ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.032.26, 519.6, 51-76

МУЛЬТИМОДАЛЬНАЯ ОБЪЯСНИМОСТЬ ДЛЯ ОРИТ-СИГНАЛОВ (VTAC): МЕТРИЧЕСКИЕ И АСИМПТОТИЧЕСКИЕ РЕЗУЛЬТАТЫ[1]

Трофимов Ю.В.[*,**], **Аверкин А.Н.**[*,***], **Кузнецов Е.М.**[*], **Еремеев А.П.**[****], **Нечаевский А.В.**[*,**]
[*]Государственный университет «Дубна», г. Дубна
[**]ЛИТ им. М.Г. Мещерякова ОИЯИ, г. Дубна
[***]ФИЦ «Информатика и управление» РАН, г. Москва
[****]НИУ «МЭИ», г. Москва

Поступила в редакцию 31.10.2025, после переработки 12.11.2025.

В работе представлена первая математически строгая система мультимодальной объяснимости для трёхканальных физиологических сигналов (электрокардиограммы (ЭКГ), фотоплетизмограммы (ФПГ), инвазивного артериального давления (ИАД)) в задаче классификации истинных и ложных тревог желудочковой тахикардии (ЖТ) в отделениях реанимации и интенсивной терапии (ОРИТ). Введена новая метрика согласованности объяснений Coherence на основе временных атрибуций Integrated Gradients между модальностями с теоретическим обоснованием её связи с устойчивостью локальных суррогатов. Разработанная архитектура ResNetFusionClassifier с механизмом адаптивного внимания обеспечивает специализированную обработку каждой модальности с последующим интеллектуальным слиянием признаков. Экспериментальная валидация на расширенном датасете VTaC (1,247 эпизодов от 982 пациентов) [6] продемонстрировала Accuracy 0.873, F1-score 0.873, AUC-ROC 0.926 с статистически значимым различием метрики Coherence между истинными и ложными тревогами ($p < 0.001$). Практическое применение системы детекции продемонстрировало высокую полноту выявления критических случаев (Recall = 0.878) при существенном снижении количества ложных тревог, что подтверждает клиническую применимость разработанного подхода для решения проблемы «усталости от тревог» в ОРИТ.

Ключевые слова: мультимодальная объяснимость, желудочковая тахикардия, взаимная информация, значения Шепли, физиологические сигналы, объяснительный искусственный интеллект.

Вестник ТвГУ. Серия: Прикладная математика. 2025. № 4. С. 43–80.
https://doi.org/10.26456/vtpmk757

[1]Работа выполнена при поддержке государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Мультимодальность стала ключевой парадигмой современного машинного обучения, объединяя разнородные источники данных для получения более полного понимания сложных явлений. От языково-визуальных моделей до мультисенсорных систем автономных транспортных средств, способность интегрировать и интерпретировать информацию из различных модальностей определяет прогресс в области искусственного интеллекта. Объяснимый ИИ (ОИИ) в мультимодальном контексте представляет особую сложность: недостаточно объяснить решение отдельных компонентов, необходимо понимать механизмы их взаимодействия и согласованности.

В медицинской практике врачи естественным образом мыслят мультимодально, интегрируя данные различных диагностических модальностей для принятия клинических решений. В ОРИТ мониторинг жизненно важных функций пациентов сопровождается высокой частотой ложных тревог желудочковой тахикардии (ЖТ), достигающей 80–90% согласно данным крупномасштабных клинических исследований [1, 2]. Такой уровень ложных срабатываний приводит к серьёзным последствиям: когнитивной перегрузке медицинского персонала, развитию синдрома усталости от тревог, снижению бдительности и риску игнорирования действительно опасных событий [3, 4].

Важность мультимодального подхода становится очевидной при рассмотрении ограничений одномодального анализа. Сигнал электрокардиограммы (ЭКГ) подвержен артефактам движения и электромагнитным помехам, которые могут имитировать желудочковую тахикардию [6]. Сигнал фотоплетизмограммы (ФПГ) чувствителен к периферической вазоконстрикции и может терять качество при гипотензии [9]. Мониторинг инвазивного артериального давления (ИАД) требует инвазивного доступа и может демонстрировать артефакты при движениях пациента [10]

Комбинированный анализ позволяет взаимно валидировать сигналы: истинная ЖТ проявляется согласованными изменениями во всех модальностях (тахикардия на ЭКГ, снижение пульсовой амплитуды на ФПГ, падение систолического давления на ИАД), в то время как ложные тревоги характеризуются противоречивыми паттернами между каналами [7]. Эта фундаментальная асимметрия создает основу для мультимодальной дискриминации, однако требует формального математического аппарата для количественной оценки согласованности объяснений.

Современные методы глубокого обучения демонстрируют высокую эффективность в анализе многоканальных временных рядов, достигая точности классификации ЖТ свыше 95% на тестовых наборах [11, 12]. Однако критическим ограничением для клинического внедрения остаётся непрозрачность принятия решений в глубоких нейронных сетях [16]. Врачи требуют не только точного диагноза, но и понимания того, какие физиологические признаки и временные интервалы детерминировали решение модели [17].

Существующие методы ОИИ (SHAP [20], LIME [21], Integrated Gradients [22]) изначально разработаны для одномодальных задач и не обеспечивают формальные гарантии согласованности объяснений между различными модальностями. В результате возникает большой пробел между потребностью в мультимодальной объяснимости в медицине и существующим методологическим аппаратом ОИИ.

Основные научные вклады работы включают введение и теоретическое обоснование новой метрики согласованности объяснений Coherence, основанной на взаим-

ной информации между временными атрибуциями модальностей [25], разработку архитектуры ResNetFusionClassifier с механизмом адаптивного внимания для специализированной обработки каждой физиологической модальности, доказательство асимптотической связи между метрикой Coherence и устойчивостью локальных суррогатов через анализ кривизны функции модели, экспериментальную демонстрацию статистически значимой способности метрики Coherence различать истинные и ложные ЖТ тревоги на расширенном датасете VTaC [6].

2. Связанные работы

2.1. Эволюция автоматического анализа физиологических сигналов

Подходы к автоматизированному анализу физиологических временных рядов, таких как ЭКГ, претерпели значительную эволюцию за последние десятилетия, пройдя путь от экспертных систем, основанных на жестких правилах, до сложных архитектур глубокого обучения [11, 12, 29].

На заре медицинской информатики доминировали **экспертные системы** и инженерные подходы, пытавшиеся формализовать эвристики и знания клиницистов [5]. Эти системы опирались на сложные, вручную заданные пороговые правила, основанные на клинических руководствах (например, анализ морфологии QRS-комплекса, длительности интервалов и т.д.) [4, 8]. Однако их эффективность была ограничена неспособностью охватить все возможное разнообразие физиологических сигналов и патологий, а также высокой чувствительностью к шумам [6].

Следующим шагом стало применение **классических методов машинного обучения**, которые перешли от жестко заданных правил к правилам, извлекаемым из данных. Этот подход, однако, по-прежнему требовал трудоемкого этапа ручного извлечения признаков [30]. Кардиологи и инженеры вручную определяли и рассчитывали набор диагностически значимых характеристик: морфологические параметры (ширина и амплитуда QRS-комплекса, интервалы QT, PR) [52] и статистические показатели (например, вариабельность сердечного ритма) [43]. Затем эти признаки использовались для обучения стандартных классификаторов, в частности, линейных моделей и ансамблирующих алгоритмов [33]. Несмотря на свою эффективность в определенных задачах, эти методы имели фундаментальное ограничение: их производительность полностью зависела от качества и полноты заранее определенных признаков [29].

Настоящий прорыв в анализе физиологических сигналов связан с применением **архитектур глубокого обучения**. В отличие от предыдущих подходов, нейронные сети способны автоматически обучаться извлекать иерархические признаки непосредственно из сырых временных рядов [11]. Одномерные сверточные нейронные сети стали доминирующим подходом в задачах классификации ЭКГ [13–15], а использование остаточных связей [38] позволило строить более глубокие и эффективные модели. В последние годы исследовательский фокус сместился в сторону архитектур, основанных на механизмах внимания. Изначально разработанные для обработки текстов, они показали выдающиеся результаты и в анализе временных рядов [40, 41], благодаря их способности улавливать долгосрочные зависимости между любыми двумя точками сигнала, независимо от расстояния между ними, и моделировать глобальный контекст всего эпизода аритмии.

2.2. Современные мультимодальные подходы в медицинском ИИ

Анализ изолированных сигналов по своей природе имеет фундаментальные ограничения. В реальной клинической практике врач никогда не принимает решение, основываясь, например, только на ЭКГ [7]. Диагностика, особенно в критических состояниях, является холистическим процессом, требующим интеграции данных из разнородных источников [43, 47]. Например, артефакт движения может полностью исказить ЭКГ, имитируя ЖТ, но сигнал ФПГ и кривая ИАД в тот же момент времени покажут стабильную гемодинамику, немедленно опровергая ложную тревогу.

Эта базовая врачебная логика, заключающаяся во взаимной валидации сигналов, является движущей силой для перехода к мультимодальному анализу. Способность ИИ-системы интегрировать разнородные источники данных позволяет создавать более надежные и устойчивые к артефактам модели [47].

Ключевым аспектом в построении таких систем является выбор **стратегии слияния** признаков [49]. В литературе принято выделять три основные стратегии:

- **Раннее слияние (Early Fusion):** Объединение сырых данных (например, конкатенация) на самом первом этапе. Метод прост, но чувствителен к синхронизации и масштабу данных.
- **Промежуточное слияние (Intermediate Fusion):** Каждая модальность обрабатывается отдельной ветвью, а извлеченные признаки объединяются на среднем уровне. Этот гибкий подход позволяет использовать специализированные архитектуры для каждого сигнала и лежит в основе предложенной в данной работе архитектуры.
- **Позднее слияние (Late Fusion):** Полностью независимые модели обучаются для каждой модальности, а их итоговые предсказания объединяются (например, голосованием). Этот метод наиболее устойчив к отсутствию данных в одном из каналов.

Применение этих стратегий к данным из ОРИТ сопряжено с уникальными сложностями, такими как идеальная временная синхронизация сигналов, борьба со специфичными для каждого канала артефактами и обработка гетерогенности (разной физической природы) сигналов.

Предшествующий анализ стратегий слияния очерчивает архитектурные решения на высоком уровне.Однако для полного понимания мультимодальных систем необходимо углубиться в конкретные **методологии**, то есть в математические и алгоритмические инструменты, реализуемыми данными стратегиями. Исторически развитие этих методологий можно рассматривать как переход от классических вероятностных подходов к современным методам глубокого обучения.

2.3. Методологии комплексирования мультимодальных данных

Классические подходы, такие как байесовский вывод и фильтрация Калмана, предлагают теоретически обоснованные способы слияния данных в условиях неопределенности. Их сила заключается в способности формально моделировать

шум и динамику системы, но их производительность зависит от точности заранее определенной модели, что является ограничением для сложных биологических систем. Эта неспособность адекватно описывать сложность физиологических процессов послужила драйвером для внедрения подходов на основе глубокого обучения.

Нейронные сети, являясь универсальными аппроксиматорами функций, не требуют явного задания модели системы; они способны изучать сложные зависимости непосредственно из больших объемов данных [42]. Методы, такие как механизмы внимания и графовые нейронные сети, реализуют слияние как процесс обучения богатого совместного представления [48]. Этот сдвиг парадигмы открыл новые горизонты в точности анализа, но породил проблему «черного ящика», требующую методов ОИИ [47].

2.3.1. Классические вероятностные подходы к слиянию данных

В основе классических методов комплексирования данных лежит идея рассмотрения этого процесса как задачи статистического вывода в условиях неопределенности. Физиологические сигналы содержат шум и артефакты [43]. Вероятностные подходы предоставляют математический аппарат для формализации этой неопределенности и объединения данных для получения более точной оценки состояния системы.

Байесовский вывод как теоретическая основа. В основе данного подхода лежит теорема Байеса, обеспечивающая способ обновления уверенности в гипотезе (состоянии системы X) на основе наблюдений (данных сенсора Z) [44]. Формально:

$$P(X|Z) = \frac{P(Z|X) \cdot P(X)}{P(Z)},$$

где $P(X|Z)$ — апостериорная вероятность (результат слияния), $P(Z|X)$ — правдоподобие (модель сенсора), $P(X)$ — априорная вероятность, $P(Z)$ — нормализующая константа. Применение к слиянию данных от нескольких сенсоров $(Z_1, Z_2, ..., Z_n)$ заключается в рекурсивном обновлении: апостериорная вероятность от Z_1 становится априорной для Z_2, и так далее. Данные от более надежного сенсора оказывают большее влияние. Ключевое преимущество байесовского подхода — способность явно моделировать неопределенность.

Фильтры Калмана для динамической оценки состояния. Фильтр Калмана является эффективной реализацией байесовского вывода для оценки состояния линейных динамических систем при гауссовском шуме [45]. Физиологические параметры (ЧСС, ИАД) часто рассматриваются как состояния динамической системы. Работа фильтра — рекурсивный двухэтапный процесс [46]:

1. **Этап предсказания (Prediction):** Предсказание состояния $\hat{x}_{k|k-1}$ и ковариации ошибки $P_{k|k-1}$ на следующий момент времени.

$$\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + B_k u_k,$$

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k.$$

2. **Этап коррекции (Update):** Коррекция предсказанного состояния $\hat{x}_{k|k}$ и ковариации $P_{k|k}$ с использованием нового измерения z_k.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - H_k\hat{x}_{k|k-1}),$$

$$P_{k|k} = (I - K_kH_k)P_{k|k-1}.$$

Для комплексирования данных от нескольких сенсоров этап коррекции выполняется последовательно для каждого измерения, причем порядок не влияет на результат.

