УДК 004.8, 004.89, 616.12-008.313

МОДАЛЬНАЯ, ВРЕМЕННАЯ И СОБЫТИЙНАЯ ХРУПКОСТЬ МУЛЬТИМОДАЛЬНОЙ МОДЕЛИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПРИ АНАЛИЗЕ ТРЕВОГ ЖЕЛУДОЧКОВОЙ ТАХИКАРДИИ

Трофимов Юрий Владиславович[1], Аверкин Алексей Николаевич[2], Кузнецов Егор Михайлович[3]

[1]*Ассистент, младший научный сотрудник;*
Государственный университет «Дубна»;
Россия, 141980, Московская область, г. Дубна, ул. Университетская, 19;
Инженер-программист;
Объединённый институт ядерных исследований;
Россия, 141980, Московская область, г. Дубна, ул. Жолио-Кюри, 6;
e-mail: ura_trofim@bk.ru.

[2]*Ведущий научный сотрудник;*
ФИЦ «Информатика и управление» РАН;
Россия, 119333, г. Москва, ул. Вавилова, 44/2;
Доцент;
Государственный университет «Дубна»;
Россия, 141980, Московская область, г. Дубна, ул. Университетская, 19;
e-mail: averkin2003@inbox.ru.

[3]*Студент;*
Государственный университет «Дубна»;
Россия, 141980, Московская область, г. Дубна, ул. Университетская, 19;
e-mail: Kot454556@yandex.ru.

Рассматривается задача проверки хрупкости мультимодальной модели анализа физиологических временных рядов при классификации истинных и ложных тревог желудочковой тахикардии. Проведена многоосевая оценка модели в полном режиме, при удалении ЭКГ, на трудных и конфликтных событиях, при временных сдвигах окна и в режимах калибровочного анализа. Показано, что высокая точность на полном тесте скрывает модальную и временную хрупкость, а отдельные стратегии дообучения дают только ограниченные эффекты. Внешняя проверка на PhysioNet/CinC Challenge 2015 подтверждает применимость протокола к независимому источнику ОРИТ-тревог.

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Ключевые слова: медицинский искусственный интеллект, физиологические временные ряды, желудочковая тахикардия, мультимодальная модель, хрупкость модели, стресс-тестирование, калибровка, ложные тревоги, PhysioNet.

Для цитирования:

Трофимов Ю. В., Аверкин А. Н., Кузнецов Е. М. Модальная, временная и событийная хрупкость мультимодальной модели искусственного интеллекта при анализе тревог желудочковой тахикардии // Системный анализ в науке и образовании. 2026. № 3. С. 79–94. EDN: GDXOFS. URL: https://sanse.ru/index.php/sanse/article/view/737.

Статья находится в открытом доступе и распространяется в соответствии с лицензией Creative Commons «Attribution» («Атрибуция») 4.0 Всемирная (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/deed.ru

MODAL, TEMPORAL AND EVENT-LEVEL FRAGILITY OF A MULTIMODAL ARTIFICIAL INTELLIGENCE MODEL FOR VENTRICULAR TACHYCARDIA ALARM ANALYSIS

Trofimov Yuri V.[1], Averkin Alexey N.[2], Kuznetsov Egor M.[3]

[1]Assistant, Junior Researcher;
Dubna State University;
19 Universitetskaya St., Dubna, Moscow Region, 141980, Russia;
Software Engineer;
Joint Institute for Nuclear Research;
6 Joliot-Curie St., Dubna, Moscow Region, 141980, Russia;
e-mail: ura_trofim@bk.ru.

[2]Leading Researcher;
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences;
44/2 Vavilova St., Moscow, 119333, Russia;
Associate Professor;
Dubna State University;
19 Universitetskaya St., Dubna, Moscow Region, 141980, Russia;
e-mail: averkin2003@inbox.ru.

[3]Student;
Dubna State University;
19 Universitetskaya St., Dubna, Moscow Region, 141980, Russia;
e-mail: Kot454556@yandex.ru.

The paper considers fragility assessment of a multimodal physiological time-series model for classifying true and false ventricular tachycardia alarms. The model is evaluated under several stress profiles: full input, ECG removal, hard and conflict events, temporal window shifts, and calibration checks. The results show that high performance on the full test set may hide modality and temporal fragility, while fine-tuning strategies provide only bounded improvements. An external check on the PhysioNet/CinC Challenge 2015 dataset demonstrates that the evaluation procedure can be applied to an independent ICU alarm source.

Keywords: medical artificial intelligence, physiological time series, ventricular tachycardia, multimodal model, model fragility, stress testing, calibration, false alarms, PhysioNet.

For citation:
Trofimov Yu. V., Averkin A. N., Kuznetsov E. M. Modal, temporal and event-level fragility of a multimodal artificial intelligence model for ventricular tachycardia alarm analysis. *System analysis in science and education*, 2026;(3):79–94 (in Russ). EDN: GDXOFS. Available from: https://sanse.ru/index.php/sanse/article/view/737.

Введение

Мультимодальная модель, обученная различать истинные и ложные тревоги желудочковой тахикардии в отделениях реанимации и интенсивной терапии (ОРИТ), может показывать высокое качество на полном тестовом наборе *VTaC* и одновременно демонстрировать хрупкость на сложных и конфликтных подмножествах данных или при контролируемых возмущениях входного контекста [1, 2]. В проведённых экспериментах на сложном подмножестве событий доля ошибок достигала 0,6197 при средней уверенности модели 0,8230. При смещении предтревожного окна на 5 секунд площадь под *ROC*-кривой снижалась с 0,9592 до 0,7788, а доля ложноотрицательных решений возрастала с 0,0511 до 0,5766. В отдельных стресс-профилях модель сохраняет высокую уверенность при пропуске истинных тревог.

Такая ситуация особенно критична для тревог желудочковой тахикардии. Ложное срабатывание создаёт избыточную нагрузку на персонал ОРИТ и может запускать ненужную клиническую реакцию, тогда как пропуск истинного эпизода связан с риском жизнеугрожающего состояния [3–7]. Поэтому задача анализа тревог ЖТ не сводится к повышению средней точности классификатора. Необходимо

понимать, при каких нарушениях входного сигнала и в каких временных условиях вывод модели ещё можно считать применимым.

Под предтревожным окном в работе понимается временной интервал физиологических сигналов, подаваемый модели на вход и предшествующий моменту срабатывания тревоги. В клиническом мониторинге такой интервал обычно включает несколько синхронизированных каналов: электрокардиограмму (ЭКГ), фотоплетизмограмму (ФПГ) и инвазивное артериальное давление (ИАД). ЭКГ может содержать артефакты движения, ФПГ зависит от качества периферической пульсовой волны, а ИАД отражает гемодинамическую реакцию. Мультимодальная модель должна использовать эти каналы совместно, однако само наличие нескольких модальностей не гарантирует устойчивости. Модель может фактически опираться на доминирующий ЭКГ-сигнал, терять качество при его удалении или резко деградировать при смещении анализируемого интервала относительно момента тревоги.

