale evaluation of humans as audio deepfake detectors // Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24). — 2024. — New York: ACM. — P. 2696–2710. — doi: https://doi.org/10.1145/3658644.3670325.

[4] *Wang X., Delgado H., Evans N. [et al.]* ASVspoof 5: evaluation of spoofing, deepfake, and adversarial attack detection using crowdsourced speech // IEEE Transactions on Audio, Speech and Language Processing. — 2026. — Vol. 34. — P. 2354–2367. — doi: https://doi.org/10.1109/TASLPRO.2026.3682962.

[5] *Momin M. S., Sufian A., Barman D. [et al.]* Explainable deepfake detection across different modalities: an overview of methods and challenges // Image and Vision Computing. — 2025. — Vol. 163. — Art. 105738. — doi: https://doi.org/10.1016/j.imavis.2025.105738.

[6] *Bisogni C., Loia V., Nappi M. [et al.]* Acoustic features analysis for explainable machine learning-based audio spoofing detection // Computer Vision and Image Understanding. — 2024. — Vol. 249. — Art. 104145. — doi: https://doi.org/10.1016/j.cviu.2024.104145.

[7] *Schwalbe G., Finzel B.* A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts // Data Mining and Knowledge Discovery. — 2024. — Vol. 38. — P. 3043–3101. — doi: https://doi.org/10.1007/s10618-022-00867-8.

[8] *Yang W., Wei Y., Wei H. [et al.]* Survey on explainable AI: from approaches, limitations and applications aspects // Human-Centric Intelligent Systems. — 2023. — Vol. 3. — P. 161–188. — doi: https://doi.org/10.1007/s44230-023-00038-y.

[9] *Budhkar A., Song Q., Su J., Zhang X.* Demystifying the black box: a survey on explainable artificial intelligence (XAI) in bioinformatics // Computational and Structural Biotechnology Journal. — 2025. — Vol. 27. — P. 346–359. — doi: https://doi.org/10.1016/j.csbj.2024.12.011.

[10] *Salih A. O. M., Emam H. M., Ahmed A. B. G. E. [et al.]* Deepfake audio detection in voice authentication: a spectral and CNN-based comprehensive review // Engineering, Technology & Applied Science Research. — 2025. — Vol. 15, №6. — P. 29824–29832. — doi: https://doi.org/10.48084/etasr.13400.

[11] *Almutairi Z., Elgibreen H.* A review of modern audio deepfake detection methods: challenges and future directions // Algorithms. — 2022. — Vol. 15, №5. — Art. 155. — doi: https://doi.org/10.3390/a15050155.

[12] *Tahaoglu G., Baracchi D., Shullani D. [et al.]* Deepfake audio detection with spectral features and ResNeXt-based architecture // Knowledge-Based Systems. — 2025. — Vol. 323. — Art. 113726. — doi: https://doi.org/10.1016/j.knosys.2025.113726.

[13] *Asuai C., Arinomor A. P., Atumah C. T. [et al.]* Hybrid CNN-LSTM architectures for deepfake audio detection using Mel frequency cepstral coefficients and spectrogram analysis // American Journal of Mathematical and Computer Modelling. — 2025. — Vol. 10, № 3. — P. 98–109. — doi: https://doi.org/10.11648/j.ajmcm.20251003.12.

[14] *Firc A., Malinka K., Hanacek P.* Evaluation framework for deepfake speech detection: a comparative study of state-of-the-art deepfake speech detectors // Cybersecurity. — 2025. — Vol. 8. — Art. 1. — doi: https://doi.org/10.1186/s42400-024-00346-1.

[15] *Mersha M., Lam K., Wood J. [et al.]* Explainable artificial intelligence: a survey of needs, techniques, applications, and future direction // Neurocomputing. — 2024. — Vol. 599. — Art. 128111. — doi: https://doi.org/10.1016/j.neucom.2024.128111.

[16] *Spinner T., Schlegel U., Schäfer H. [et al.]* explAIner: a visual analytics framework for interactive and explainable machine learning // IEEE Transactions on Visualization and Computer Graphics. — 2020. — Vol. 26, № 1. — P. 1064–1074. — doi: https://doi.org/10.1109/TVCG.2019.2934629.

[17] *Naresh B., Ch R., Reddy K. S. [et al.]* Enhancing the detection of deepfake audio using explainable AI techniques for secure and trustworthy applications // 2025 International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA). — Aurangabad, 2025. — P. 1–6. — doi: https://doi.org/10.1109/ICAMIDA64673.2025.11209741.