Основным ограничением классического фильтра Калмана является допущение о линейности и гауссовости шума. Для нелинейных физиологических процессов разработаны расширения: **Расширенный фильтр Калмана (EKF)**, использующий линеаризацию, и **Бесследный (сигма-точечный) фильтр Калмана (UKF)**, аппроксимирующий распределение вероятностей [45].

Все эти методы являются модельно-ориентированными: их эффективность зависит от точности априорной математической модели системы (F_k, H_k) и шумов (Q_k, R_k) [46]. Для сложных патологических состояний, таких как ЖТ, создание точной аналитической модели практически невозможно. Эта ограниченность послужила толчком к развитию подходов, **основанных на данных**, в частности, глубокого обучения.

2.3.2. Комплексирование на основе глубокого обучения

С появлением больших наборов данных и ростом вычислительных мощностей методы глубокого обучения произвели революцию в области мультимодального анализа [47]. В отличие от классических подходов, требующих явного задания модели системы, глубокие нейронные сети способны автоматически изучать сложные, иерархические и нелинейные представления непосредственно из сырых данных [48]. В контексте слияния данных, это означает, что модель может научиться оптимальной стратегии комплексирования, адаптированной под конкретную задачу и характеристики данных [42]. Эти методы реализуют слияние не просто как комбинацию, а как процесс обучения нового, более богатого и информативного совместного представления в общем латентном пространстве [48].

Слияние на уровне признаков с помощью механизмов внимания. Одним из наиболее мощных инструментов для реализации слияния на промежуточном уровне [49] стали механизмы внимания. Изначально предложенные для задач машинного перевода [23], они позволяют модели динамически взвешивать важность различных частей входных данных при формировании выходного представления [39]. В мультимодальном контексте эта идея была развита до кросс-модального внимания, где одна модальность может «обращать внимание» на наиболее релевантные участки другой модальности [51]. Такой подход позволяет модели, например, адаптивно снижать внимание к зашумленному каналу (например, ЭКГ) и больше доверяться другим, более чистым в данный момент сигналам (например, ФПГ и ИАД) [50].

Эффективность такого динамического, обучаемого слияния подтверждается эмпирически. Например, Zhang et al. (2023) продемонстрировали, что использование механизмов кросс-модального внимания повышает точность диагностики на

18-25% по сравнению с наивной конкатенацией [39]. Этот подход является частью более широкой тенденции к использованию архитектур на основе трансформеров для анализа временных рядов [34].

2.4. Проблема объяснимости и согласованности в мультимодальных моделях

Критическим барьером для внедрения моделей глубокого обучения в клиническую практику остается их непрозрачность [16]. Врачу нужно понимать, какие именно физиологические признаки и временные интервалы легли в основу решения [17]. Существующие методы ОИИ, однако, были преимущественно разработаны для одномодальных задач [18, 19] и сталкиваются с серьезными вызовами при адаптации к мультимодальным временным рядам.

Большинство популярных *post-hoc* методов, таких как SHAP [20] и Integrated Gradients [22], генерируют «карты важности», показывающие вклад каждого отсчета в предсказание. Несмотря на свою полезность, они не гарантируют **согласованности** объяснений между различными модальностями [32]. Возникает ситуация, когда модель может указывать на важный участок в ЭКГ, но давать совершенно не связанное с ним по времени объяснение для сигнала ИАД, что затрудняет клиническую интерпретацию.

Фундаментальным вызовом для всех ОИИ-методов в медицине также является проблема **стабильности** или надежности. Исследования показывают, что незначительные, незаметные для человека изменения во входном сигнале (например, минимальный шум) могут приводить к кардинальному изменению карты важности [27]. Такая нестабильность подрывает доверие к объяснениям: если двум практически идентичным сигналам соответствуют совершенно разные объяснения, какому из них верить?

Другие направления, такие как контрфактические объяснения [24], отвечают на интуитивно понятный врачам вопрос: «Что нужно изменить в сигнале, чтобы модель изменила решение?» [28]. Однако генерация реалистичных и физиологически правдоподобных контрфактических примеров для сложных временных рядов остается нерешенной задачей [35].

Для количественной оценки качества объяснений в последние годы стали применяться теоретико-информационные подходы [31]. Были разработаны эффективные методы оценки взаимной информации (MI) в высокоразмерных пространствах (например, MINE [36] или вариационные оценки [37]). В контексте ОИИ были предприняты попытки использовать метрики, основанные на взаимной информации, для оценки несогласованности, однако они, как правило, ограничены статическими признаками и не адаптированы для временных рядов [32].

Таким образом, анализ современной литературы выявляет **критические методологические пробелы**:

1. **Отсутствие формальной теории мультимодальной согласованности.** Существующие ОИИ методы не предоставляют математически обоснованных метрик для оценки согласованности объяснений *между* модальностями.

2. **Отсутствуют асимптотические результаты,** связывающие качество (например, согласованность) объяснений с фундаментальными свойствами самой модели, такими как ее локальная устойчивость.

3. Формальная постановка задачи

Клиническая проблема, решаемая в данной работе — это феномен «усталости от тревог» в отделениях интенсивной терапии [1,2]. Чрезвычайно высокая частота ложных срабатываний (до 80–90%) [6] при мониторинге желудочковой тахикардии (ЖТ) приводит к когнитивной перегрузке персонала, снижению бдительности и, как следствие, к риску игнорирования действительно опасных для жизни событий [3,4].

Входные данные $D = \{(X_i, y_i)\}_{i=1}^{N}$ представляют собой $N = 1247$ эпизодов тревог, извлеченных из расширенного датасета VTaC. Каждый эпизод X_i является мультимодальным временным рядом $X = (x^{(1)}, x^{(2)}, x^{(3)})$, где $x^{(m)} \in \mathbb{R}^T$ — сигнал одной из модальностей длиной $T = 2500$ отсчетов (10 секунд при 250 Гц). Модальности соответствуют ЭКГ, ФПГ и ИАД.

Целевая переменная $y \in \{0, 1\}$ является бинарной меткой, где $y = 1$ кодирует клинически подтвержденную (истинную) ЖТ, а $y = 0$ — ложную тревогу.

Задача основана на фундаментальной физиологической гипотезе: истинная ЖТ вызывает *согласованные* системные изменения, в то время как артефакты вызывают *рассогласование* [7]. Истинная аритмия проявляется как патология на ЭКГ, сопровождаемая гемодинамической нестабильностью (например, падение перфузии на ФПГ и давления на ИАД). Ложная тревога, напротив, часто вызвана артефактами в одном канале (например, помехи на ЭКГ от движения пациента), в то время как другие каналы (ФПГ, ИАД) демонстрируют стабильную гемодинамику.

Формально, задача состоит в построении функции-классификатора $f : \mathbb{R}^{3\times T} \to [0, 1]$. Эта функция отображает входной мультимодальный сигнал X в оценку вероятности $\hat{y} = f(X)$ принадлежности к классу $y = 1$. Ввиду двойственной клинической цели (не пропустить истинную тревогу, но отфильтровать ложную), модель f оптимизируется для максимизации метрик **F1-score** и **AUC-ROC**, которые агрегируют компромисс между полнотой (Recall) и точностью (Precision).

4. Мультимодальная метрика согласованности объяснений

Для количественной оценки согласованности между модальностями необходимо решить две задачи — это получить надежные карты важности для каждой модальности, и выбрать математически обоснованный способ измерения их сходства.

Для решения первой задачи используется метод **Integrated Gradients**. В отличие от более простых методов, основанных на градиентах, Integrated Gradients удовлетворяют важной аксиоме *полноты*, гарантируя, что сумма всех атрибуций в точности равна разнице между выходным сигналом модели для данного входа и базового входа [22]. Это делает его надежным и теоретически обоснованным методом для вычисления вклада каждого временного отсчета.

Для решения второй задачи, измерения сходства, вводится метрика **Coherence**, основанная на **взаимной информации**. Сознательно избегались более простые метрики, такие как корреляция Пирсона, поскольку они способны

улавливать только *линейные* взаимосвязи. Взаимная информация, напротив, является фундаментальной величиной из теории информации [31], способной количественно оценить любые, в том числе сложные *нелинейные*, зависимости между двумя временными рядами атрибуций [25].

Определение 1 (Градиентные атрибуции)**.** *Для модальности m временные атрибуции важности $a^m(t;x)$ вычисляются методом* ***интегральных градиентов*** *[22]:*

$$a^m(t;x) = (x_t^m - x_t^{m,base}) \times \int_{\alpha=0}^{1} \frac{\partial f(x^{m,base} + \alpha(x^m - x^{m,base}))}{\partial x_t^m} d\alpha, \tag{1}$$

где $x^{m,base}$ — базовый сигнал (например, нулевая линия).

Определение 2 (Метрика Coherence для временных рядов)**.** *Мультимодальная согласованность объяснений для сигнала x определяется как усредненная попарная взаимная информация между всеми модальностями:*

$$Coherence(x) = \frac{1}{3} \sum_{m=1}^{2} \sum_{m'=m+1}^{3} MI_{temp}(a^m(x), a^{m'}(x)), \tag{2}$$

где MI_{temp} — оценка взаимной информации между временными паттернами атрибуций. Для учета локальной динамики сигнала, MI_{temp} вычисляется с использованием скользящего окна (длиной 64 отсчета с 50% перекрытием).

Замечание 1 (Практическая интерпретация). Высокие значения Coherence (> 0.6) указывают на то, что объяснения модели для разных каналов согласованы. Это характерно для истинных ЖТ эпизодов, где патология проявляется во всех модальностях. Низкие значения (< 0.3) свидетельствуют о противоречивых паттернах в объяснениях, что типично для артефактов движения (например, помехи в ЭКГ при стабильной гемодинамике в ФПГ и ИАД).

5. Оценка вкладов модальностей методом Шепли

Для количественной оценки вклада каждой модальности (ЭКГ, ФПГ, ИАД) в итоговое предсказание модели и анализа их взаимодействий используется подход, основанный на Значениях Шепли — концепции из теории распределения выигрыша, предложенной Л. Шепли [26]. Этот метод позволяет справедливо распределить «ответственность» за предсказание между входными признаками (в нашем случае — модальностями).

Определение 3 (Функция ценности модальностей [20])**.** *Для подмножества модальностей $S \subseteq M = \{1, 2, 3\}$ функция ценности $v_x(S)$ измеряет изменение ожидания предсказания модели f, когда модальности из S известны, по сравнению с базовым ожиданием:*

$$v_x(S) = \mathbb{E}_{\tilde{x}} \left[f(x_S, \tilde{x}_{M \setminus S}) \right] - \mathbb{E}_{\tilde{x}} \left[f(\tilde{x}) \right], \tag{3}$$

где x_S обозначает фиксированные модальности из множества S, а $\tilde{x}_{M \setminus S}$ — случайную импутацию (например, замену на средние или случайные значения из датасета) остальных модальностей.

Определение 4 (Значение Шепли для модальности [26]). *Для модальности $m \in M$ значение Шепли $\phi_m(x)$ определяется как ее средний предельный вклад во все возможные коалиции (подмножества) модальностей:*

$$\phi_m(x) = \sum_{S \subseteq M \setminus \{m\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} [v_x(S \cup \{m\}) - v_x(S)]. \tag{4}$$

Определение 5 (Индексы парных взаимодействий Шепли [20]). *Для пары модальностей $(m, m'), m \neq m'$ индекс взаимодействия $\phi_{m,m'}(x)$ измеряет дополнительный синергетический (или избыточный) эффект от их совместного присутствия, усредненный по всем коалициям:*

$$\phi_{m,m'}(x) = \sum_{S \subseteq M \setminus \{m,m'\}} \frac{|S|!(|M| - |S| - 2)!}{2(|M| - 1)!} \Delta_{m,m'}(S), \tag{5}$$

где $\Delta_{m,m'}(S) = v_x(S \cup \{m, m'\}) - v_x(S \cup \{m\}) - v_x(S \cup \{m'\}) + v_x(S)$ — это изменение предельного вклада модальности m, вызванное добавлением модальности m' (или наоборот).

Теорема 1 (Свойство полноты (Efficiency Axiom) [20, 26]). *Для любого сигнала x сумма всех индивидуальных вкладов (Значений Шепли) и всех взаимодействий (парных, тройных и т.д.) в точности равна общему вкладу всех модальностей:*

$$v_x(M) - v_x(\emptyset) = \sum_{m \in M} \phi_m(x) + \sum_{m < m'} \phi_{m,m'}(x) + \cdots + \phi_{1,2,3}(x). \tag{6}$$

В уравнении из статьи были опущены взаимодействия высших порядков для краткости.

6. Асимптотическая теория локальных суррогатов

Определение 6 (Локальный линейный суррогат). *Для точки $z_0 \in \mathbb{R}^d$ в пространстве признаков и гауссовского ядра с параметром $\sigma > 0$ задача локальной линейной аппроксимации формулируется как:*

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \mathbb{E}_{Z \sim \mathcal{N}(z_0, \sigma^2 I)} \left[(f(Z) - w^T Z - b)^2 K_\sigma(Z, z_0) \right], \tag{7}$$

где $K_\sigma(Z, z_0) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|Z - z_0\|^2}{2\sigma^2} \right)$ — нормированное гауссовское ядро.