Ранее для задачи анализа ОРИТ-сигналов была рассмотрена мультимодальная объяснимость модели *ResNetFusionClassifier* и предложена метрика согласованности объяснений *Coherence* для трёхканальных физиологических сигналов *VTaC* [8]. В той работе основной акцент был сделан на согласованности объяснений между модальностями и метрическом анализе мультимодальных атрибуций. Настоящая работа решает другую задачу: не вводит новую метрику объяснений, а проверяет границы применимости уже обученной мультимодальной модели в стрессовых режимах.

В качестве базовой модели используется *ResNetFusion* – архитектура, в которой отдельные одномерные остаточные сверточные ветви обрабатывают физиологические сигналы, после чего их признаки объединяются в общем классификационном контуре. Такая архитектура подходит для анализа ЭКГ, ФПГ и ИАД, но требует отдельной проверки на хрупкость. Высокое качество в полном режиме не отвечает на вопросы о поведении модели без ЭКГ, при смещении временного окна, на трудных событиях или при конфликте между электрокардиографическими и гемодинамическими признаками.

В работе рассматриваются две группы проверок. Первая группа – диагностические стресс-профили, которые не изменяют модель, а показывают, в каких режимах она теряет устойчивость: при удалении модальностей, смещении предтревожного окна, переходе к сложным событиям, нарушении калибровки и проверке на независимом наборе *CinC2015*. Вторая группа — интервенции, то есть попытки изменить обучение или архитектурный контур модели. К ним относятся *ECG-dropout,* перевзвешивание трудных событий, *XAI*/*conflict-reweighting,* латентно-предиктивный JEPA-подобный контур и механизм обучаемого внимания. Для интервенций недостаточно показать локальный прирост метрики: если метод не превосходит отрицательный контроль, его нельзя считать подтверждённым улучшением модели. Физиологический профиль *PhysProfile* рассматривается отдельно: он не изменяет модель и не является новым классификатором, а используется как способ предметного описания конфликтных и уверенно ошибочных событий — конфликта ЭКГ и гемодинамики, слабой поддержки ФПГ/ИАД, низкого качества сигнала и временной несогласованности каналов.

Цель работы состоит в проверке устойчивости мультимодальной модели анализа тревог желудочковой тахикардии при нарушении входного контекста: потере модальностей, смещении временного окна и переходе к сложным событиям.

В работе выделены конкретные режимы отказа модели: ограниченное допустимое смещение предтревожного окна, рост ложноотрицательных и уверенно ложноотрицательных решений, зависимость от ЭКГ и непереносимость части интервенций через отрицательные контроли. Высокая метрика рассматривается не как доказательство надёжности, а как исходная точка для поиска измеримых условий отказа; *PhysProfile* используется только как физиологически понятный язык описания конфликтных событий, а не как новый предиктор ошибок.

1. Данные и базовая модель

Основной экспериментальный набор данных в работе – *VTaC* [1, 2]. Используется фиксированное разбиение: 2048 обучающих, 495 валидационных и 482 тестовых события; в тестовой части содержится 137 истинных и 345 ложных тревог ЖТ. Для каждого события доступны две ЭКГ-деривации, фотоплетизмограмма и инвазивное артериальное давление. Предтревожное окно имеет длительность

10 секунд: 2500 отсчётов при частоте дискретизации 250 Гц. Для внешней проверки используется подвыборка *PhysioNet*/*CinC Challenge* 2015, включающая только тревоги желудочковой тахикардии: 100 записей, из них 26 истинных и 74 ложные тревоги [3]. Каналы во внешней подвыборке доступны не для всех записей: инвазивное артериальное давление представлено в 49 случаях, фотоплетизмограмма – в 87 случаях. Поэтому *CinC2015* используется не как полноценное воспроизведение всех мультимодальных профилей *VTaC*, а как ограниченная внешняя проверка ключевых признаков хрупкости модели: зависимости от ЭКГ, деградации при смещении окна и нарушения калибровки уверенности. Сводная характеристика наборов данных приведена в таблице 1.

Табл. 1. Характеристика экспериментальных наборов данных

Набор данных	**Задача**	**Число событий**	**Истинные тревоги**	**Ложные тревоги**	**Каналы**	**Роль в работе**
VTaC	Классификация истинных и ложных тревог ЖТ	482	137	345	ЭКГ, ФПГ, ИАД	основной тестовый набор
CinC2015 VT-only	Внешняя проверка профилей хрупкости на ОРИТ-тревогах	100	26	74	ЭКГ, ФПГ частично, ИАД частично	ограниченная внешняя проверка

На исходном тестовом режиме *VTaC*, без удаления каналов и без смещения окна, *ResNetFusion* показывает высокое качество различения истинных и ложных тревог ЖТ. Для пороговых показателей использовался один фиксированный порог, выбранный на чистой валидационной выборке до проведения стресс-тестов; положительным классом считалась истинная тревога ЖТ. Базовые значения метрик приведены в таблице 2 и далее используются как точка отсчёта для анализа модальной, временной и событийной деградации.

Табл. 2. Базовое качество ResNetFusion на исходном тестовом режиме VTaC

Метрика	**Значение**
Площадь под *ROC*-кривой	0,9592
Площадь под *PR*-кривой	0,8938
Сбалансированная точность	0,9150
Чувствительность	0,9489
Специфичность	0,8812
Доля пропущенных истинных тревог	0,0511
Доля ложных срабатываний	0,1188
Ошибка калибровки уверенности	0,0382
Оценка Брайера	0,0746
Средняя уверенность	0,8951

Дополнительно было выполнено контрольное сравнение нескольких архитектур, работающих с исходными сигналами, совместимых с тем же протоколом временного сдвига окна. Эта проверка не рассматривалась как полноценный бенчмарк архитектур или поиск лучшей модели: её задача состояла в том, чтобы проверить, является ли выявленная хрупкость особенностью только *ResNetFusion*. Для сопоставимости все модели оценивались на одном тестовом разбиении *VTaC* из 482 событий, а порог классификации для каждой модели выбирался только на чистой валидационной части.

Таблица 3 показывает, что хрупкость не сводится к одной архитектуре: у всех моделей удаление ЭКГ снижает ROC-AUC, а допустимое смещение окна не превышает 5 с. По смешанному критерию отказа часть моделей выходит из устойчивой области уже при 2,5 с, что указывает на общую зависимость от временного положения окна и ЭКГ-сигнала.