[18] *Sundararajan M., Taly A., Yan Q.* Axiomatic attribution for deep networks // Proceedings of the 34th International Conference on Machine Learning (PMLR). — 2017. — Vol. 70. — P. 3319–3328.

[19] *Cheng Z., Wu Y., Li Y. [et al.]* A comprehensive review of explainable artificial intelligence (XAI) in computer vision // Sensors. — 2025. — Vol. 25, № 13. — Art. 4166. — doi: https://doi.org/10.3390/s25134166.

[20] *Lim S.-Y., Chae D.-K., Lee S.-C.* Detecting deepfake voice using explainable deep learning techniques // Applied Sciences. — 2022. — Vol. 12, № 8. — Art. 3926. — doi: https://doi.org/10.3390/app12083926.

[21] *Govindu A., Kale P., Hullur A. [et al.]* Deepfake audio detection and justification with explainable artificial intelligence (XAI). — Preprint (Version 1). — Research Square, 2023. — doi: https://doi.org/10.21203/rs.3.rs-3444277/v1.

[22] *Holzinger A., Saranti A., Molnar C. [et al.]* Explainable AI methods — a brief overview // xxAI — Beyond Explainable AI. Lecture Notes in Computer Science. — 2022. — Vol. 13200. — P. 13–38. — doi: https://doi.org/10.1007/978-3-031-04083-2_2.

[23] *Ge W., Patino J., Todisco M. [et al.]* Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations // 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — 2022. — P. 6387–6391. — doi: https://doi.org/10.1109/ICASSP43922.2022.9747476.

[24] *Kwak I.-Y., Kwak S., Lee J. [et al.]* Voice spoofing detection through residual network, max feature map, and depthwise separable convolution // IEEE Access. — 2023. — Vol. 11. — P. 49140–49152. — doi: https://doi.org/10.1109/ACCESS.2023.3275790.

[25] *Akman A., Schuller B. W.* Audio explainable artificial intelligence: a review // Intelligent Computing. — 2024. — Vol. 3. — Art. 0074. — doi: https://doi.org/10.34133/icomputing.0074.

[26] *Grinberg P., Kumar A., Koppisetti S. [et al.]* What does an audio deepfake detector focus on? A study in the time domain // ICASSP 2025 — 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. — 2025. — Hyderabad. — P. 1–5. — doi: https://doi.org/10.1109/ICASSP49660.2025.10887568.

[27] *Cheng Z., Yang H., Xiong Y. [et al.]* Explainable AI for forensic speech authentication within cognitive and computational neuroscience // Frontiers in Neuroscience. — 2025. — Vol. 19. — Art. 1692122. — doi: https://doi.org/10.3389/fnins.2025.1692122.

[28] *Baniecki H., Biecek P.* Adversarial attacks and defenses in explainable artificial intelligence: a survey // Information Fusion. — 2024. — Vol. 107. — Art. 102303. — doi: https://doi.org/10.1016/j.inffus.2024.102303.

[29] *Alvarez-Melis D., Jaakkola T. S.* Towards robust interpretability with self-explaining neural networks // Advances in Neural Information Processing Systems (NeurIPS 2018). — 2018. — Vol. 31. — P. 7786–7795.

[30] *Ridley M.* Human-centered explainable artificial intelligence: an Annual Review of Information Science and Technology (ARIST) paper // Journal of the Association for Information Science and Technology. — 2025. — Vol. 76, №1. — P. 98–120. — doi: https://doi.org/10.1002/asi.24889.

[31] *Kim J., Maathuis H., Sent D.* Human-centered evaluation of explainable AI applications: a systematic review // Frontiers in Artificial Intelligence. — 2024. — Vol. 7. — Art. 1456486. — doi: https://doi.org/10.3389/frai.2024.1456486.

[32] *Yang T., Sun C., Lyu S. [et al.]* Forensic deepfake audio detection using segmental speech features // Forensic Science International. — 2026. — Vol. 379. — Art. 112768. — doi: https://doi.org/10.1016/j.forsciint.2025.112768.

[33] *Padín-Torrente H., Carneiro-Diaz V., Ortega-Fernandez I.* Toward human-centered explainability: natural language explanations for anomaly detection // Information Systems Frontiers. — 2026. — doi: https://doi.org/10.1007/s10796-026-10717-3.