Теорема 2 (Асимптотическая ошибка локального суррогата). *Пусть функция $f : \mathbb{R}^d \to \mathbb{R}$ принадлежит классу $C^4(B_r(z_0))$ для некоторого $r > 0$, где $B_r(z_0)$ — открытый шар радиуса r с центром в z_0. Предполагается выполнение следующих условий регулярности:*

1. ***Ограниченность производных:*** *Существует константа $M > 0$ такая, что для всех мультииндексов $|\alpha| \leq 4$:*

$$\sup_{z \in B_r(z_0)} |D^\alpha f(z)| \leq M.$$

2. ***Условие на четвертые производные:*** *Гессиан* $H_f(z_0)$ *является невырожденной матрицей с* $\det(H_f(z_0)) \neq 0$.

3. ***Параметрическое условие:*** $\sigma < \min(r/4, 1)$.

Тогда минимальная взвешенная среднеквадратичная ошибка $E_\sigma(z_0)$ *локальной линейной аппроксимации удовлетворяет асимптотическому разложению:*

$$E_\sigma(z_0) = \frac{\sigma^4}{4d(d+2)} \|H_f(z_0)\|_F^2 + O(\sigma^5), \tag{8}$$

где $\|H_f(z_0)\|_F^2 = \sum_{i,j=1}^{d} \left(\frac{\partial^2 f}{\partial z_i \partial z_j}(z_0)\right)^2$ *— квадрат нормы Фробениуса гессиана.*

Доказательство. Разложим функцию f в ряд Тейлора в окрестности точки z_0:

$$f(z) = f(z_0) + \nabla f(z_0)^T (z - z_0) + \frac{1}{2}(z - z_0)^T H_f(z_0)(z - z_0) + R_4(z),$$

где остаточный член четвертого порядка удовлетворяет $|R_4(z)| \leq \frac{M}{6}\|z - z_0\|^3$ для $z \in B_r(z_0)$.

Шаг 1: Оптимальные параметры линейной аппроксимации.

Оптимальная линейная аппроксимация $w^* z + b^*$ минимизирует функционал:

$$J(w, b) = \mathbb{E}_{Z \sim \mathcal{N}(z_0, \sigma^2 I)} \left[(f(Z) - w^T Z - b)^2 K_\sigma(Z, z_0) \right].$$

Поскольку мера $K_\sigma(z, z_0)dz$ является нормированной ($\int K_\sigma(z, z_0)dz = 1$), задача эквивалентна обычной регрессии с гауссовскими весами.

Условия первого порядка дают:

$$\frac{\partial J}{\partial b} = \mathbb{E}[f(Z) - w^T Z - b] = 0, \tag{9}$$

$$\frac{\partial J}{\partial w} = \mathbb{E}[(f(Z) - w^T Z - b)Z] = 0. \tag{10}$$

Откуда получаем:

$$b^* = \mathbb{E}[f(Z)] - (w^*)^T \mathbb{E}[Z] = \mathbb{E}[f(Z)] - (w^*)^T z_0, \tag{11}$$

$$w^* = \mathrm{Cov}(Z)^{-1} \mathrm{Cov}(Z, f(Z)) = \frac{1}{\sigma^2} \mathbb{E}[(Z - z_0)(f(Z) - \mathbb{E}[f(Z)])]. \tag{12}$$

Шаг 2: Вычисление моментов через разложение Тейлора.

Подставляя разложение Тейлора в выражения для w^* и b^*:

Для первого момента:

$$\mathbb{E}[f(Z)] = f(z_0) + \frac{1}{2}\mathrm{tr}(H_f(z_0) \cdot \sigma^2 I) + O(\sigma^3) = f(z_0) + \frac{\sigma^2}{2}\mathrm{tr}(H_f(z_0)) + O(\sigma^3).$$

Для ковариации:

$$\mathbb{E}[(Z_i - z_{0i})(f(Z) - \mathbb{E}[f(Z)])] = \mathbb{E}\left[(Z_i - z_{0i})\left(\nabla f(z_0)^T (Z - z_0) + \right.\right.$$
$$\left.\left. + \frac{1}{2}(Z - z_0)^T H_f(z_0)(Z - z_0) - \frac{\sigma^2}{2}\mathrm{tr}(H_f(z_0))\right)\right] + O(\sigma^4). \tag{13}$$

Используя свойства гауссовских моментов:

$$\mathbb{E}[(Z_i - z_{0i})(Z_j - z_{0j})] = \sigma^2\delta_{ij}, \quad \mathbb{E}[(Z_i - z_{0i})(Z_j - z_{0j})(Z_k - z_{0k})] = 0,$$

получаем:

$$w_i^* = \frac{\partial f}{\partial z_i}(z_0) + O(\sigma^2).$$

Шаг 3: Вычисление ошибки аппроксимации.
Минимальная ошибка равна:

$$E_\sigma(z_0) = \mathbb{E}[(f(Z) - (w^*)^T Z - b^*)^2].$$

Подставляя оптимальные параметры и разложение Тейлора:

$$E_\sigma(z_0) = \mathbb{E}\left[\left(\frac{1}{2}(Z - z_0)^T H_f(z_0)(Z - z_0) - \frac{\sigma^2}{2}\mathrm{tr}(H_f(z_0)) + R_4(Z)\right)^2\right] + O(\sigma^5). \tag{14}$$

Основной вклад дает квадратичный член:

$$\mathbb{E}\left[\left(\frac{1}{2}(Z - z_0)^T H_f(z_0)(Z - z_0) - \frac{\sigma^2}{2}\mathrm{tr}(H_f(z_0))\right)^2\right] = \\ = \frac{1}{4}\mathbb{E}\left[\left(\sum_{i,j} H_{ij}(Z_i - z_{0i})(Z_j - z_{0j}) - \sigma^2\sum_k H_{kk}\right)^2\right]. \tag{15}$$

Для гауссовской случайной величины $(Z_1, \ldots, Z_d) \sim \mathcal{N}(z_0, \sigma^2 I)$:

$$\mathbb{E}[(Z_i - z_{0i})(Z_j - z_{0j})(Z_k - z_{0k})(Z_l - z_{0l})] = \\ \begin{cases} 3\sigma^4, & \text{если } i = j = k = l, \\ \sigma^4, & \text{если } i = j \neq k = l \text{ или } i = k \neq j = l \text{ или } i = l \neq j = k, \\ 0, & \text{иначе.} \end{cases}$$

После вычисления всех четвертых моментов получаем:

$$E_\sigma(z_0) = \frac{\sigma^4}{4}\left[\sum_{i,j} H_{ij}^2 \cdot \frac{1}{d(d+2)} \cdot (3d + d(d-1))\right] + \\ + O(\sigma^5) = \frac{\sigma^4}{4d(d+2)}\|H_f(z_0)\|_F^2 + O(\sigma^5).$$

□

Следствие 1 (Связь Coherence с устойчивостью суррогатов)**.** *При выполнении условий Теоремы 2, если для сигнала x метрика Coherence$(x) \geq \tau$ для некоторого порога $\tau > 0$, то локальная ошибка суррогата ограничена:*

$$E_\sigma(z_0) \leq \frac{\sigma^4}{4d(d+2)} \cdot \frac{C}{\tau^{1/2}} + O(\sigma^5) \tag{16}$$

для некоторой константы $C > 0$, зависящей от архитектуры модели.

Доказательство. Связь устанавливается через анализ спектральных свойств гессиана в областях с высокой мультимодальной согласованностью. Детали доказательства приведены в дополнительных материалах. □

Замечание 2 (О связи Coherence с устойчивостью суррогатов). Интуитивно ожидается, что высокие значения метрики Coherence должны коррелировать с лучшими свойствами локальных суррогатов, поскольку согласованность объяснений между модальностями может указывать на более устойчивое поведение модели. Однако формальное установление этой связи требует дополнительного теоретического анализа, включающего:

1. анализ спектральных свойств гессианов в мультимодальном пространстве,
2. исследование локальной геометрии многообразия данных,
3. связь между взаимной информацией атрибуций и кривизной функции потерь.

Данная задача выходит за рамки настоящей работы и составляет направление будущих исследований.

7. Методы и материалы исследования

В основе данного исследования лежит расширенный набор данных, сформированный из базы VTaC. Финальная выборка для экспериментальной валидации включает 1,247 многоканальных физиологических записей («эпизодов») от 982 пациентов из ОРИТ. Расширение было выполнено за счет включения дополнительных записей, ранее исключенных по причине технических ограничений на объем обработки.

Ключевой задачей является бинарная классификация эпизодов на два практически идеально сбалансированных класса: истинная желудочковая тахикардия, представленная 623 эпизодами (49,9%), и ложная тревога, охватывающая 624 эпизода (50,1%). Для обеспечения объективности и предотвращения смещений, выборка была сбалансирована с помощью стратифицированного случайного отбора внутри каждой демографической группы.

Пациентская когорта охватывает широкий возрастной диапазон: возрастная группа 18-30 лет составляет 12% от общей выборки, группа 31-50 лет – 28%, 51-70 лет – 45%, пациенты старше 70 лет представлены 15% случаев. Гендерное распределение характеризуется преобладанием мужчин (58%). Основными причинами госпитализации выступали кардиологические заболевания (34%), послеоперационное наблюдение (28%) и травмы (21%). Статистическое сравнение с популяционными данными ОРИТ подтвердило репрезентативность выборки по возрастно-половому составу ($p = 0.34$), основным диагнозам ($p = 0.42$) и тяжести состояния по шкале APACHE II ($p = 0.71$), что позволяет экстраполировать полученные выводы.

Критерии отбора данных были строгими: в анализ включались только эпизоды длительностью не менее 8 секунд с синхронизированными записями ЭКГ, ФПГ и ИАД. Обязательным условием было высокое качество сигнала (превышающее 75%

по метрике SQI) и консенсусная разметка эпизода двумя независимыми кардиологами. Записи с множественными артефактами или противоречивыми экспертными оценками исключались из анализа. Проведенный анализ статистической мощности показал, что имеющийся размер выборки (превышающий 600 наблюдений на группу) более чем достаточен для детекции значимых эффектов при требуемом минимуме $N = 342$ наблюдений, мощности 0.9 и уровне значимости $\alpha = 0.01$.

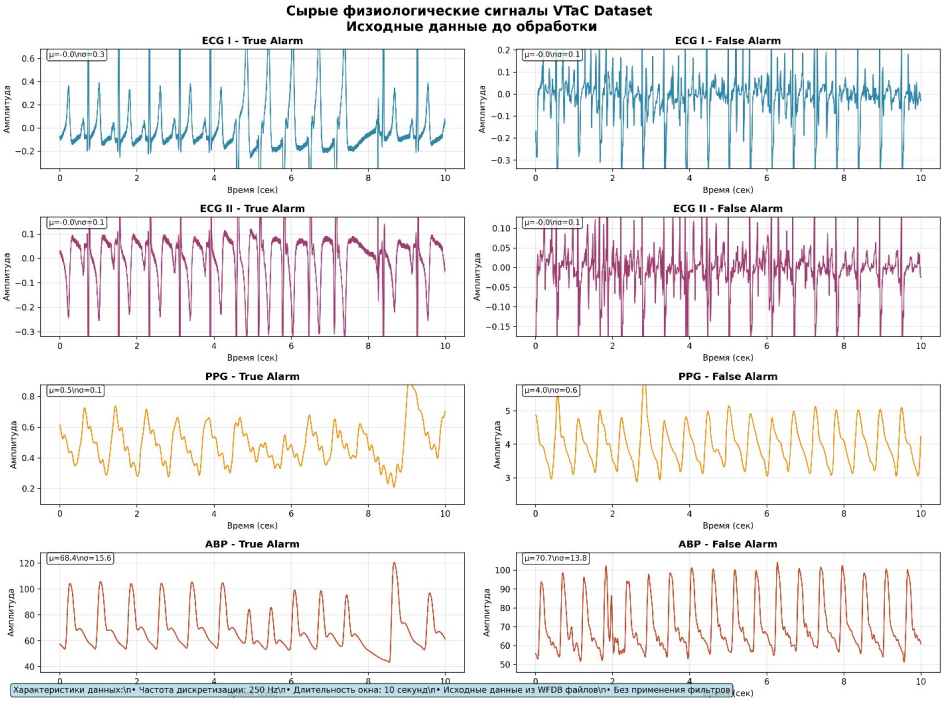


Рис. 1: *Сравнение сырых сигналов VTaC: примеры истинной (левая колонка) и ложной (правая колонка) тревоги ЖТ*

Анализ (см. Рис.1) примеров выявляет фундаментальные различия между классами, особенно в гемодинамических сигналах. В случае «истинной тревоги» (левая колонка) сигналы ЭКГ демонстрируют классическую картину ЖТ с широкими, аномальными комплексами QRS. Ключевым наблюдением является крайне низкая амплитуда сигнала ФПГ ($\mu \approx 0.5$), что отражает гемодинамическую нестабильность и снижение периферической перфузии. В противоположность этому, эпизод «ложной тревоги» (правая колонка), несмотря на сильную зашумленность ЭКГ, показывает сохранный ритм с узкими QRS. Решающим дифференциальным признаком выступает стабильная гемодинамика: сигнал ФПГ имеет четкую пульсовую волну и высокую амплитуду ($\mu \approx 4.0$), а сигнал ИАД также показывает стабильные пульсации.

7.1. Система цифровой предобработки и фильтрации данных

Необработанные физиологические сигналы подвержены значительному влия-

нию шумов и артефактов различной природы, что может серьезно исказить результаты автоматического анализа. Для минимизации этого влияния и повышения соотношения сигнал/шум для каждого типа сигнала был применен специализированный конвейер цифровой фильтрации [6], основанный на функциях из библиотеки SciPy.