Табл. 3. Контрольное сравнение временной и модальной хрупкости архитектур на VTaC

Модель	***ROC-AUC* исходный режим**	***ROC-AUC* без ЭКГ**	***ΔROC-AUC* без ЭКГ**	**Допустимое смещение по *ROC-AUC*, с**	**Первый критический сдвиг, с**	**Причина первого отказа**	**Доля пропущенных истинных тревог при 5 с**	***ECE* при 5 с**
ResNetFusion	0,959	0,765	−0,194	2,5	5,0	несколько критериев	0,5766	0,122
CNN-RNN с механизмом внимания	0,805	0,463	−0,342	5,0	7,5	несколько критериев	0,1460	0,060
CNN-1D	0,900	0,589	−0,310	2,5	2,5	высокая доля пропусков	0,4818	0,047
GRU	0,897	0,662	−0,236	2,5	5,0	несколько критериев	0,4964	0,068
TCN	0,926	0,633	−0,292	2,5	2,5	высокая доля пропусков	0,5328	0,046

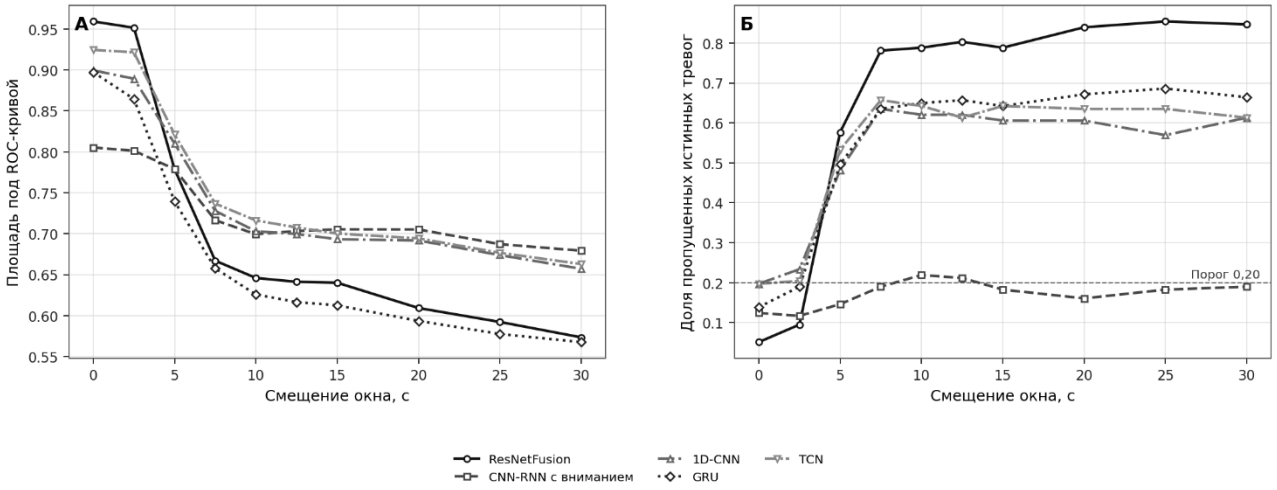


Рис. 1. Временная хрупкость разных архитектур при смещении предтревожного окна: А – изменение площади под ROC-кривой; Б – изменение доли пропущенных истинных тревог

Рисунок 1 дополняет таблицу 3: форма деградации различается, но общий профиль сохраняется – при смещении окна растёт доля пропущенных истинных тревог, а качество выходит за устойчивую область. Далее подробно анализируется *ResNetFusion* как основная модель эксперимента; временные профили остальных архитектур построены по той же сетке смещений и приведены в сводном виде.

Отдельно была выполнена контрольная проверка классических моделей машинного обучения на ручных признаках временных рядов. Эти модели не включались в таблицу 3 и на рисунок 1, поскольку используют другое представление данных и не сопоставимы напрямую с архитектурами, работающими с исходными сигналами. Признаки извлекались из ЭКГ, ФПГ и ИАД и включали статистические, спектральные, пиковые и простые межканальные характеристики. Лучшая классическая модель по площади под *ROC*-кривой на валидационной выборке – *LightGBM* – на тестовой части *VTaC* достигла *ROC-AUC* = 0,9246, *PR-AUC* = 0,8427 и сбалансированной точности 0,8328. Для *ResNetFusion* в том же разбиении значения составили ROC-AUC = 0,9592, PR-AUC = 0,8938, сбалансированная точность – 0,9150. Более высокие показатели качества по сравнению с *LightGBM* обосновывают выбор *ResNetFusion* в качестве базовой конфигурации для последующего стресс-тестирования.

2. Многоосевая схема проверки хрупкости

В работе хрупкость модели рассматривается как ухудшение качества, рост числа ошибок или потеря надёжности вывода при изменении входного контекста. Проверка строится по двум ветвям: диагностической и интервенционной. Диагностическая ветвь не изменяет модель и показывает, в каких режимах она теряет устойчивость. Интервенционная ветвь проверяет, дают ли модификации обучения или архитектуры реальный эффект по сравнению с отрицательными контролями. Диагностическая ветвь включает четыре основные оси, а также физиологический профиль событий *PhysProfile*.

- Первая ось – модальная хрупкость. Сравниваются три режима: полный набор каналов; конфигурация без ЭКГ, где остаются только ФПГ и ИАД; конфигурация только с ЭКГ, эквивалентная удалению гемодинамических каналов. Такая проверка показывает, насколько модель зависит от ЭКГ и сохраняет ли качество при исключении гемодинамических сигналов.

- Вторая ось – временная хрупкость. Предтревожное окно последовательно смещается назад относительно момента срабатывания тревоги. По изменению метрик определяется скорость деградации модели и граница устойчивой работы при уходе от исходного временного контекста.

- Третья ось – событийная хрупкость. Отдельно анализируются сложные события, конфликтные события и уверенно ошибочные решения. Такой разбор нужен для перехода от средней оценки по всему тестовому набору к наиболее проблемным эпизодам.

- Четвёртая ось – калибровочная хрупкость. Анализируются средняя уверенность, ошибка калибровки уверенности, оценка Брайера и доля уверенно ошибочных решений. Эта ось фиксирует режимы, в которых число ошибок возрастает, но модель сохраняет высокую уверенность.

- Физиологический профиль событий *PhysProfile* завершает диагностическую ветвь. Он не используется как основной предиктор ошибок и не подменяет клиническую интерпретацию. Его задача – предметно описывать проблемные случаи: несогласованность между ЭКГ и гемодинамическим откликом, слабую поддержку со стороны ФПГ или ИАД, низкое качество сигнала и временную несогласованность каналов.