UDC 004.934:004.8

COMPARATIVE ANALYSIS OF XAI APPROACHES FOR SYNTHETIC SPEECH DETECTION

Kim K.K.

kamilla.k.kim@gmail.com

Research Institute for the Development of Digital Technologies and Artificial Intelligence, 17A, Buz-2, Tashkent, 100125 Uzbekistan

To date, speech synthesis systems have advanced to the point where the automatic detection of synthetic and deepfake speech has become increasingly challenging. At the same time, most detection models remain largely opaque "black boxes", which limits their applicability in digital forensics, biometrics, and other critical systems. This study reviews and comparatively analyzes explainable artificial intelligence (XAI) methods applied specifically to synthetic and deepfake speech detection. Gradient-based, relevance-based, and model-agnostic approaches are examined, including Integrated Gradients, Deep Taylor Decomposition, Layer-wise Relevance Propagation, LIME, SHAP, and Grad-CAM. Particular attention is paid to the principles underlying explanation generation and the interpretability of explanations in relation to the acoustic characteristics of speech signals. Approaches to evaluating explanation quality are discussed, including faithfulness,

stability, and sensitivity to perturbations, and the limitations of published studies are analyzed. The comparison provides a structured overview of the strengths and limitations of different XAI methods, identifies challenges in comparing existing results, and outlines promising directions for future research.

Keywords: explainable artificial intelligence, XAI, synthetic speech detection, audio deepfake detection, interpretability

Citation: Kim K.K. 2026. Comparative analysis of XAI approaches for synthetic speech detection. *Problems of Computational and Applied Mathematics.* 5(76): 72-94.

DOI: 10.71310/pcam.5_76.2026.06

№ 5(76) 2026

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

AiRI
RAQAMLI TEXNOLOGIYALAR VA SUN'IY INTELLEKTNI RIVOJLANTIRISH ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 5(76) 2026

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Главный редактор:
Равшанов Н.

Заместители главного редактора:
Арипов М.М., Шадиметов Х.М., Ахмедов Д.Д.

Ответственный секретарь:
Убайдуллаев М.Ш.

Редакционный совет:
Азамов А.А., Алоев Р.Д., Амиргалиев Е.Н. (Казахстан), Арушанов М.Л., Абдуллозода С.Ф. (Таджикистан), Боборахимов Б.И., Бурнашев В.Ф., Джумаёзов У.З., Загребина С.А. (Россия), Задорин А.И. (Россия), Игнатьев Н.А., Ильин В.П. (Россия), Иманкулов Т.С. (Казахстан), Исмагилов И.И. (Россия), Кабанихин С.И. (Россия), Курбонов Н.М., Маматов Н.С., Мирзаев Н.М., Мурадов Ф.А., Назирова Э.Ш., Нормуродов Ч.Б., Нуралиев Ф.А., Нуралиев Ф.М., Опанасенко В.Н. (Украина), Расулмухамедов М.М., Садуллаева Ш.А., Старовойтов В.В. (Беларусь), Хаётов А.Р., Халджигитов А., Хамдамов Р.Х., Хужаев И.К., Хужаеров Б.Х., Эшмаматова Д.Б., Дустмуродова Ш.Ж., Чье Ен Ун (Россия), Шабозов М.Ш. (Таджикистан), Dimov I. (Болгария), Li Y. (США), Mascagni M. (США), Min A. (Германия), Singh M. (Южная Корея).

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

При перепечатке материалов ссылка на журнал обязательна.
За точность фактов и достоверность информации ответственность несут авторы.

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(998) 71 263-41-98.
Э-почта: journals@airi.uz.
Веб-сайт: https://journals.airi.uz.

Дизайн и вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 28.10.2026 г.
Формат 60x84 1/8. Заказ № 6. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 5(76) 2026

The journal was established in 2015.
6 issues are published per year.

Founder:
Digital Technologies and Artificial Intelligence Development Research Institute.

Editor-in-Chief:
Ravshanov N.

Deputy Editors:
Aripov M.M., Shadimetov Kh.M., Akhmedov D.D.

Executive Secretary:
Ubaydullaev M.Sh.