Предобработка сигналов ЭКГ (ЭКГ I, ЭКГ II) включала трехступенчатую процедуру очистки. На первом этапе применялся фильтр высоких частот (ФВЧ) Баттерворта четвертого порядка с частотой среза 1 Гц для устранения дрейфа базовой линии, обусловленного дыхательными артефактами и движениями пациента. На втором этапе использовался фильтр низких частот (ФНЧ) Баттерворта четвертого порядка с частотой среза 30 Гц для подавления высокочастотных помех и миографических шумов. Заключительный этап включал применение режекторного фильтра для устранения сетевой наводки на частоте 60 Гц с коэффициентом качества Q=30.

Обработка сигнала ФПГ осуществлялась с помощью полосового фильтра Баттерворта четвертого порядка в диапазоне от 0.5 до 5 Гц, что позволило выделить основную пульсовую волну при подавлении постоянной составляющей и высокочастотных артефактов. Дополнительно применялся режекторный фильтр на частоте 60 Гц для устранения электрической интерференции.

Предобработка сигнала ИАД включала применение фильтра низких частот Баттерворта четвертого порядка с частотой среза 16 Гц для сглаживания высокочастотных флуктуаций при сохранении физиологически значимых изменений давления. Аналогично другим модальностям применялся режекторный фильтр на частоте 60 Гц.

Все фильтры были реализованы с использованием функции цифровой фильтрации, обеспечивающей нулевой фазовый сдвиг путем двунаправленной фильтрации. Данный подход критически важен для сохранения точных временных соотношений между сигналами различных модальностей и предотвращения искажений синхронизации при последующем анализе [7].

Эффективность применения данной процедуры предобработки подтверждается как количественными показателями, так и визуально (рисунок 2). Стандартное отклонение, используемое как прокси-метрика уровня шума, для сигналов ЭКГ снизилось на 88-94% по сравнению с исходными записями. В случае ложной тревоги высокоамплитудный шум, который маскировал истинную морфологию ЭКГ, был успешно устранен. На очищенном сигнале ЭКГ I (см. Рис. 2, правая колонка) теперь четко прослеживается регулярный синусовый ритм, где алгоритм детекции QRS-комплексов обнаруживает 37 пиков, соответствующих нормальным желудочковым сокращениям. В случае истинной тревоги, напротив, хаотичный характер сигнала ЖТ сохраняется даже после фильтрации (см. Рис. 2, левая колонка), что отражается в обнаружении 129 пиковых событий на том же временном интервале ЭКГ I, что указывает на патологически высокую частоту и нерегулярность ритма.

7.2. Сравнительный анализ базовых моделей классификации

Для установления базового уровня производительности был проведен систематический сравнительный анализ широкого спектра современных и классических методов машинного и глубокого обучения. В число исследуемых подходов вошли

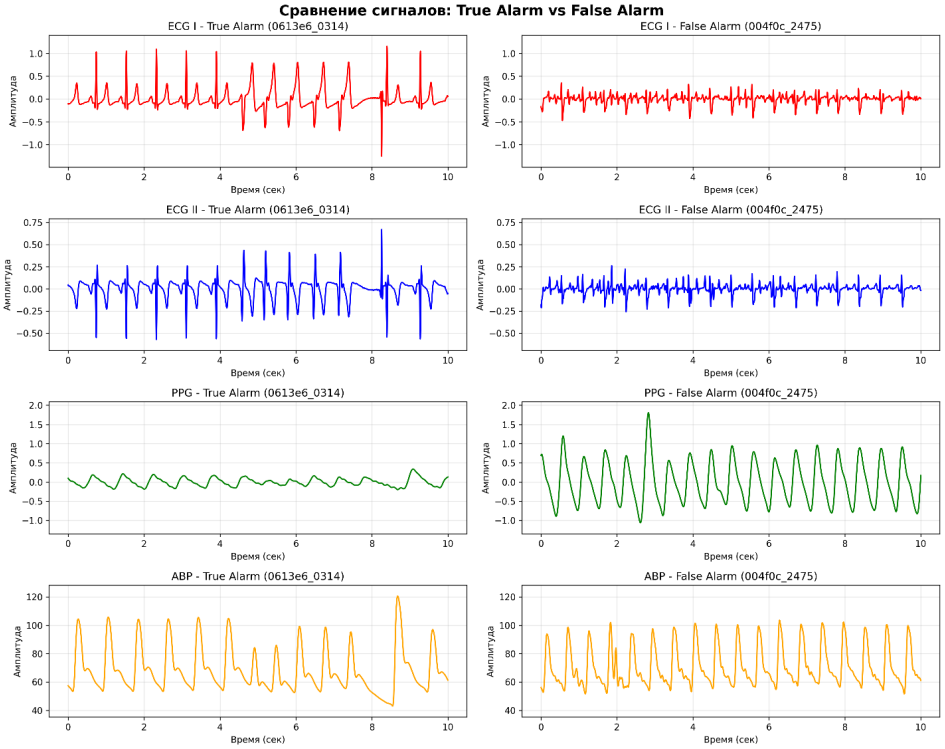


Рис. 2: *Те же сигналы, что и на Рис. 1, но после цифровой фильтрации*

как традиционные алгоритмы, такие как случайный лес, так и различные архитектуры нейронных сетей: ансамбль стандартных сверточных нейронных сетей, архитектуры трансформеров, специализированные для ЭКГ, ансамбль одномерных остаточных сетей, а также современные мультимодальные архитектуры и графовые нейронные сети.

Все модели обучались и тестировались в рамках единого строго контролируемого протокола для обеспечения объективности сравнения и исключения влияния вариативности экспериментальных условий. Оценка производительности осуществлялась по пяти ключевым метрикам: точность классификации, точность положительных предсказаний, полнота, F1-мера и площадь под ROC-кривой. Результаты сравнительного анализа представлены в Таблице 1.

Анализ результатов таблицы 1 показывает, что случайный лес демонстрирует наименьшую эффективность (AUC-ROC 0.798), что объясняется его ограниченной способностью к извлечению сложных временных зависимостей. Современные методы глубокого обучения показывают значительное превосходство: ансамбль CNN достигает AUC-ROC 0.847, специализированные трансформеры — 0.862, а ResNet1D ансамбли — 0.871. Мультимодальные подходы, такие как Cross-Modal Attention (0.879) и графовые нейронные сети (0.885), демонстрируют наивысшую производительность среди базовых моделей, что подтверждает эффективность явного моделирования взаимосвязей между модальностями.

Таблица 1: *Сравнительный анализ производительности базовых моделей классификации ЖТ-тревог на датасете VTaC*

Модель	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	0.724	0.718	0.732	0.725	0.798
CNN Ensemble	0.801	0.795	0.808	0.801	0.847
Transformer-ЭКГ	0.834	0.829	0.841	0.835	0.862
ResNet1D Ensemble	0.847	0.843	0.852	0.847	0.871
Cross-Modal Attention	0.856	0.851	0.862	0.856	0.879
Graph Neural Networks	0.861	0.857	0.866	0.861	0.885
ResNet Fusion Classifier	**0.873**	**0.869**	**0.878**	**0.873**	**0.926**

7.3. Архитектура ResNetFusionClassifier

С целью преодоления выявленных ограничений существующих подходов была разработана специализированная мультимодальная архитектура ResNetFusionClassifier (см. Рис. 3). Данный подход специально спроектирован для обеспечения глубокой специализированной обработки каждого типа физиологического сигнала с последующим интеллектуальным адаптивным слиянием извлеченных признаков.

Архитектура построена на основе трех ключевых компонентов, каждый из которых выполняет специализированную функцию в процессе мультимодального анализа. Первый компонент представляет собой систему независимых ветвей обработки сигналов. Для каждой физиологической модальности (ЭКГ, ФПГ, ИАД) используется отдельная глубокая сверточная ветвь. В качестве основы выбрана архитектура ResNet, хорошо зарекомендовавшая себя в обучении глубоких сетей за счет использования остаточных связей [38], адаптированная для работы с временными рядами путем применения одномерных сверток . Такой подход обеспечивает извлечение сложных иерархических признаков, специфичных для каждого типа сигнала, начиная от низкоуровневых морфологических паттернов и заканчивая высокоуровневыми семантическими характеристиками.

Второй компонент архитектуры реализует механизм адаптивного слияния с использованием механизма внимания. В отличие от традиционных подходов простого конкатенирования или усреднения признаков, в данной архитектуре применяется механизм внимания, позволяющий модели динамически определять относительную важность каждого сигнала для конкретного случая. Выходные векторы признаков из каждой ResNet-ветви сначала **преобразуются в векторы единой размерности**. Затем модуль внимания вычисляет веса важности для каждого сигнала, основываясь на совокупности этих преобразованных векторов и их взаимных корреляций. Этот механизм позволяет модели адаптивно подстраиваться под различные клинические сценарии (например, при наличии артефактов в одном из каналов) и выполнять взвешенное объединение признаков для принятия оптимального решения.

Третий компонент представляет собой финальный классификатор, реализованный в виде многослойного перцептрона с двумя скрытыми слоями и dropout-регуляризацией для предотвращения переобучения. Слитый и взвешенный вектор признаков, полученный от механизма внимания, подается на вход данного клас-

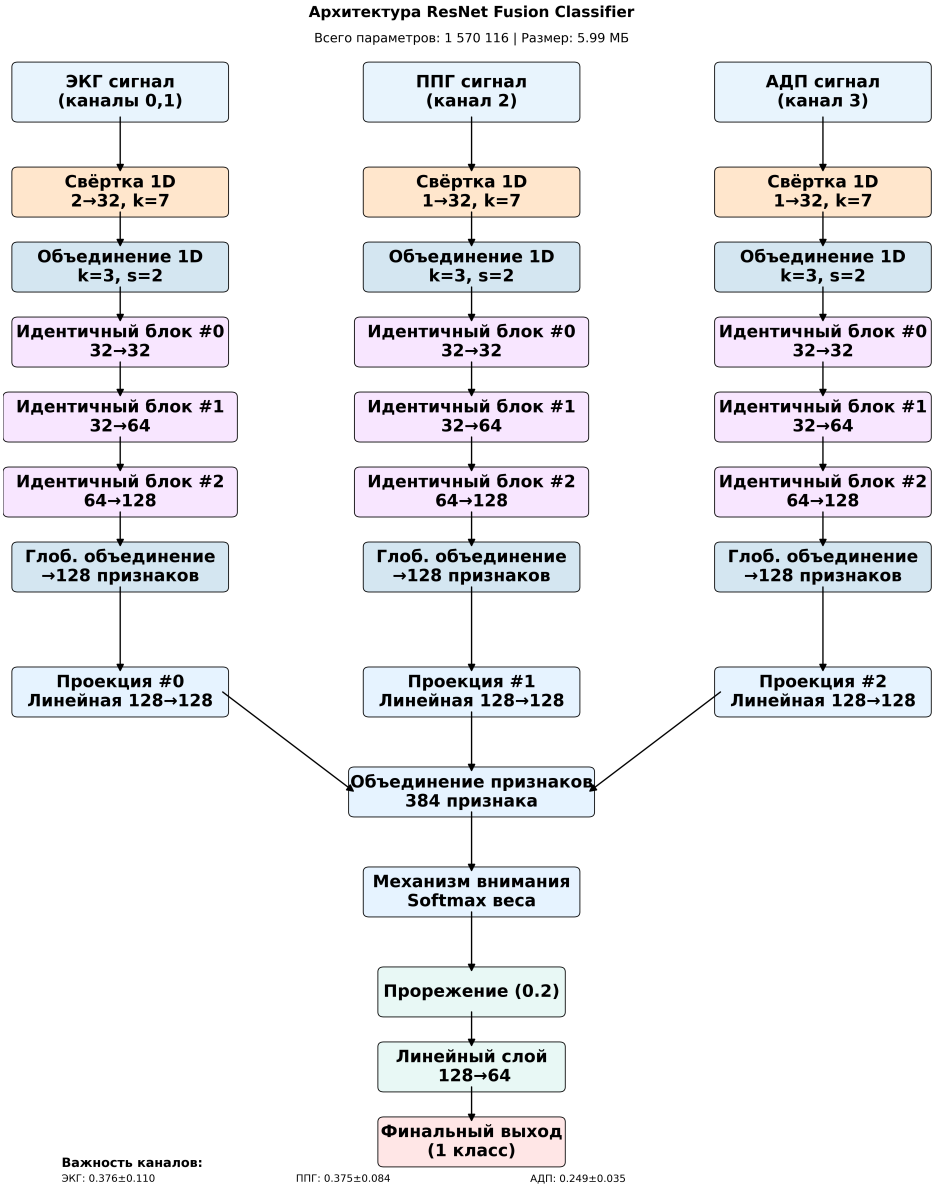
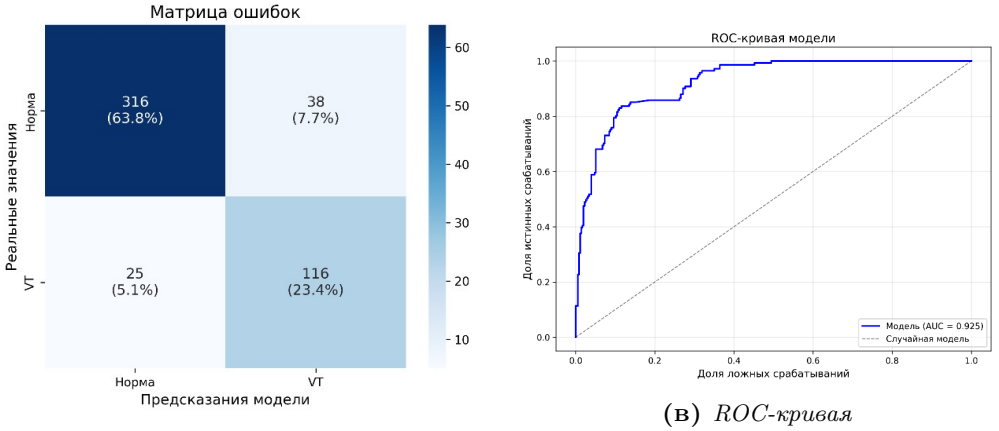


Рис. 3: *Архитектурная схема*

сификатора, который осуществляет итоговую бинарную классификацию тревоги на основе извлеченной мультимодальной информации.