Интервенционная ветвь включает *ECG-dropout* [9], перевзвешивание трудных событий, *XAI*-перевзвешивание [11–13], *JEPA*-подобный контур и модификации с механизмом внимания [14–18]. Эти модификации не считаются улучшением только по факту локального прироста метрики. Каждая из них сопоставляется с отрицательным контролем: случайным *dropout*, перемешанным вниманием, упрощённым латентным контуром или случайным перевзвешиванием. Если модификация не превосходит соответствующий контроль, её эффект не считается подтверждённым.

Общая логика проверки представлена на рисунке 2. Диагностическая ветвь выявляет режимы отказа модели, а интервенционная ветвь отделяет подтверждённые улучшения от эффектов, воспроизводимых техническими или случайными контролями.

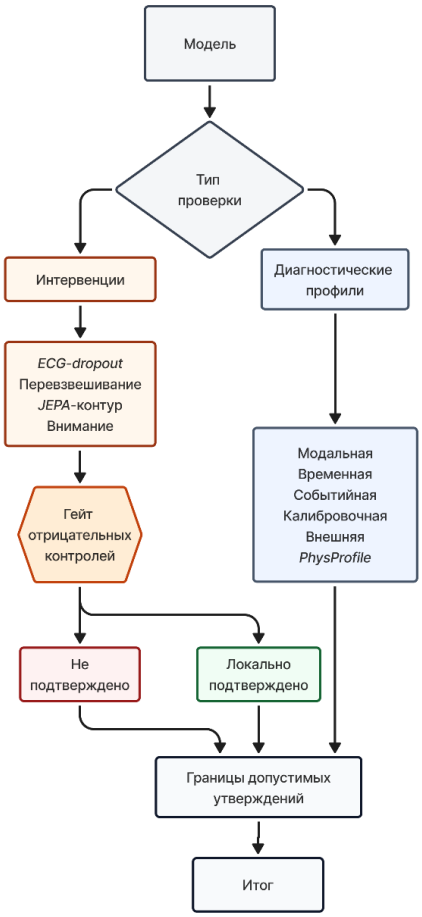


Рис. 2. Схема многоосевой проверки хрупкости: диагностические профили и интервенции

3. Модальная хрупкость и зависимость от ЭКГ

Одной из ключевых проверок является оценка зависимости модели от ЭКГ. В полной конфигурации *ResNetFusion* использует несколько синхронизированных физиологических каналов, однако высокая метрика в таком режиме не показывает, насколько равномерно модель опирается на разные модальности. Для этого дополнительно рассматривается режим удаления ЭКГ, в котором на вход модели подаются только ФПГ и ИАД. Если качество резко снижается, это указывает на модальную хрупкость и возможную доминирующую роль ЭКГ в принятии решения.

Для снижения такой зависимости была проверена модификация *ECG-dropout*, связанная с идеей модального *dropout* [9]. При обучении она реализует модальное отключение сигналов: часть входных каналов временно исключается без добавления шума и без временного маскирования. В варианте *ECG-dropout* отключение направлено преимущественно на ЭКГ: вероятность применения процедуры составляет 0,65, а целевая вероятность отключения ЭКГ – 0,45. ФПГ и ИАД в этой конфигурации не отключаются. В контрольной конфигурации используется тот же тип регуляризации, но отключение распределено равномерно между каналами: для ЭКГ, ФПГ и ИАД задано по 0,15. Поэтому сравнение с контролем позволяет отделить эффект направленного отключения ЭКГ от общего эффекта модального возмущения.

В режиме без ЭКГ площадь под *ROC*-кривой у модели с *ECG-dropout* составляет 0,8092 против 0,7893 у базовой модели и 0,8008 у контрольной конфигурации. Разрыв между полным режимом и режимом без ЭКГ уменьшается с 0,1680 до 0,1494, что указывает на снижение зависимости от ЭКГ. Локальный прирост ROC-AUC в режиме без ЭКГ не переносится на временные сдвиги и сложные подмножества данных. Для ОРИТ такой режим сохраняет практический смысл: ЭКГ может временно теряться или зашумляться. Сводные результаты приведены в таблице 4.

Табл. 4. Влияние ECG-dropout на модальную хрупкость модели

Сравнение	Профиль	Метрика	Δ	95% CI	Интерпретация
ECG-dropout / базовая модель	без ЭКГ	*ROC-AUC*	+0,0185	[0,0039; 0,0327]	устойчивый локальный эффект
ECG-dropout / случайный контроль	без ЭКГ	*ROC-AUC*	+0,0100	[0,0005; 0,0195]	небольшой эффект выше случайного контроля
ECG-dropout / базовая модель	чистый режим	Сбалансированная точность	−0,0200	[−0,0355; −0,0040]	компромиссная деградация в исходном режиме
ECG-dropout / базовая модель	разрыв качества между полным режимом и режимом без ЭКГ	Разрыв качества	+0,0209	[0,0063; 0,0354]	разрыв качества при потере ЭКГ сократился

4. Временная хрупкость и граница применимости модели

Временная хрупкость оценивалась на тестовой части *VTaC*, описанной в разделе 1. Для каждого смещения использовался тот же фиксированный порог классификации, выбранный на чистой валидационной выборке до проведения стресс-тестов. Смещение окна означало выбор более раннего предтревожного интервала из расширенного контекста записи; циклический сдвиг сигнала не применялся. Исходная точка при смещении 0 с соответствует базовой оценке модели из таблицы 2, поэтому далее анализируется изменение метрик относительно исходного режима.

Уверенным пропуском истинной тревоги считается ложноотрицательное решение, при котором модель относит истинную тревогу ЖТ к ложным тревогам с уверенностью не ниже 0,8. Этот показатель выделяет подмножество ложноотрицательных решений, для которых апостериорная вероятность выбранного ложного класса составляет не менее 0,8.

Количественные результаты временного стресс-теста приведены в таблице 5 и на рисунке 3.