Editorial Council:
Azamov A.A., Aloev R.D., Amirgaliev E.N. (Kazakhstan), Arushanov M.L., Abdullozoda S.F. (Tajikistan), Boborakhimov B.I., Burnashev V.F., Djumayozov U.Z., Zagrebina S.A. (Russia), Zadorin A.I. (Russia), Ignatiev N.A., Ilyin V.P. (Russia), Imankulov T.S. (Kazakhstan), Ismagilov I.I. (Russia), Kabanikhin S.I. (Russia), Kurbonov N.M., Mamatov N.S., Mirzaev N.M., Muradov F.A., Nazirova E.Sh., Normurodov Ch.B., Nuraliev F.A., Nuraliev F.M., Opanasenko V.N. (Ukraine), Sadullaeva Sh.A., Starovoitov V.V. (Belarus), Khayotov A.R., Khaldjigitov A., Khamdamov R.Kh., Khujaev I.K., Khujayorov B.Kh., Eshmamatova D.B., Dustmurodova Sh.J., Chye En Un (Russia), Shabozov M.Sh. (Tajikistan), Dimov I. (Bulgaria), Li Y. (USA), Mascagni M. (USA), Min A. (Germany), Singh M. (South Korea).

The journal is registered by Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan.
Certificate of Registration No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

At a reprint of materials the reference to the journal is obligatory.
Authors are responsible for the accuracy of the facts and reliability of the information.

Address:
100125, Tashkent, Buz-2, 17A.
Tel.: +(998) 71 263-41-98.
E-mail: journals@airi.uz.
Web-site: https://journals.airi.uz.

Layout design:
Sharipov Kh.D.

DTAIRI printing office.
Signed for print 28.10.2026
Format 60x84 1/8. Order No. 6. Print run of 100 copies.

Содержание

Расулов А.С., Раимова Г.М., Рахматов М.Й.
Метод Монте-Карло для решения нелинейного кинетического уравнения Больцмана 8

Эшкулов М.У., Хамдамов Р.Х.
Многокритериальная оптимизация системы подъёма воды с использованием солнечных фотоэлектрических панелей 22

Абдуллаев К.Д.
Принцип разделения классов и зависимость результатов кластеризации от меры расстояния 33

Рамазонов Ш.Ш.
Селекция объектов по отношениям с ближайшими объектами противоположного класса и сохранение локальной структуры обучающей выборки . . 46

Исомиддинов А.И., Акбарова А.А., Акбаров Б.Х.
Марковская переключающаяся векторная авторегрессия для скрытых гликемических режимов в реальных данных самоконтроля диабета: оценка на отложенной выборке 60

Ким К.К.
Сравнительный анализ XAI-подходов для обнаружения синтетической речи 72

Эшмаматова Д.Б., Солижанова К.А.
Пфаффианова структура и спектральная динамика дискретного оператора Лотки–Вольтерры: подход к интеграции геологических данных 95

Джумаёзов У.З., Абдирахмонова М.Н., Усмонов Л.С.
Численное решение трёхмерной связанной динамической задачи термоупругости в деформациях 119

Бакаев И.И., Ахмеджанова Д.А.
α-справедливая оптимизационная модель и численные алгоритмы выбора общей лексической нормы тюркских языков 137

Равшанов Н., Камолиддинова Ф., Lu Y.
Моделирование переноса и диффузии загрязняющих веществ в пограничном слое атмосферы с учётом орографии местности и метеорологических факторов 148

Contents

Rasulov A.S., Raimova G.M., Rakhmatov M.Y.
Monte Carlo solution of the nonlinear kinetic Boltzmann equation 8
Eshkulov M.U., Khamdamov R.Kh.
Multi-criteria optimization of solar photovoltaic-based water lifting systems . . . 22
Abdullaev K.D.
The principle of class separation and the dependence of clustering results on the distance measure 33
Ramazonov Sh.Sh.
Instance selection based on nearest-enemy relations and preservation of the local structure of a training sample 46
Isomiddinov A.I., Akbarova A.A., Akbarov B.X.
Markov-switching vector autoregression for latent glycaemic regimes in real-world diabetes self-monitoring data: an out-of-sample evaluation 60
Kim K.K.
Comparative analysis of XAI approaches for synthetic speech detection 72
Eshmamatova D.B., Solijanova K.A.
Pfaffian structure and spectral dynamics of a discrete Lotka–Volterra operator: a framework for geological evidence integration 95
Djumayozov U.Z., Abdirakhmonova M.N., Usmonov L.S.
Numerical solution of a three-dimensional coupled dynamic thermoelasticity problem in terms of strains 119
Bakayev I.I., Akhmedjanova D.A.
An α-fair optimization model and numerical algorithms for selecting a common lexical norm of Turkic languages 137
Ravshanov N., Kamoliddinova F., Lu Y.
Modeling pollutant transport and diffusion in the atmospheric boundary layer considering terrain orography and meteorological factors 148