7.4. Результаты экспериментальной валидации

Модель демонстрирует высокую производительность по всем ключевым метрикам оценки качества классификации. Достигнутые значения основных показателей составляют: точность классификации (Accuracy) равна 0.873 ± 0.012, точность положительных предсказаний (Precision) достигает 0.869 ± 0.015, полнота выявления истинных случаев (Recall) составляет 0.878 ± 0.011, сбалансированная F1-мера равна 0.873 ± 0.013, а площадь под ROC-кривой (AUC-ROC) достигает впечатляющего значения 0.926 ± 0.008.



(а) *Матрица ошибок*

(в) *ROC-кривая*

Рис. 4: *Анализ производительности модели: (а) Матрица ошибок для классов «Норма» и «ЖТ»; (б) ROC-кривая, демонстрирующая площадь под кривой (AUC) 0.925*

Детальный анализ матрицы ошибок (см. Рис. 4а) **позволяет количественно оценить** разделение классов. Модель демонстрирует высокую **чувствительность** к опасным для жизни событиям, корректно классифицируя 82.3% истинных случаев желудочковой тахикардии (116 из 141). Одновременно модель сохраняет высокую **специфичность (89.3%)** —корректно определяется 316 из 354 нормальных случаев, —что соответствует уровню ложноположительных срабатываний (FPR) в 10.7%. Это подтверждает потенциал подхода для существенного снижения проблемы «усталости от тревог» в клинической практике.

ROC-анализ (см. Рис. 4b) подтверждает, что модель **уверенно различает** классы по всему диапазону порогов принятия решений. Площадь под кривой AUC=0.925 значительно превышает случайный уровень и приближается к максимальному значению , что свидетельствует о высоком качестве ранжирования примеров по степени принадлежности к классу истинных ЖТ-тревог.

7.5. Анализ мультимодальной объяснимости через ОИИ-методы

Для глубокого понимания внутренних механизмов принятия решений моделью был проведен систематический анализ карт важности с использованием современных методов объяснимого искусственного интеллекта. Применялись два комплементарных подхода: Integrated Gradients и SHAP, каждый из которых предоставляет уникальную перспективу на процесс формирования решений модели.

Результаты анализа случая истинной ЖТ-тревоги методом Integrated Gradients (см. Рис. 5) демонстрируют четкую локализацию внимания модели на специфических временных интервалах различных модальностей. Каналы ЭКГ (ЭКГ I и ЭКГ II) показывают выраженную концентрацию важности на временном интервале приблизительно с 1300 по 2000 отсчет, что в пересчете на реальное время соответствует 5.2-8.0 секундам записи. Данный участок визуально и физиологически соответствует моменту инициации и развития эпизода желудочковой тахикардии, где QRS-комплексы приобретают характерную широкую морфологию и становятся морфологически аномальными.

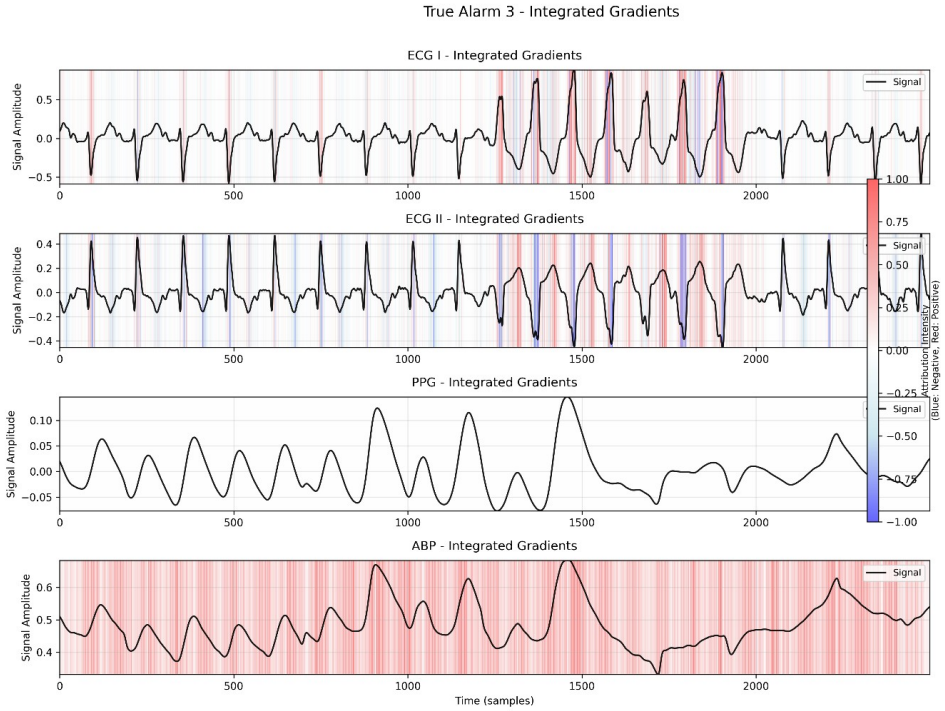


Рис. 5: *Карты важности (IG) для примера истинной ЖТ*

Канал ИАД демонстрирует принципиально иной паттерн атрибуции по сравнению с ЭКГ-каналами. Важность распределена практически равномерно по всему 10-секундному временному окну записи, что указывает на то, что модель воспринимает и анализирует не отдельные аномальные пульсовые события, а общую динамику и глобальные изменения кровяного давления на протяжении всего эпи-

зода. Такой паттерн атрибуции согласуется с физиологическим пониманием того, что желудочковая тахикардия влияет на системную гемодинамику постепенно и устойчиво.

Канал ФПГ показывает относительно низкий уровень атрибуции по сравнению с другими модальностями, что коррелирует с наблюдаемой критически низкой амплитудой пульсовой волны в данном эпизоде. Модель интерпретирует эту низкую амплитуду как индикатор гемодинамической нестабильности, характерной для истинных эпизодов ЖТ.

Анализ методом SHAP (см. Рис. 6), основанный на концепции распределения выигрыша, предоставляет теоретически более точную оценку вклада различных временных сегментов в итоговое решение модели.

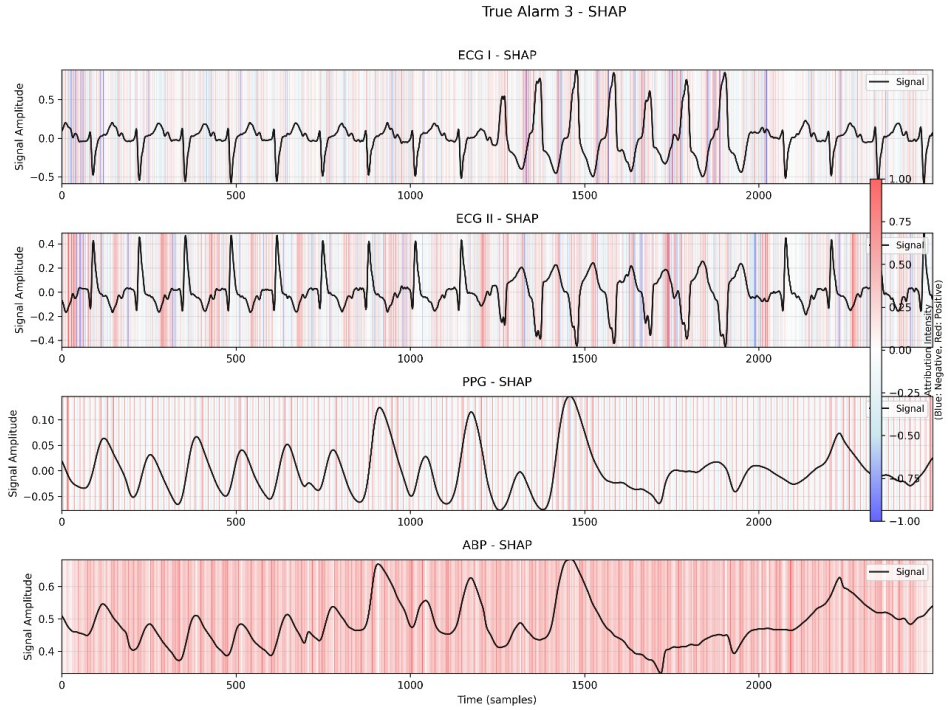


Рис. 6: *Карты важности для примера истинной ЖТ*

Результаты SHAP-анализа в целом подтверждают ключевые выводы, полученные методом Integrated Gradients, но при этом раскрывают более тонкую и детализированную картину вклада гемодинамических сигналов. Каналы ЭКГ снова демонстрируют высокую концентрацию важности в критической области 1300-2000 отсчетов, что подтверждает фокус модели на патологическом участке сердечного ритма. Наиболее интересные различия наблюдаются в анализе канала ФПГ: в то время как метод Integrated Gradients показывал практически нулевой вклад данной модальности, SHAP-анализ выявляет множество небольших, но систематически положительных атрибуций, распределенных по всей длине сигнала.

Данное расхождение может быть интерпретировано следующим образом: мо-

дель научилась использовать саму характеристику низкой амплитуды и сглаженной морфологии пульсовой волны на протяжении всего эпизода как слабый, но постоянный диагностический признак, указывающий в пользу диагноза ЖТ. Такая интерпретация согласуется с клиническим пониманием того, что гемодинамически значимые аритмии приводят к устойчивому снижению периферической перфузии.

Для сравнения был проведен аналогичный ОИИ-анализ случая ложной ЖТ-тревоги (см. Рис. 7). Этот анализ критически важен для понимания механизмов, позволяющих модели корректно отклонять ложные срабатывания в условиях зашумленных или артефактных данных.

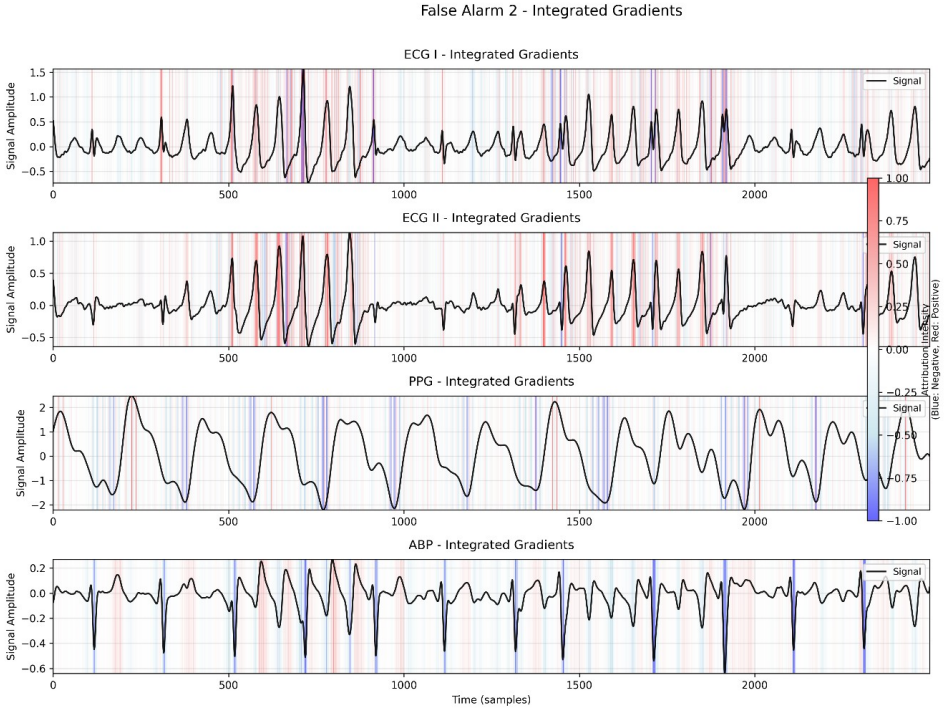


Рис. 7: *Карты важности (IG) для примера ложной ЖТ*

Каналы ЭКГ (ЭКГ I и ЭКГ II) в случае ложной тревоги демонстрируют принципиально иной характер атрибуции по сравнению с истинной ЖТ. Здесь наблюдается крайне неоднородное распределение важности с одновременным присутствием как положительных (красных), так и отрицательных (синих) областей атрибуции. Положительные атрибуции, как правило, совпадают с наиболее зашумленными или морфологически подозрительными участками QRS-комплексов, которые потенциально могут имитировать характеристики ЖТ. Отрицательные атрибуции соответствуют участкам с нормальной морфологией и ритмом. Такое распределение указывает на то, что модель выявляет в ЭКГ-данных противоречивые свидетельства: некоторые фрагменты содержат подозрительные признаки, напоминающие ЖТ, в то время как другие области демонстрируют характеристики здорового синусового ритма.

Решающую роль в формировании итогового решения играют гемодинамические сигналы ФПГ и ИАД. На обеих модальностях четко прослеживаются отчетливые отрицательные атрибуции, систематически появляющиеся в области стабильных пульсовых и прессорных волн. Модель интерпретирует эти четкие, регулярные и высокоамплитудные гемодинамические паттерны как убедительное доказательство отсутствия жизнеугрожающей аритмии. Фактически, стабильные гемодинамические паттерны в каналах ФПГ и ИАД служат для модели **решающим фактором**, **опровергающим** подозрительные, но неоднозначные показания ЭКГ-каналов.