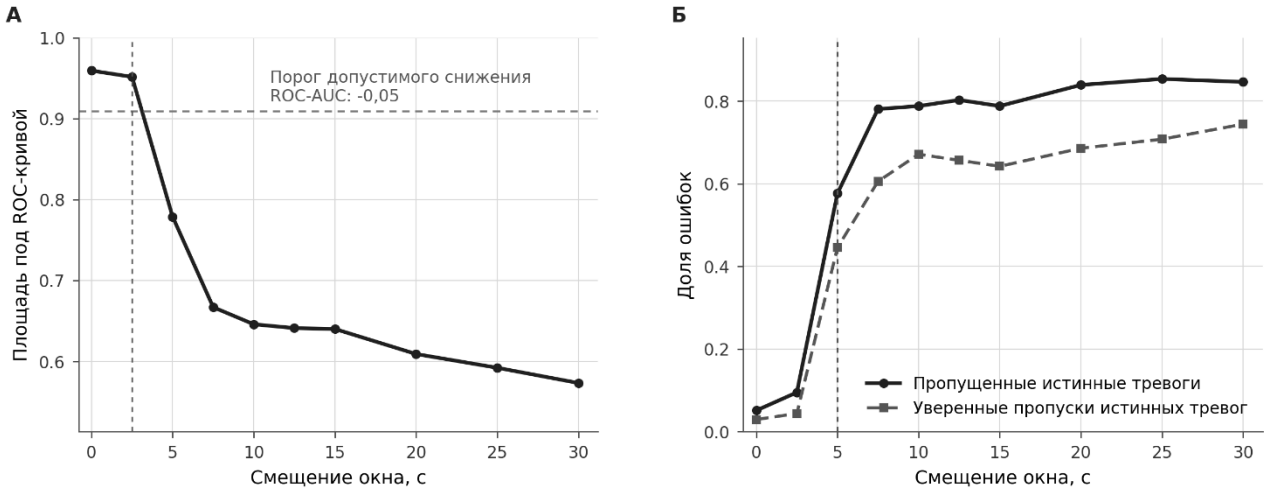


Рис. 3. Временная деградация качества и уверенности модели при смещении предтревожного окна

Табл. 5. Временной профиль хрупкости модели

Смещение окна, с	**Площадь под *ROC*-кривой**	**Площадь под *PR*-кривой**	**Сбалансированная точность**	**Доля пропущенных истинных тревог**	**Уверенные пропуски истинных тревог**	**Средняя уверенность**	**Ошибка калибровки уверенности**
0,0	0,9592	0,8938	0,9150	0,0511	0,0292	0,8951	0,0382
2,5	0,9515	0,8823	0,8888	0,0949	0,0438	0,8877	0,0382
5,0	0,7788	0,6217	0,6696	0,5766	0,4453	0,8909	0,1223
7,5	0,6671	0,4914	0,5704	0,7810	0,6058	0,8958	0,1778
10,0	0,6460	0,4691	0,5696	0,7883	0,6715	0,9024	0,1844
12,5	0,6413	0,4383	0,5594	0,8029	0,6569	0,8997	0,1903
15,0	0,6402	0,4382	0,5638	0,7883	0,6423	0,8959	0,1802
20,0	0,6093	0,4095	0,5397	0,8394	0,6861	0,9009	0,1995
25,0	0,5923	0,3734	0,5353	0,8540	0,7080	0,8986	0,2004
30,0	0,5735	0,3596	0,5317	0,8467	0,7445	0,8971	0,2091

Анализ временного профиля выявляет участок резкой деградации между 2,5 и 5 с. Рабочая граница применимости задавалась условием снижения *ROC-AUC* не более чем на 0,05 относительно исходного режима; по этому правилу допустимое смещение предтревожного окна составляет 2,5 с. При смещении на 5 с модель выходит за эту границу: *ROC-AUC* падает до 0,7788, доля ложноотрицательных ответов возрастает до 0,5766, а ошибка калибровки увеличивается с 0,0382 до 0,1223. При дальнейшем смещении деградация усиливается: на 10 с *ROC-AUC* снижается до 0,6460, доля пропущенных истинных тревог достигает 0,7883, а средняя уверенность остаётся около 0,90. Такой профиль указывает на рассогласование дискриминативной способности и калибровки; без предварительной проверки временной устойчивости модель не должна использоваться в медицинских системах поддержки принятия врачебных решений.

Внешняя проверка выполнялась на подвыборке *PhysioNet/CinC Challenge* 2015 с тревогами желудочковой тахикардии, описанной в таблице 1. Из-за неполной доступности ФПГ и ИАД этот набор не использовался для полного воспроизведения мультимодальных профилей *VTaC*. Проверялись только два признака хрупкости: зависимость от ЭКГ и деградация при смещении предтревожного окна. В полном режиме *ROC-AUC* составил 0,9600, без ЭКГ – 0,8181, только с ЭКГ – 0,9345, при смещении окна на 20 с – 0,5759. *CinC2015* не подтверждает универсальность модели, но показывает переносимость схемы оценки на независимый источник ОРИТ-тревог.

5. Сложные события и уверенные ошибки модели

Для анализа ошибок с высокой уверенностью использовались два фиксированных подмножества: сложные и конфликтные события. Сложные события выделялись по признакам нестабильности модели. Конфликтные события выделялись по межканальному и объяснительному рассогласованию. Дальнейший анализ проверяет, как эти подмножества связаны с ошибками, уверенностью и калибровкой.

Сложное подмножество формировалось по априорно рассчитанному числовому показателю трудности. Показатель строился как агрегат индикаторов модельной нестабильности: ошибки в исходном режиме, уверенной ошибки, ложного принятия ложной тревоги за истинную, изменения решения при удалении ЭКГ, выраженного изменения вероятности между полным режимом и режимом без ЭКГ, а также высокой уверенности при модальной хрупкости. Событие включалось в сложное подмножество, если значение этого показателя было не ниже 2,75. По этому правилу в тестовой части *VTaC* было выделено 71 событие из 482.

Конфликтное подмножество формировалось отдельно по композитному показателю конфликтности объяснения. Этот показатель рассчитывался до анализа ошибок тестового набора и объединял четыре компонента: низкую согласованность атрибуций, расхождение между объяснением и решением модели, модальную хрупкость и неопределённость вывода. В конфликтное подмножество включались верхние 20% событий по значению этого показателя; для тестовой части *VTaC* это дало 97 событий из 482. Метки классов и ошибки тестового набора использовались только для последующей оценки характеристик подмножества, но не для его формирования.

Для анализа уверенных пропусков используется тот же порог уверенности 0,8, заданный в разделе временной хрупкости. Этот показатель выделяет подмножество ложноотрицательных решений, для которых апостериорная вероятность выбранного ложного класса составляет не менее 0,8.

Калибровочные характеристики исходного и стрессовых режимов приведены в таблице 6.

Табл. 6. Калибровка модели в исходном и стрессовых режимах

Профиль	**n**	**Доля ошибок**	**Средняя уверенность**	**Ошибка калибровки**	**Уверенно ошибочные решения**	**Доля пропущенных истинных тревог**
Исходный режим	482	0,0996	0,8951	0,0371	0,0290	0,0511
Без ЭКГ	482	0,2863	0,7365	0,0688	0,0187	0,3358
Сложное подмножество событий	71	0,6197	0,8230	0,4428	0,1972	0,4444
Конфликтное подмножество событий	97	0,1856	0,8038	0,1408	0,0206	0,0789
Смещение окна 10 с	482	0,2759	0,9024	0,1854	0,1929	0,7883
Смещение окна 20 с	482	0,2967	0,9009	0,2005	0,2012	0,8394

На сложном подмножестве доля ошибок достигает 0,6197 при средней уверенности 0,8230 и ошибке калибровки 0,4428, что соответствует режиму уверенных ошибок, а не только снижению средней точности. Во временных стресс-профилях основной риск связан с пропусками истинных тревог: при смещении на 10–20 с их доля достигает 0,7883–0,8394 при средней уверенности около 0,90. Такой режим не выявляется по одной метрике качества на исходном тесте. Для предметного разбора были выбраны три события из тестовой части *VTaC*: панели А и Б представляют ошибки с высокой уверенностью модели, а панель В – конфликтный контрольный случай без ошибки классификации. Такой выбор разделяет два явления: ошибочное решение модели и физиологическую проблемность события.

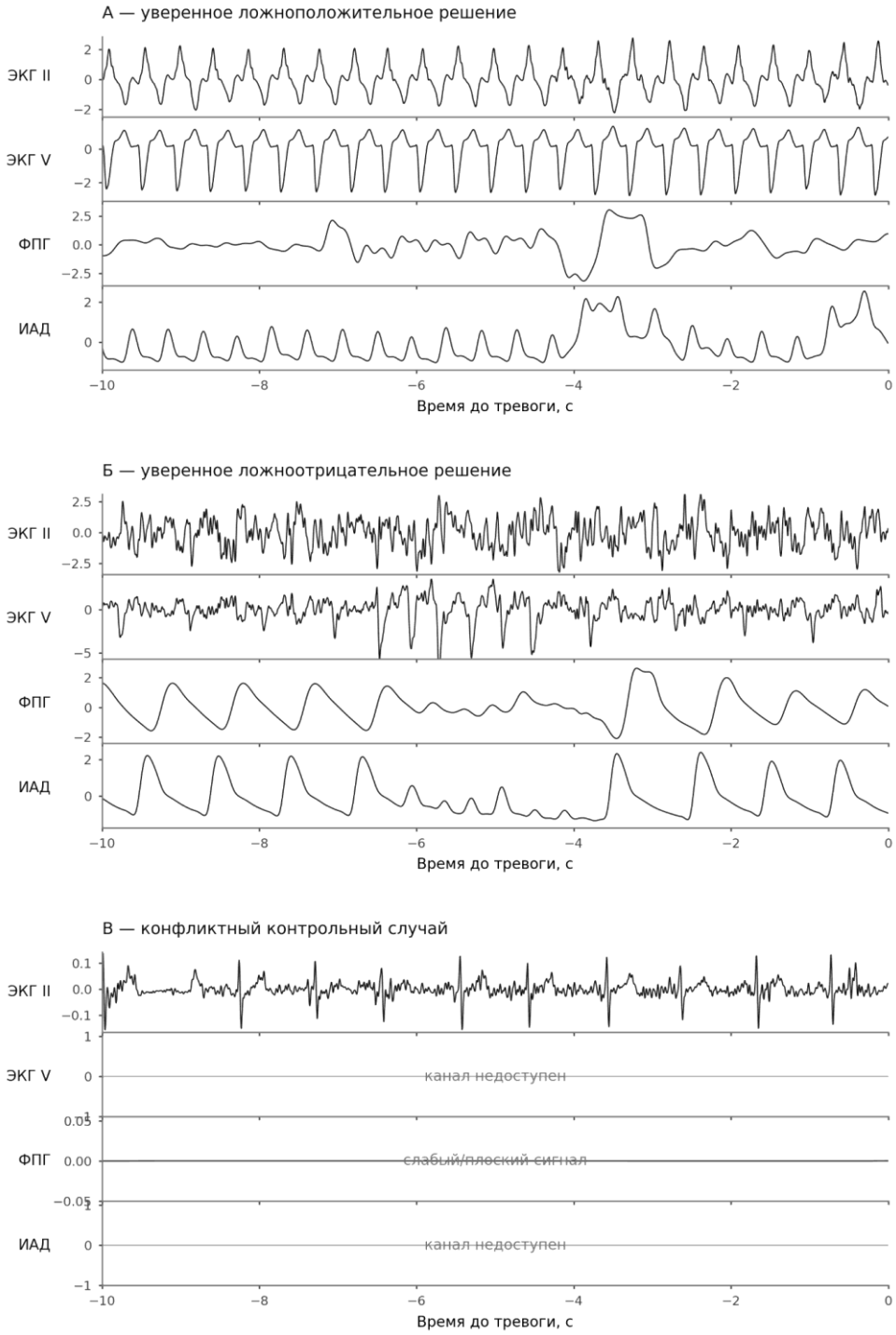


Рис. 4. Примеры событий VTaC: А – уверенное ложноположительное решение; Б – уверенное ложноотрицательное решение; В – физиологически конфликтный контрольный случай без ошибки классификации

В событии А модель относит ложную тревогу к истинным с вероятностью 0,9803. Это уверенное ложноположительное решение. ЭКГ имеет тахикардоподобный рисунок, но разметка события соответствует ложной тревоге. По *PhysProfile* сильный конфликт не выделяется: оценки частоты составляют 135 уд/мин по ЭКГ, 66 уд/мин по ФПГ и 132 уд/мин по ИАД. Отсутствие выраженного физиологического конфликта само по себе не гарантирует правильность вывода модели.

Событие Б относится к наиболее опасному типу ошибки: истинная тревога ЖТ классифицируется как ложная с уверенностью 1,0000. При этом оценки частоты резко расходятся: ЭКГ – 213 уд/мин, ФПГ

– 54 уд/мин, ИАД – 60 уд/мин. В этом случае модель подавляет истинную тревогу именно там, где между электрическим и гемодинамическим контурами есть выраженное рассогласование.

Событие B используется как контрольный пример. Модель правильно относит ложную тревогу к ложным, причём без высокой уверенности: вероятность истинной тревоги равна 0,4182, уверенность – 0,5818. *PhysProfile* указывает на проблемность записи: ЭКГ оценивается на уровне 105 уд/мин, ФПГ почти не поддерживает ритмический вывод, канал ИАД отсутствует. Физиологический конфликт или неполнота каналов не равны ошибке классификации, но помечают событие как требующее предметного разбора.

Рисунок 4 разделяет три ситуации: уверенное ложноположительное решение, уверенный пропуск истинной тревоги и конфликтный случай без ошибки классификации. Это задаёт границу применения *PhysProfile*: он не детектирует ошибки автоматически, а описывает физиологический контекст проблемных событий – слабую гемодинамическую поддержку, расхождение ЭКГ с ФПГ/ИАД, отсутствие или плоский характер каналов и локализованность признаков во времени.

6. Стратегии модификации и отрицательные контроли

После диагностики модальной, временной и событийной хрупкости проверялись способы снизить выявленные ограничения. Рассматривались *ECG-dropout*, перевзвешивание трудных событий, *XAI*/*conflict*-перевзвешивание, *JEPA*-подобный латентно-предиктивный контур, механизм внимания и *PhysProfile*.