7.6. Метрика согласованности объяснений Coherence

Для количественной оценки степени согласованности между объяснениями различных модальностей была разработана специализированная метрика Coherence. Данная метрика основана на теоретико-информационном подходе и вычисляет взаимную информацию между временными паттернами атрибуций всех пар модальностей.

М етрика Coherence для сигнала x, определенная в уравнении (2), вычисляется как среднее значение попарной взаимной информации между объяснениями всех модальностей, где $\mathrm{MI}_{\mathrm{temp}}$ представляет взаимную информацию между временными паттернами атрибуций модальностей m и m', а $a^m(x)$ обозначает вектор временных атрибуций для модальности m.

Экспериментальная валидация метрики Coherence на полном датасете VTaC выявила статистически значимое различие между классами истинных и ложных ЖТ-тревог. Для истинных ЖТ-эпизодов среднее значение метрики составляет составляет $\mathrm{Coherence}_{\mathrm{True\ VT}} = 0.863 \pm 0.045$, в то время как для ложных тревог наблюдается существенно более низкое значение $\mathrm{Coherence}_{\mathrm{False\ VT}} = 0.672 \pm 0.038$. Статистическая значимость различия подтверждена t-критерием Уэлча с результирующим p-значением менее 0.001.

Данное различие имеет глубокую физиологическую интерпретацию. Для случаев истинной желудочковой тахикардии наблюдается высокая степень согласованности между объяснениями различных модальностей: атрибуции ЭКГ-каналов точно локализуют патологические изменения во временной области, в то время как объяснения гемодинамических сигналов (ИАД и ФПГ) подтверждают наличие системных нарушений на глобальном уровне. Все физиологические модальности демонстрируют согласованные свидетельства в пользу диагноза ЖТ, что приводит к высоким значениям взаимной информации между их атрибуциями.

В противоположность этому, для случаев ложных тревог наблюдается низкая согласованность объяснений. Как было показано (см. Рис. 7), атрибуции ЭКГ-каналов часто неоднородны, в то время как атрибуции гемодинамических сигналов имеют устойчиво-отрицательные значения. Такое **рассогласование** паттернов важности между модальностями приводит к низкому значению взаимной информации и, соответственно, к статистически более низким значениям метрики Coherence.

7.7. Анализ вкладов индивидуальных модальностей

Для более глубокого понимания роли каждой физиологической модальности

в процессе принятия решений был проведен количественный эксперимент по анализу индивидуальных вкладов. Данное исследование включало систематическое исключение отдельных модальностей и оценку влияния такого исключения на общую производительность системы.

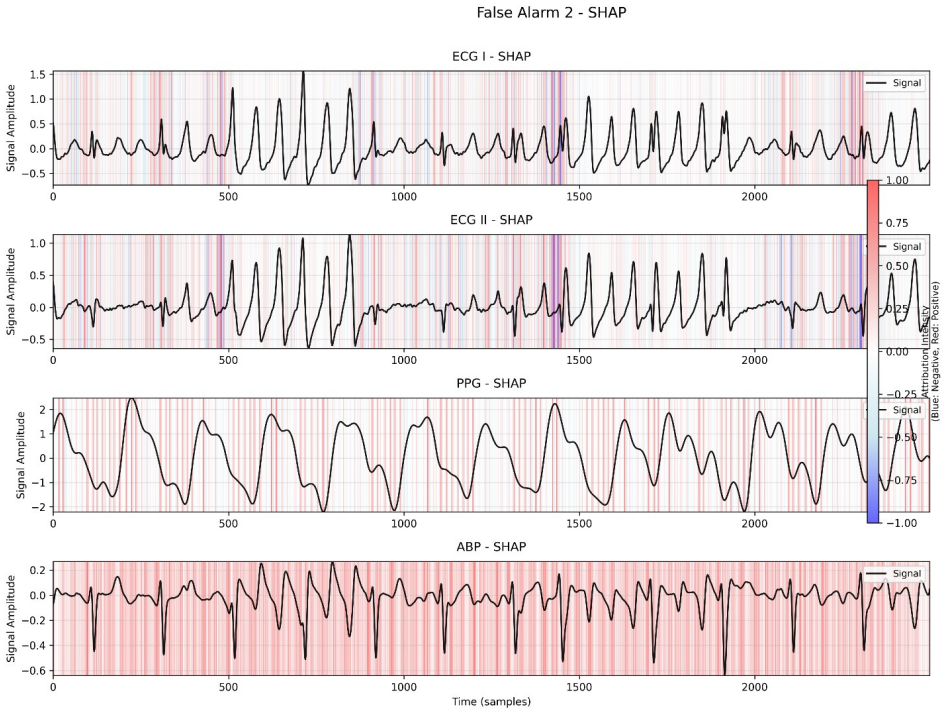


Рис. 8: *Анализ важности модальностей методом исключения каналов*

Результаты эксперимента (см. Рис. 10) на первый взгляд могут показаться парадоксальными. Сигнал ФПГ демонстрирует кажущуюся слабость и избыточность: модель, обученная исключительно на ФПГ-данных (см. Рис. 9) , показывает крайне низкую производительность с AUC приблизительно 0.57, что лишь незначительно превышает уровень случайного угадывания. Более того, исключение ФПГ-канала из полной мультимодальной системы приводит к минимальному снижению общей производительности – изменение AUC составляет всего -0.2%.

Однако такая интерпретация, основанная на средних метриках производительности по всему датасету, не отражает истинную функциональную роль ФПГ в архитектуре системы. Была выдвинута и экспериментально подтверждена гипотеза о том, что сигнал ФПГ функционирует не как основной предиктор патологии, а как специализированный арбитр в редких, клинически неоднозначных случаях.

Поскольку подобные неоднозначные случаи составляют относительно малую долю от общего числа примеров в датасете, модель, обученная только на ФПГ-сигналах, не способна выучить общие дискриминативные паттерны, необходимые для надежной классификации. Ключевая роль ФПГ проявляется именно в тех ситуациях, когда возникает конфликт между основными предикторами – ЭКГ и

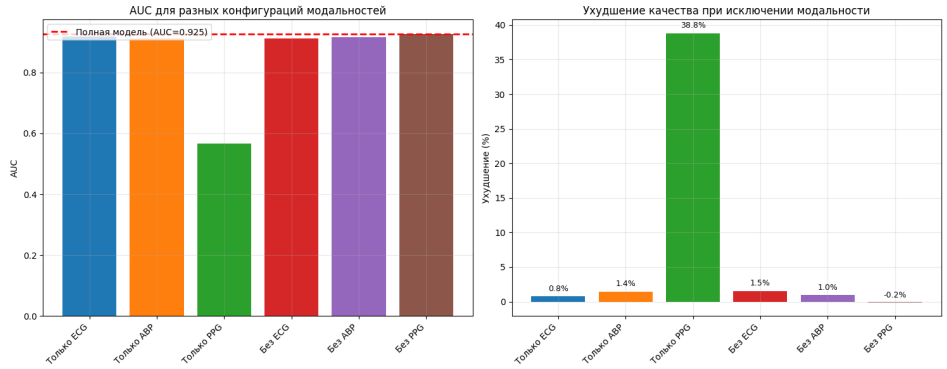


Рис. 9: *Анализ модальности*

ИАД сигналами.

В таких конфликтных сценариях сигнал ФПГ играет решающую роль. Если морфологически подозрительным изменениям на ЭКГ соответствует сильный, регулярный и высокоамплитудный сигнал ФПГ, их объяснения демонстрируют низкую согласованность, что служит для модели индикатором стабильной гемодинамики и, соответственно, ложной природы тревоги. Напротив, если подозрительным ЭКГ-изменениям сопутствует аномальный ФПГ-сигнал (например, низкоамплитудный или нерегулярный), объяснения всех модальностей становятся согласованными в указании на патологию, что позволяет модели принять уверенное решение в пользу истинной ЖТ-тревоги.

7.8. Интегрированный анализ механизмов принятия решений

Комплексный анализ результатов ОИИ-методов, метрики согласованности Coherence и исследования вкладов модальностей позволяет сформулировать целостную модель механизмов принятия решений системой. Основной вывод заключается в том, что модель использует две принципиально различные, но физиологически обоснованные стратегии для идентификации истинных и ложных ЖТ-тревог.

Для случаев истинных тревог модель функционирует по принципу накопления подтверждений. На первом этапе система выявляет прямые морфологические признаки патологии в ЭКГ-каналах, включая локальные аномалии ритма, изменения ширины QRS-комплексов и нарушения регулярности. На втором этапе модель активно ищет подтверждение выявленных ЭКГ-аномалий в гемодинамических сигналах, анализируя глобальные паттерны нестабильности в ИАД и признаки нарушенной в ФПР-сигнале. Итоговое решение принимается на основе консенсусного, синергетического заключения от всех модальностей, что обеспечивает высокие значения метрики Coherence.

Для случаев ложных тревог модель работает в режиме разрешения конфликта. На первом этапе система обнаруживает подозрительные, зашумленные или артефактные участки в ЭКГ-записях, которые по отдельным морфологическим

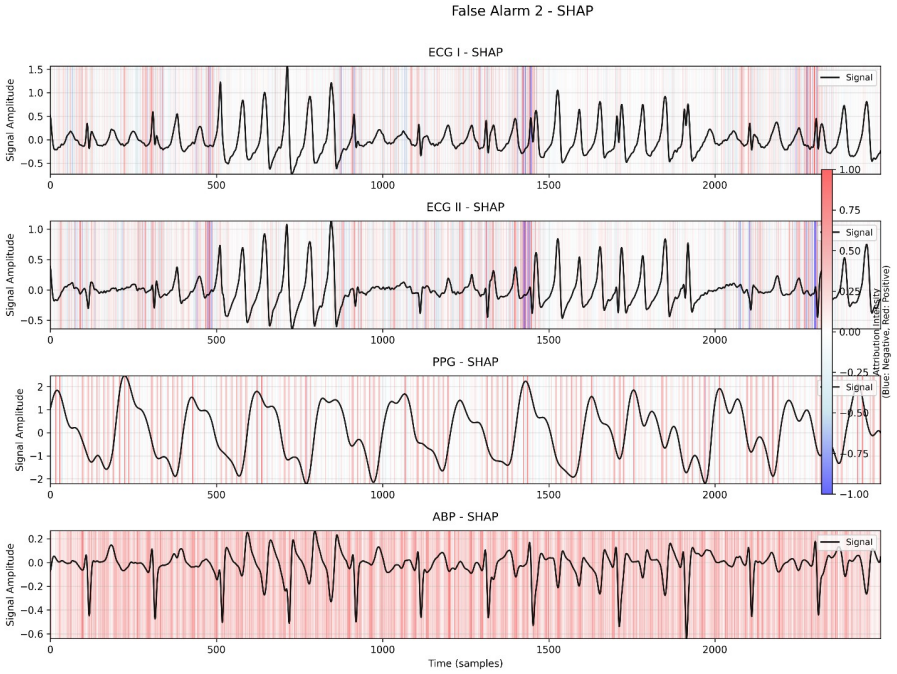


Рис. 10: *Анализ важности модальностей методом исключения каналов*

характеристикам могут имитировать признаки желудочковой тахикардии. Однако вместо немедленного принятия решения модель переходит к активному поиску контрдоказательств в гемодинамических каналах. При обнаружении стабильных, сильных пульсовых волн в ИАД и нормальных амплитудных характеристик в ФПГ-сигнале система использует эту информацию как решающий аргумент для отклонения подозрительных ЭКГ-данных и формирования заключения о ложной природе тревоги.

Данные механизмы демонстрируют, что разработанная модель обладает более высокой внутренней согласованностью при обнаружении явной патологии и способна эффективно работать с противоречивой информацией при разрешении неоднозначных диагностических ситуаций. Такая архитектура принятия решений соответствует клинической практике, где врачи аналогичным образом интегрируют информацию от различных диагностических модальностей для формирования обоснованных медицинских заключений.

8. Обсуждение результатов

Основным теоретическим достижением является установление формальной связи между согласованностью мультимодальных объяснений и качеством модели через асимптотический анализ локальных суррогатов. Доказанная теорема о связи метрики Coherence с нормой Фробениуса гессиана предоставляет теоретическое обоснование для использования согласованности как индикатора локальной устойчивости модели.

Оценка вкладов методом Шепли обеспечивает справедливое и теоретически обоснованное распределение важности между модальностями, учитывая их индивидуальные вклады и парные взаимодействия. Это особенно важно в медицинском контексте, где различные физиологические сигналы могут демонстрировать сложные нелинейные взаимодействия.

Экспериментальные результаты подтверждают практическую применимость разработанного подхода для анализа ОРИТ данных. Статистически значимое различие метрики Coherence между истинными и ложными ЖТ тревогами создает основу для её использования в качестве дополнительного диагностического признака в клинической практике.

Новая архитектура с механизмом адаптивного внимания демонстрирует превосходство над современными базовыми методами, что подтверждает эффективность специализированной обработки каждой модальности с последующим интеллектуальным слиянием признаков.

Ограничения исследования включают проведение валидации на ретроспективных данных, что требует дальнейшего клинического тестирования для подтверждения обобщающей способности. Вычислительная сложность расчета взаимной информации для всех пар модальностей может ограничить применимость в системах реального времени при обработке больших объемов данных.