Каждую стратегию проверяли по одной логике: улучшает ли она целевой профиль и превосходит ли соответствующий отрицательный контроль. Если локальный прирост воспроизводится случайным, равномерным, перемешанным или упрощённым вариантом того же подхода, то утверждение о специфичном для метода механизме отклоняется. Отрицательный результат в таком протоколе ограничивает допустимую интерпретацию эффекта: локальный прирост метрики не считается подтверждённым повышением устойчивости, если он воспроизводится отрицательным контролем.

Сводная матрица проверенных стратегий приведена в таблице 7.

Табл. 7. Проверяемые стратегии и границы допустимых выводов

Стратегия	**Отрицательный контроль**	**Результат проверки**	**Допустимый вывод**	**Недопустимый вывод**
ECG-dropout	Случайное модальное отключение	В режиме без ЭКГ качество растёт относительно базовой модели и контрольной конфигурации, но качество в исходном режиме снижается	*ECG-dropout* локально снижает зависимость от ЭКГ	*ECG-dropout* является новой универсальной базовой конфигурацией
Перевзвешивание трудных событий	Случайные или равномерные веса	Устойчивого преимущества над контрольным перевзвешиванием не получено	Метод не подтверждён как самостоятельное средство повышения устойчивости	Трудные события можно исправить простым увеличением веса в обучении
XAI/*conflict*-перевзвешивание	Перемешанные или случайные веса	Эффект не превосходит контрольные варианты	В данной постановке метод не даёт основания для сильного утверждения	*XAI*-веса являются обоснованным механизмом повышения устойчивости модели
JEPA-подобный контур	Упрощённый контур только с дистилляцией и JEPA-контроли	Приросты не превосходят контрольные варианты	*JEPA*-контур не подтверждён как способ снижения хрупкости	Латентно-предиктивное обучение решает проблему стрессовой деградации

Механизм внимания	Равномерное, случайное или перемешанное внимание	Локальный выигрыш в отдельных режимах не превосходит контрольные варианты; часть профилей устойчивости ухудшается	Механизм обучаемого внимания не подтверждён как источник устойчивого улучшения	Карты внимания можно трактовать как объяснение решения модели
PhysProfile	Нейросетевые признаки ошибки	*PhysProfile* уступает нейросетевым признакам при прямом предсказании ошибок, но даёт предметно читаемое описание конфликтных событий	*PhysProfile* полезен как язык разбора фенотипов отказа	*PhysProfile* является лучшим предиктором ошибок или клиническим правилом

Таблица 7 ограничивает допустимые выводы: *ECG-dropout* даёт подтверждённый локальный эффект, тогда как перевзвешивание, *XAI*-перевзвешивание, *JEPA*-подобный контур и механизм внимания не подтверждаются как способы снижения хрупкости, поскольку не превосходят отрицательные контроли или ухудшают отдельные профили. *PhysProfile* не используется как критерий правильности решения; при меньшей точности по сравнению с нейросетевыми признаками ошибки он сохраняет роль языка описания фенотипов отказа — слабой гемодинамической поддержки, межканального расхождения, низкого качества сигнала и пограничных событий.

7. Дополнительная проверка устойчивости

Для сопоставления многоосевой схемы с более общими проверками надёжности была выполнена дополнительная стресс-проверка *ResNetFusion* на *VTaC*. Она включала физиологически интерпретируемые возмущения сигналов и отдельные алгоритмические атаки чувствительности. Эта проверка не является полной оценкой устойчивости к состязательным возмущениям или внекорпусным примерам: синтетические возмущения применялись после нормализации, а *FGSM*/*PGD* использовались только для оценки чувствительности выходов модели к малым направленным возмущениям входа.

Табл. 8. Дополнительная проверка устойчивости ResNetFusion

Проверка	**ROC-AUC**	**Доля пропущенных истинных тревог**	**Средняя уверенность**	**Интерпретация**
Исходный режим	0,9592	0,0511	0,8951	базовый уровень
Без ЭКГ	0,7650	0,3358	0,7360	выраженная модальная хрупкость
Сдвиг окна 5 с	0,7788	0,5766	0,8909	временная хрупкость и уверенные пропуски
Сдвиг окна 20 с	0,6093	0,8394	0,9009	выход из устойчивой области
Шум ЭКГ 10 дБ	0,8516	0,9781	0,9679	чувствительность к зашумлению ЭКГ
Шум всех каналов 10 дБ	0,8661	0,9927	0,9854	уверенная деградация при общем шуме
Обнуление участка ЭКГ 2 с	0,8936	0,2336	0,8603	локальная потеря ЭКГ ухудшает качество
Межканальный лаг ФПГ/ИАД 1 с	0,9578	0,0511	0,8920	слабый эффект в данном протоколе
FGSM $\varepsilon = 0{,}01$	0,9339	0,1022	0,8760	алгоритмическая чувствительность, не физиологический сценарий

Шум с отношением сигнал/шум 10 дБ использовался как контролируемое возмущение нормализованных входных сигналов и не моделирует конкретный физиологический артефакт. Наиболее сильная деградация связана с потерей или искажением ЭКГ, смещением предтревожного окна и сильным синтетическим шумом; при таких режимах модель может сохранять высокую

уверенность в ложноотрицательных решениях. Межканальный лаг ФПГ/ИАД почти не меняет качество, что согласуется с выводом о доминирующей роли ЭКГ.

8. Обсуждение результатов

Сравнение архитектур использовалось для оценки характера деградации, а не для максимизации метрик. Поскольку исходное качество моделей различается, основной интерес представляет не абсолютный *ROC-AUC*, а изменение качества при нарушении входного контекста. Для временного сдвига дополнительно учитывалась дискретная скорость деградации — отношение падения метрики к величине смещения окна.

Феномен модального доминирования ЭКГ может быть связан с архитектурой позднего слияния модальностей. В *ResNetFusion* одномерные остаточные сверточные ветви извлекают признаки раздельно для каждого канала; затем латентные векторы ЭКГ, ФПГ и ИАД конкатенируются непосредственно перед финальным линейным слоем. Такая топология ограничивает раннее кодирование кросс-модальных динамических зависимостей: фазового сдвига между ЭКГ и пульсовой волной, задержки ФПГ/ИАД относительно QRS-комплекса и гемодинамического подтверждения тахикардического ритма. В результате гемодинамические каналы могут не компенсировать деградацию ЭКГ-ветви, хотя формально присутствуют во входном наборе. Перспективным направлением дальнейших исследований является переход от позднего слияния к раннему или промежуточному слиянию модальностей, при котором межканальные отношения доступны модели до финальной классификации.

Заключение

Установлено, что высокая метрика *ResNetFusion* на исходном тестовом режиме не гарантирует устойчивости модели при изменении входного контекста. При *ROC-AUC* = 0,9592 модель демонстрирует выраженную зависимость от ЭКГ, резкую деградацию при смещении предтревожного окна и рост уверенно ошибочных решений на сложных событиях.