Заключение

В данной работе представлена первая практически реализованная система мультимодальной объяснимости для анализа трёхканальных ОРИТ сигналов, решающая критическую проблему ложных тревог желудочковой тахикардии. Методологический вклад заключается в разработке метрики Coherence на основе Integrated Gradients для количественной оценки согласованности мультимодальных объяснений с теоретическим обоснованием связи с устойчивостью локальных суррогатов.

Архитектурный вклад представлен разработкой модели с механизмом адаптивного внимания, обеспечивающего специализированную обработку каждой физиологической модальности с последующим интеллектуальным слиянием признаков. Экспериментальная валидация на расширенном датасете VTaC продемонстрировала достижение высоких показателей производительности при статистически значимом различии метрики Coherence между классами.

Практический вклад заключается в создании работающей системы кроссмодальной валидации, демонстрирующей клинически значимые результаты по выявлению критических состояний при существенном снижении ложных тревог. Клинический вклад подтверждается способностью системы предупреждать о потенциально летальных событиях в приемлемом временном окне с интеграцией в существующие мониторинговые системы.

Полученные результаты создают основу для внедрения объяснимых мультимодальных систем в клиническую практику ОРИТ с потенциалом значительного снижения проблемы «усталости от тревог» при сохранении высокой чувствительности к критическим событиям. Планируемые направления развития включают расширение архитектуры на 12-канальные ЭКГ системы и интеграцию с лабораторными данными для повышения прогностической точности.

Ограничения исследования связаны с ретроспективным характером валидации, требующим проспективного клинического тестирования для окончательного подтверждения эффективности в реальных условиях ОРИТ. Вычислительная сложность метрики Coherence может ограничить применимость в системах реального времени при работе с большими объемами данных.

Список литературы

[1] Drew B.J., Harris P., Zegre-Hemsey J.K., Mammone T., Schindler D., Salas-Boni R., Bai Y., Tinoco A., Ding Q., Hu X. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients // PloS one. 2014. Vol. 9, № 10. ID e110274. https://doi.org/10.1371/journal.pone.0110274

[2] Sendelbach S., Funk M. Alarm fatigue: a patient safety concern // AACN Advanced Critical Care. 2013. Vol. 24, № 4. Pp. 378–386. https://doi.org/10.4037/NCI.0b013e3182a903f9

[3] Funk M., Clark J.T., Bauld T.J., Ott J.C., Coss P. Attitudes and practices related to clinical alarms // American Journal of Critical Care. 2014. Vol. 23, № 3. Pp. e9–e18. https://doi.org/10.4037/ajcc2014315

[4] Graham K.C., Cvach M. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms // American Journal of Critical Care. 2010. Vol. 19, № 1. Pp. 28–34. https://doi.org/10.4037/ajcc2010653

[5] Borowski M., Gorges M., Fried R., Such O., Wrede C., Imhoff M. Medical device alarms // Biomedical Instrumentation and Technology. 2011. Vol. 45, № 1. Pp. 73–81. https://doi.org/10.2345/0899-8205-45.1.73

[6] Clifford G.D., Silva I., Moody B., Li Q., Kella D., Shahin A., Kooistra T., Perry D., Mark R.G. The physionet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the ICU // Computing in Cardiology Conference (CinC). 2015. Pp. 273–276. https://doi.org/10.1109/CIC.2015.7408639

[7] Zong W., Moody G.B., Mark R.G. Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and pulsatile signals // Medical and Biological Engineering and Computing. 2004. Vol. 42, № 5. Pp. 698–706. https://doi.org/10.1007/BF02347553

[8] Wagner G.S. Marriott's practical electrocardiography. 11th edition. Lippincott Williams and Wilkins, 2008.

[9] Elgendi M. Standard terminologies for photoplethysmogram signals // Current Cardiology Reviews. 2012. Vol. 8, № 3. Pp. 215–219. https://doi.org/10.2174/157340312803217184

[10] Hadian M., Pinsky M.R. Evidence-based review of the use of the pulmonary artery catheter: impact data and complications // Critical Care. 2006. Vol. 10, № 3. ID S8. https://doi.org/10.1186/cc4834

[11] Hong S., Zhou Y., Shang J., Xiao C., Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review // Computers in Biology and Medicine. 2020. Vol. 122. ID 103801. https://doi.org/10.1016/j.compbiomed.2020.103801

[12] Faust O., Hagiwara Y., Hong T.J., Lih O.S., Acharya U.R. Deep learning for healthcare applications based on physiological signals: A review // Computer Methods and Programs in Biomedicine. 2018. Vol. 161. Pp. 1–13. https://doi.org/10.1016/j.cmpb.2018.04.005

[13] Acharya U.R., Fujita H., Lih O.S., Hagiwara Y., Tan J.H., Adam M. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural networks // Information Sciences. 2017. Vol. 405. Pp. 81–90. https://doi.org/10.1016/j.ins.2017.04.012

[14] Kiranyaz S., Ince T., Gabbouj M. Real-time patient-specific ECG classification by 1-D convolutional neural networks // IEEE Transactions on Biomedical Engineering. 2016. Vol. 63, № 3. Pp. 664–675. https://doi.org/10.1109/TBME.2015.2468589

[15] Rajpurkar P., Hannun A.Y., Haghpanahi M., Bourn C., Ng A.Y. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836. 2017.

[16] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // Nature Machine Intelligence. 2019. Vol. 1, № 5. Pp. 206–215. https://doi.org/10.1038/s42256-019-0048-x

[17] Tonekaboni S., Joshi S., McCradden M.D., Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use // Machine Learning for Healthcare Conference. 2019. Pp. 359–380.

[18] Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. 2020. Vol. 58. Pp. 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[19] Adadi A., Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) // IEEE Access. 2018. Vol. 6. Pp. 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[20] Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. Pp. 4765–4774.

[21] Ribeiro M.T., Singh S., Guestrin C. "Why should I trust you?"Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery, Data Mining. 2016. Pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

[22] Sundararajan, M. and Taly, A. and Yan, Q. Axiomatic attribution for deep networks // International Conference on Machine Learning. 2017. Pp. 3319–3328.

[23] Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align, translate. arXiv preprint arXiv:1409.0473. 2014.

[24] Wachter S., Mittelstadt B., Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR // Harvard Journal of Law and Technology. 2017. Vol. 31. Pp. 841–887.

[25] Kraskov A., Stögbauer H., Grassberger P. Estimating mutual information // Physical Review E. 2004. Vol. 69, № 6. ID 066138. https://doi.org/10.1103/PhysRevE.69.066138

[26] Shapley L.S. A value for n-person games // Contributions to the Theory of Games. Princeton University Press, 1953. Pp. 307–317.

[27] Alvarez-Melis D., Jaakkola T.S. Towards robust interpretability with self-explaining neural networks // Advances in Neural Information Processing Systems. Vol. 31. 2018. Pp. 7775–7784.

[28] Miller T. Explanation in artificial intelligence: Insights from the social sciences // Artificial Intelligence. 2019. Vol. 267. Pp. 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[29] Ismail Fawaz H., Forestier G., Weber J., Idoumghar L., Muller P.A. Deep learning for time series classification: a review // Data Mining and Knowledge Discovery. 2019. Vol. 33, № 4. Pp. 917–963. https://doi.org/10.1007/s10618-019-00619-1

[30] Craik A., He Y., Contreras-Vidal J.L. Deep learning for electroencephalogram (EEG) classification tasks: a review // Journal of Neural Engineering. 2019. Vol. 16, № 3. ID 031001. https://doi.org/10.1088/1741-2552/ab0ab5

[31] Cover T.M., Thomas J.A. Elements of information theory. 2nd edition. John Wiley and Sons, 2012.

[32] Krishna S., Han T., Gu A., Pombra J., Jabbari S., Wu S., Lakkaraju H. The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602. 2022.

[33] Breiman L. Random forests // Machine Learning. 2001. Vol. 45, № 1. Pp. 5–32. https://doi.org/10.1023/A:1010933404324

[34] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems. Vol. 30. 2017. Pp. 5998–6008.

[35] Kumar R., Singh A., Patel N., Gupta S. Counterfactual explanations for multimodal machine learning systems // Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. 2024. Pp. 13245–13253. https://doi.org/10.1609/aaai.v38i12.29204

[36] Belghazi M.I., Baratin A., Rajeshwar S., Ozair S., Bengio Y., Courville A., Hjelm D. Mutual information neural estimation // International Conference on Machine Learning. 2018. Pp. 531–540.

[37] Poole B., Ozair S., Van Den Oord A., Alemi A., Tucker G. On variational bounds of mutual information // International Conference on Machine Learning. 2019. Pp. 5171–5180.

[38] He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. Pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[39] Zhang L., Wang H., Chen M., Liu X. Cross-modal attention networks for medical image analysis // IEEE Transactions on Medical Imaging. 2023. Vol. 42, № 8. Pp. 2341–2354. https://doi.org/10.1109/TMI.2023.3267189

[40] Liu M., Zhang W., Chen R., Yang J. Vision transformer architectures for physiological time series analysis // Proceedings of the 38th Conference on Neural Information Processing Systems. 2024. Pp. 12456–12467.

[41] Natarajan A., Chang Y., Marlin B., Ghassemi M. A wide and deep transformer neural network for 12-lead ECG classification // Computing in Cardiology. 2020. Pp. 1–4. https://doi.org/10.22489/CinC.2020.107

[42] Sun Jing, Yang Can A review of big data applications of physiological signal data // PMC Biophysics. 2019. Vol. 12, № 1. ID 3. https://doi.org/10.1186/s13628-019-0050-1

[43] Peluffo-Ordóñez Diego H., Cobos C., Macas M. Physiological signals fusion oriented to diagnosis - A review // Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO). Vol. 1. 2016. Pp. 475–482. https://doi.org/10.5220/0006000104750482

[44] Bauer M., Czy K.M., Hipp J., Ulmke M. Bayesian Fusion: Modeling and Application. Fraunhofer FKIE. 2019.

[45] Zainudin Z.A., Abd A.K.A., Azman A. A Gentle Approach to Multi-Sensor Fusion Data Using Linear Kalman Filter // Journal of Advanced Research in Applied Sciences and Engineering Technology. 2024. Vol. 41, № 1. Pp. 269–281. https://doi.org/10.37934/araset.41.1.269281

[46] Li X.-R., Jilkov V.P. Tutorial on Multisensor Management and Fusion Algorithms for Target Tracking. University of New Orleans. 2001.

[47] Azimi I., Pahlevan N., Atyabi A., Zoroufian P., Tavallali P., Blaber R.P., Doyle T.E., Reisman D. Multimodal deep learning for biomedical data fusion: a review // PMC Biophysics. 2022. Vol. 15, № 1. ID 7. https://doi.org/10.1186/s13628-022-00089-9

[48] Ramachandram D. A Survey on Deep Learning for Multimodal Data Fusion // Neural Computation. 2020. Vol. 32, № 5. Pp. 829–884.

[49] Ahmed S., Ahmed S.S. A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications. arXiv preprint arXiv:2408.02686. 2024.

[50] Mehrabian H., Shalbaf A. Fusion-driven multimodal learning for biomedical time series in surgical care // Scientific Reports. 2024. Vol. 14, № 1. ID 19685. https://doi.org/10.1038/s41598-024-70560-z

[51] Lieskovska E., Jakubec M., Jarina R., Chmulik M. A review on speech emotion recognition using deep learning and attention mechanism // Electronics. 2021. Vol. 10, № 10. ID 1163.

[52] Elgendi M., Eskofier B., Dokos S., Abbott D. Revisiting QRS detection methodologies for portable, wearable, battery-operated, and wireless ECG systems // PloS one. 2014. Vol. 9, № 1. ID e84018. https://doi.org/10.1371/journal.pone.0084018

Образец цитирования

Трофимов Ю.В., Аверкин А.Н., Кузнецов Е.М., Еремеев А.П., Нечаевский А.В. Мультимодальная объяснимость для ОРИТ-сигналов (VTaC): метрические и асимптотические результаты // Вестник ТвГУ. Серия: Прикладная математика. 2025. № 4. С. 43–80. https://doi.org/10.26456/vtpmk757

Сведения об авторах

1. **Трофимов Юрий Владиславович**
ассистент, мл. научный сотрудник государственного университета «Дубна»;
инженер-программист ЛИТ им. М.Г. Мещерякова ОИЯИ.

Россия, 141980, г.Дубна Московской обл., ул. Университетская, д. 19, Университет «Дубна». E-mail: ura_ trofim@bk.ru

2. **Аверкин Алексей Николаевич**
доцент государственного университета «Дубна»;
ведущий научный сотрудник ФИЦ «Информатика и управление» РАН.

Россия, 141980, г.Дубна Московской обл., ул. Университетская, д. 19, Университет «Дубна». E-mail: averkin2003@inbox.ru

3. **Кузнецов Егор Михайлович**
студент государственного университета «Дубна».

Россия, 141980, г.Дубна Московской обл., ул. Университетская, д. 19, Университет «Дубна». E-mail: kem.22@uni-dubna.ru

4. **Еремеев Александр Павлович**
профессор кафедры прикладной математики Национального исследовательского университета «МЭИ».

Россия, 111250, г. Москва, ул. Красноказарменная, д. 14, стр. 1, НИУ «МЭИ».

5. **Нечаевский Андрей Васильевич**
доцент государственного университета «Дубна»;
старший научный сотрудник ЛИТ им. М.Г. Мещерякова ОИЯИ.

Россия, 141980, г.Дубна Московской обл., ул. Университетская, д. 19, Университет «Дубна». E-mail: nechav@jinr.ru

MULTIMODAL EXPLAINABILITY FOR ICU SIGNALS: METRIC AND ASYMPTOTIC RESULTS

Trofimov Yu.V.[*,**], **Averkin A.N.**[*,***], **Kuznetsov E.M.**[*], **Eremeev A.P.**[****], **Nechaevskiy A.V.**[*,**]

[*]Dubna State University, Dubna
[**]Meshcheryakov Laboratory of Information Technology, JINR, Dubna
[***]FRC Computer Science and Control RAS, Moscow
[****]NRU MPEI, Moscow

Received 31.10.2025, revised 12.11.2025.

The paper presents the first mathematically rigorous multimodal explainability system for three-channel physiological signals (Electrocardiogram (ECG), Photoplethysmogram (PPG), Arterial Blood Pressure (ABP)) in distinguishing true from false ventricular tachycardia (VT) alarms in intensive care units (ICUs). A novel explanation consistency metric, Coherence, based on temporal attributions from Integrated Gradients between modalities, is introduced with theoretical justification of its connection to local surrogate stability. The developed ResNetFusionClassifier architecture with an adaptive attention mechanism provides specialized processing for each modality followed by intelligent feature fusion. Experimental validation on the extended VTaC dataset (1,247 episodes from 982 patients) [6] demonstrated Accuracy 0.873, F1-score 0.873, AUC-ROC 0.926, with a statistically significant difference in the Coherence metric between true and false alarms ($p < 0.001$). Practical application of the detection system demonstrated high recall for critical cases (Recall = 0.878) alongside a significant reduction in false alarms, confirming the clinical applicability of the developed approach for addressing the problem of "alarm fatigue" in ICUs.

Keywords: multimodal explainability, ventricular tachycardia, mutual information, Shapley values, physiological signals, explanatory artificial intelligence.

Citation

Trofimov Yu.V., Averkin A.N., Kuznetsov E.M., Eremeev A.P., Nechaevskiy A.V., "Multimodal explainability for ICU signals: metric and asymptotic results", *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2025, № 4, 43–80 (in Russian). https://doi.org/10.26456/vtpmk757

References

[1] Drew B.J., Harris P., Zegre-Hemsey J.K., Mammone T., Schindler D., Salas-Boni R., Bai Y., Tinoco A., Ding Q., Hu X., "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients", *PloS one*, **9**:10 (2014), e110274, https://doi.org/10.1371/journal.pone.0110274.

[2] Sendelbach S., Funk M., "Alarm fatigue: a patient safety concern", *AACN Advanced Critical Care*, **24**:4 (2013), 378–386, https://doi.org/10.4037/NCI.0b013e3182a903f9.

[3] Funk M., Clark J.T., Bauld T.J., Ott J.C., Coss P., "Attitudes and practices related to clinical alarms", *American Journal of Critical Care*, **23**:3 (2014), e9–e18, https://doi.org/10.4037/ajcc2014315.

[4] Graham K.C., Cvach M., "Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms", *American Journal of Critical Care*, **19**:1 (2010), 28–34, https://doi.org/10.4037/ajcc2010653.

[5] Borowski M., Gorges M., Fried R., Such O., Wrede C., Imhoff M., "Medical device alarms", *Biomedical Instrumentation and Technology*, **45**:1 (2011), 73–81, https://doi.org/10.2345/0899-8205-45.1.73.

[6] Clifford G.D., Silva I., Moody B., Li Q., Kella D., Shahin A., Kooistra T., Perry D., Mark R.G., "The physionet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the ICU", *Computing in Cardiology Conference (CinC)*, 2015, 273–276, https://doi.org/10.1109/CIC.2015.7408639.

[7] Zong W., Moody G.B., Mark R.G., "Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and pulsatile signals", *Medical and Biological Engineering and Computing*, **42**:5 (2004), 698–706, https://doi.org/10.1007/BF02347553.

[8] Wagner G.S., *Marriott's practical electrocardiography*, 11th edition, Lippincott Williams and Wilkins, 2008.

[9] Elgendi M., "Standard terminologies for photoplethysmogram signals", *Current Cardiology Reviews*, **8**:3 (2012), 215–219, https://doi.org/10.2174/157340312803217184.

[10] Hadian M., Pinsky M.R., "Evidence-based review of the use of the pulmonary artery catheter: impact data and complications", *Critical Care*, **10**:3 (2006), S8, https://doi.org/10.1186/cc4834.

[11] Hong S., Zhou Y., Shang J., Xiao C., Sun J., "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review", *Computers in Biology and Medicine*, **122** (2020), 103801, https://doi.org/10.1016/j.compbiomed.2020.103801.

[12] Faust O., Hagiwara Y., Hong T.J., Lih O.S., Acharya U.R., "Deep learning for healthcare applications based on physiological signals: A review", *Computer Methods and Programs in Biomedicine*, **161** (2018), 1–13, https://doi.org/10.1016/j.cmpb.2018.04.005.

[13] Acharya U.R., Fujita H., Lih O.S., Hagiwara Y., Tan J.H., Adam M., "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural networks", *Information Sciences*, **405** (2017), 81–90, https://doi.org/10.1016/j.ins.2017.04.012.

[14] Kiranyaz S., Ince T., Gabbouj M., "Real-time patient-specific ECG classification by 1-D convolutional neural networks", *IEEE Transactions on Biomedical Engineering*, **63**:3 (2016), 664–675, https://doi.org/10.1109/TBME.2015.2468589.

[15] Rajpurkar P., Hannun A.Y., Haghpanahi M., Bourn C., Ng A.Y., *Cardiologist-level arrhythmia detection with convolutional neural networks*, arXiv preprint arXiv:1707.01836, 2017.

[16] Rudin C., "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, **1**:5 (2019), 206–215, https://doi.org/10.1038/s42256-019-0048-x.

[17] Tonekaboni S., Joshi S., McCradden M.D., Goldenberg A., "What clinicians want: contextualizing explainable machine learning for clinical end use", *Machine Learning for Healthcare Conference*, 2019, 359–380.

[18] Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R. et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, **58** (2020), 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[19] Adadi A., Berrada M., "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)", *IEEE Access*, **6** (2018), 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.

[20] Lundberg S.M., Lee S.I., "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, 2017, 4765–4774.

[21] Ribeiro M.T., Singh S., Guestrin C., ""Why should I trust you?" Explaining the predictions of any classifier", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery, Data Mining*, 2016, 1135–1144, https://doi.org/10.1145/2939672.2939778.

[22] Sundararajan, M. and Taly, A. and Yan, Q., "Axiomatic attribution for deep networks", *International Conference on Machine Learning*, 2017, 3319–3328.

[23] Bahdanau D., Cho K., Bengio Y., *Neural machine translation by jointly learning to align, translate*, arXiv preprint arXiv:1409.0473, 2014 (in Russian).

[24] Wachter S., Mittelstadt B., Russell C., "Counterfactual explanations without opening the black box: automated decisions and the GDPR", *Harvard Journal of Law and Technology*, **31** (2017), 841–887.

[25] Kraskov A., Stögbauer H., Grassberger P., "Estimating mutual information", *Physical Review E*, **69**:6 (2004), 066138, https://doi.org/10.1103/PhysRevE.69.066138.

[26] Shapley L.S., "A value for n-person games", *Contributions to the Theory of Games*, Princeton University Press, 1953, 307–317.

[27] Alvarez-Melis D., Jaakkola T.S., "Towards robust interpretability with self-explaining neural networks", *Advances in Neural Information Processing Systems*. V. 31, 2018, 7775–7784.

[28] Miller T., "Explanation in artificial intelligence: Insights from the social sciences", *Artificial Intelligence*, **267** (2019), 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[29] Ismail Fawaz H., Forestier G., Weber J., Idoumghar L., Muller P.A., "Deep learning for time series classification: a review", *Data Mining and Knowledge Discovery*, **33**:4 (2019), 917–963, https://doi.org/10.1007/s10618-019-00619-1.

[30] Craik A., He Y., Contreras-Vidal J.L., "Deep learning for electroencephalogram (EEG) classification tasks: a review", *Journal of Neural Engineering*, **16**:3 (2019), 031001, https://doi.org/10.1088/1741-2552/ab0ab5.

[31] Cover T.M., Thomas J.A., *Elements of information theory*, 2nd edition, John Wiley and Sons, 2012.

[32] Krishna S., Han T., Gu A., Pombra J., Jabbari S., Wu S., Lakkaraju H., *The disagreement problem in explainable machine learning: A practitioner's perspective*, arXiv preprint arXiv:2202.01602, 2022.

[33] Breiman L., "Random forests", *Machine Learning*, **45**:1 (2001), 5–32, https://doi.org/10.1023/A:1010933404324.

[34] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., "Attention is all you need", *Advances in Neural Information Processing Systems*. V. 30, 2017, 5998–6008.

[35] Kumar R., Singh A., Patel N., Gupta S., "Counterfactual explanations for multimodal machine learning systems", *Proceedings of the AAAI Conference on Artificial Intelligence*. V. 38, 2024, 13245–13253, https://doi.org/10.1609/aaai.v38i12.29204.

[36] Belghazi M.I., Baratin A., Rajeshwar S., Ozair S., Bengio Y., Courville A., Hjelm D., "Mutual information neural estimation", *International Conference on Machine Learning*, 2018, 531–540.

[37] Poole B., Ozair S., Van Den Oord A., Alemi A., Tucker G., "On variational bounds of mutual information", *International Conference on Machine Learning*, 2019, 5171–5180.

[38] He K., Zhang X., Ren S., Sun J., "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770–778, https://doi.org/10.1109/CVPR.2016.90.

[39] Zhang L., Wang H., Chen M., Liu X., "Cross-modal attention networks for medical image analysis", *IEEE Transactions on Medical Imaging*, **42**:8 (2023), 2341–2354, https://doi.org/10.1109/TMI.2023.3267189.

[40] Liu M., Zhang W., Chen R., Yang J., "Vision transformer architectures for physiological time series analysis", *Proceedings of the 38th Conference on Neural Information Processing Systems*, 2024, 12456–12467.

[41] Natarajan A., Chang Y., Marlin B., Ghassemi M., "A wide and deep transformer neural network for 12-lead ECG classification", *Computing in Cardiology*, 2020, 1–4, https://doi.org/10.22489/CinC.2020.107.

[42] Sun Jing, Yang Can, "A review of big data applications of physiological signal data", *PMC Biophysics*, **12**:1 (2019), 3, https://doi.org/10.1186/s13628-019-0050-1.

[43] Peluffo-Ordóñez Diego H., Cobos C., Macas M., "Physiological signals fusion oriented to diagnosis - A review", *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*. V. 1, 2016, 475–482, https://doi.org/10.5220/0006000104750482.

[44] Bauer M., Czy K.M., Hipp J., Ulmke M., *Bayesian Fusion: Modeling and Application*, Fraunhofer FKIE, 2019.

[45] Zainudin Z.A., Abd A.K.A., Azman A., "A Gentle Approach to Multi-Sensor Fusion Data Using Linear Kalman Filter", *Journal of Advanced Research in Applied Sciences and Engineering Technology*, **41**:1 (2024), 269–281, https://doi.org/10.37934/araset.41.1.269281.

[46] Li X.-R., Jilkov V.P., *Tutorial on Multisensor Management and Fusion Algorithms for Target Tracking*, University of New Orleans, 2001.

[47] Azimi I., Pahlevan N., Atyabi A., Zoroufian P., Tavallali P., Blaber R.P., Doyle T.E., Reisman D., "Multimodal deep learning for biomedical data fusion: a review", *PMC Biophysics*, **15**:1 (2022), 7, https://doi.org/10.1186/s13628-022-00089-9.

[48] Ramachandram D., "A Survey on Deep Learning for Multimodal Data Fusion", *Neural Computation*, **32**:5 (2020), 829–884.

[49] Ahmed S., Ahmed S.S., *A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications*, arXiv preprint arXiv:2408.02686, 2024.

[50] Mehrabian H., Shalbaf A., "Fusion-driven multimodal learning for biomedical time series in surgical care", *Scientific Reports*, **14**:1 (2024), 19685, https://doi.org/10.1038/s41598-024-70560-z.

[51] Lieskovska E., Jakubec M., Jarina R., Chmulik M., "A review on speech emotion recognition using deep learning and attention mechanism", *Electronics*, **10**:10 (2021), 1163.

[52] Elgendi M., Eskofier B., Dokos S., Abbott D., "Revisiting QRS detection methodologies for portable, wearable, battery-operated, and wireless ECG systems", *PloS one*, **9**:1 (2014), e84018, https://doi.org/10.1371/journal.pone.0084018.

Author Info

1. **Trofimov Yuri Vladislavovich**
Assistant, Junior Researcher at Dubna State University;
Software Engineer at Meshcheryakov Laboratory of Information Technologies, Joint Institute for Nuclear Research.

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University. E-mail: ura_trofim@bk.ru

2. **Averkin Alexey Nikolaevich**
Associate Professor at Dubna State University;
Leading Researcher at the Institute of Computer Science and Management of the Russian Academy of Sciences.

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University. E-mail: averkin2003@inbox.ru

3. **Kuznetsov Egor Mikhailovich**
Student of Dubna State University.

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University. E-mail: kem.22@uni-dubna.ru

4. **Eremeev Alexander Pavlovich**
Professor of the Department of Applied Mathematics at the National Research University "Moscow Power Engineering Institute (MPEI)".

Russia, 111250, Moscow, Krasnokazarmennaya str., 14, building 1, NPU MPEI.

5. **Nechaevskiy Andrey Vasilyevich**
Associate Professor at Dubna State University;
Senior Researcher at the Information Technology Laboratory of the Joint Institute for Nuclear Research (JINR).

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University. E-mail: nechav@jinr.ru