Для временной хрупкости выявлена практическая граница применимости модели. При смещении окна на 2,5 с качество остаётся близким к исходному, но уже при смещении на 5 с *ROC-AUC* снижается до 0,7788, а доля пропущенных истинных тревог возрастает до 0,5766. Более ранние предтревожные интервалы нельзя использовать без отдельной проверки качества и калибровки.

Выявлено, что направленная регуляризация *ECG-dropout* снижает модальное доминирование ЭКГ: в режиме без ЭКГ *ROC-AUC* возрастает на +0,0185 относительно базовой модели и на +0,0100 относительно случайного контроля. Этот эффект сопряжён со снижением сбалансированной точности в исходном режиме на −0,0200 и задаёт границу применимости *ECG-dropout*: метод работает как селективный способ повышения устойчивости при потере доминирующего сигнала, а не как универсальное улучшение модели.

Анализ сложных событий выявил режим уверенных ошибок. На сложном подмножестве доля ошибок достигает 0,6197 при средней уверенности 0,8230. При временных сдвигах основной риск связан с уверенными пропусками истинных тревог ЖТ.

Проверка модификаций с отрицательными контролями ограничила допустимые выводы. Перевзвешивание, *XAI*-перевзвешивание, *JEPA*-подобный контур и механизм внимания в данной постановке не подтвердились как самостоятельные способы снижения хрупкости. Их локальные эффекты не превосходили соответствующие контрольные варианты либо сопровождались ухудшением других профилей.

PhysProfile не используется как самостоятельный предиктор ошибок и не заменяет клиническую интерпретацию; его роль ограничена описанием физиологического контекста проблемных событий — расхождения ЭКГ и гемодинамики, слабой поддержки ФПГ/ИАД, неполноты каналов и локализованности признаков во времени. Внешняя проверка на подвыборке *PhysioNet*/*CinC Challenge* 2015 не доказывает универсальность модели, но подтверждает применимость схемы оценки к независимому источнику ОРИТ-тревог. Основной вывод работы состоит в том, что средняя метрика медицинской мультимодальной модели должна дополняться проверкой измеримых условий отказа:

потери модальностей, смещения временного окна, трудных событий и нарушения калибровки. Без такой проверки модель с *ROC-AUC* около 0,96 в исходном режиме может в стрессовом режиме пропускать более 80% истинных тревог желудочковой тахикардии при сохранении средней уверенности около 0,90.

Список источников

1. VTaC: a benchmark dataset of ventricular tachycardia alarms from ICU monitors / L.-W. H. Lehman., B. Moody, H. Deep [et al.] // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 38827–38843. https://doi.org/10.52202/075280-1686.
2. VTaC: a benchmark dataset of ventricular tachycardia alarms from ICU monitors. Version 1.0.0 / L.-W. H. Lehman., B. Moody, H. Deep [et al.] // PhysioNet. 2024. https://doi.org/10.13026/z4f3-1f07.
3. The PhysioNet/Computing in Cardiology Challenge 2015: reducing false arrhythmia alarms in the ICU / G. D. Clifford, I. Silva, B. Moody [et al.] // Computing in Cardiology. 2015. Vol. 42. P. 273–276. https://doi.org/10.1109/cic.2015.7408639.
4. Fallet S., Yazdani S., Vesin J.-M. A multimodal approach to reduce false arrhythmia alarms in the intensive care unit // Computing in Cardiology. 2015. Vol. 42. P. 277–280. https://doi.org/10.1109/cic.2015.7408640.
5. Reducing false arrhythmia alarms in the ICU / S.-K. Teo, J. C. Wong, B. Yang B. [et al.] // Computing in Cardiology. 2015. Vol. 42. P. 1177–1180. https://doi.org/10.1109/cic.2015.7411126.
6. Reducing false arrhythmia alarms in the ICU using multimodal signals and robust QRS detection / N. Sadr, J. Huvanandana, D. T. Nguyen [et al.] // Physiological Measurement. 2016. Vol. 37, Iss. 8. P. 1340–1354. https://doi.org/10.1088/0967-3334/37/8/1340.
7. Mousavi S., Fotoohinasab A., Afghah F. Single-modal and multi-modal false arrhythmia alarm reduction using attention-based convolutional and recurrent neural networks // PLOS ONE. 2020. Vol. 15, Iss.1. P. e0226990. https://doi.org/10.1371/journal.pone.0226990.
8. Мультимодальная объяснимость для ОРИТ-сигналов (VTaC): метрические и асимптотические результаты / Ю. В. Трофимов, А. Н. Аверкин, Е. М. Кузнецов [и др.] // Вестник Тверского государственного университета. Серия: Прикладная математика. 2025. № 4. С. 43–80. https://doi.org/10.26456/vtpmk757.
9. ModDrop: adaptive multi-modal gesture recognition / N. Neverova, C. Wolf, G. W. Taylor, F. Nebout // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016. Vol. 38, Iss. 8. P. 1692–1706. https://doi.org/10.1109/TPAMI.2015.2461544.
10. On calibration of modern neural networks / C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger // Proceedings of the 34th International Conference on Machine Learning. 2017. Vol. 70. P. 1321–1330.
11. Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks // Proceedings of the 34th International Conference on Machine Learning. 2017. Vol. 70. P. 3319–3328.
12. Sanity checks for saliency maps / J. Adebayo, J. Gilmer, M. Muelly [et al.] // Advances in Neural Information Processing Systems. 2018. Vol. 32. P. 9525–9536.
13. A benchmark for interpretability methods in deep neural networks / S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim // Advances in Neural Information Processing Systems. 2019. Vol. 32. P. 9737–9748.
14. Jain S., Wallace B. C. Attention is not Explanation // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019. P. 3543–3556. https://doi.org/10.18653/v1/n19-1357.
15. Wiegreffe S., Pinter Y. Attention is not not Explanation // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. P. 11–20. https://doi.org/10.18653/v1/D19-1002.

16. Self-supervised learning from images with a joint-embedding predictive architecture / M. Assran, Q. Duval, I. Misra [et al.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. P. 15619–15629. https://doi.org/10.1109/cvpr52729.2023.01499.

17. Revisiting feature prediction for learning visual representations from video / A. Bardes, Q. Garrido, J. Ponce J. [et al.] // arXiv : [electronic archive]. arXiv:2404.08471. https://doi.org/10.48550/arXiv.2404.08471. Submitted on 15 Feb 2024.

18. TS2Vec: towards universal representation of time series / Z. Yue, Y. Wang, J. Duan. [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. 2022. Vol. 36, Iss. 8. P. 8980–8987. https://doi.org/10.1609/aaai.v36i8.20881.