ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.8, 004.89, 61:004.8

ВРЕМЕННАЯ АТРИБУЦИЯ В СТОХАСТИЧЕСКИХ ДИФФУЗИОННЫХ ПРОЦЕССАХ: МАТЕМАТИЧЕСКАЯ ФОРМАЛИЗАЦИЯ, ОБЪЯСНИМОСТЬ И КАУЗАЛЬНАЯ ВАЛИДАЦИЯ[1]

Трофимов Ю.В.[*,**], **Аверкин А.Н.**[***,*], **Лопатин М.А.**[*], **Трусов И.А.**[*], **Муравьев И.П.**[*], **Ильин А.С.**[****,*], **Шевченко А.В.**[*]

[*]Государственный университет «Дубна», г. Дубна

[**]Объединённый институт ядерных исследований, г. Дубна

[***]ФИЦ «Информатика и управление» РАН, г. Москва

[****]Университет Иннополис, г. Иннополис

Поступила в редакцию 20.09.2025, после переработки 29.03.2026.

В работе решается проблема объяснимости диффузионных моделей типа DDPM для медицинской диагностики через разработку математически строгой системы временной атрибуции. Существующие методы XAI не учитывают стохастическую динамику итеративных генеративных процессов и специфику формирования диагностически значимых признаков в процессе денойзинга. Предложена адаптация фреймворка TimeSHAP [54] для диффузионных процессов — метод TimeSHAP-Diff, расширяющий классические значения Шепли на временную размерность стохастических генеративных моделей через временную функцию ценности $v^t : 2^T \to \mathbb{R}$. Доказана математическая консистентность подхода через выполнение аксиом эффективности, симметрии, нулевого игрока и аддитивности. Экспериментальная валидация на датасете ISIC2018 (10,015 дерматоскопических изображений, 7 классов) продемонстрировала эффективность диффузионной компенсации дисбаланса классов: коэффициент дисбаланса снижен с 58.30 до 1.49 (улучшение 97.4%), accuracy повышена с 93.2% до 97.1%. Оптимально сбалансированный датасет обеспечил 100% выполнение аксиом Шепли против 42.9% для синтетических данных. Каузальная валидация через контрафактуальные интервенции подтвердила статистически значимую разность влияния регионов важности (ΔCFI=0.711, p$<$0.001). Выявлены класс-специфичные временные профили формирования диагностических признаков: раннее проявление для меланомы ($t \approx 900$) и запаздывающая консолидация для редких классов (t$>$950). Практическая значимость подтверждается созданием открытых систем SYNT_ISIC и CAS_ISIC с полной воспроизводимостью результатов для клинических приложений.

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

Ключевые слова: диффузионные процессы, причинно-следственный вывод, временной сдвиг, интервенционный анализ, стохастическая аппроксимация, медицинская диагностика.

Вестник ТвГУ. Серия: Прикладная математика. 2026. № 1. С. 89–137.
https://doi.org/10.26456/vtpmk773

Введение

Развитие генеративных моделей глубокого обучения в последние годы привело к созданию высокоэффективных алгоритмов синтеза изображений, среди которых диффузионные модели типа Denoising Diffusion Probabilistic Models (DDPM) демонстрируют превосходное качество генерации [16]. Однако для клинических приложений ключевым аспектом становится интеграция предсказательной мощности с формализованной системой верифицированной интерпретируемости результатов [1, 2].

Современные методы объяснимого искусственного интеллекта (Explainable AI, XAI), такие как Integrated Gradients [3], SHAP [4] и Grad-CAM [5], разработаны преимущественно для дискриминативных моделей и не учитывают специфику итеративных генеративных процессов. Диффузионные модели характеризуются сложной временной динамикой: процесс денойзинга включает множество последовательных шагов, на каждом из которых формируются различные уровни семантической информации [6].

Существующие XAI подходы для генеративных моделей [7, 8] фокусируются на финальном результате, игнорируя временную эволюцию важности признаков. Такой разрыв приводит к недостаточности объяснительных возможностей: репрезентативная интерпретация решений генеративной архитектуры возможна лишь при формальном учёте динамических аспектов формирования диагностических структур на последовательных этапах денойзинга.

В медицинской диагностике дерматологических заболеваний точная интерпретация решений ИИ-систем критична для клинического применения [9]. Для решения данных задач широко используется датасет ISIC (International Skin Imaging Collaboration), который содержит высококачественные дерматоскопические изображения с экспертной разметкой семи классов поражений кожи, включая злокачественные новообразования [10]. Тем не менее, применение традиционных подходов машинного обучения к этим данным осложняется существенным дисбалансом классов, характерным для реальных медицинских датасетов. Таким образом, разработка синтетических генеративных техник должна сочетать адекватное восполнение репрезентативности редких классов и обеспечивать прозрачность, позволяющую гарантировать клиническую надежность и интерпретируемость решений.

Цель настоящей работы — разработка математически строгой системы объяснимого анализа диффузионных моделей для медицинской диагностики, включающей: (1) адаптацию фреймворка TimeSHAP [54] для анализа динамики важности признаков в диффузионных процессах; (2) каузальную валидацию XAI результатов через контрафактуальные интервенции; (3) интегрированную архитектуру генерация-классификация-сегментация с комплексной объяснимостью.

Основные научные вклады работы: адаптация временного анализа важности TimeSHAP для диффузионных процессов через формулировку функции ценности для стохастических генеративных переходов; доказательство консистентности предложенного подхода согласно теореме Шепли [13]; экспериментальная валидация каузальных связей между выделенными регионами и диагностическими решениями; создание открытой программной реализации для воспроизводимости результатов.

Наш метод следует общей идее TimeSHAP — рассматривать временные элементы как игроков кооперативной игры Шепли и измерять их маргинальный вклад к целевой функции. Первоначально TimeSHAP был предложен Bento et al. [54] для рекуррентных моделей и последовательных данных с маскированием событий и отсечением малозначимых префиксов. В нашей работе мы переносим этот фреймворк на генеративные диффузионные процессы, где "временные игроки"— это шаги обратной диффузии, а функция ценности определяется через частичный денойзинг и интервенции на оставшихся шагах. Мы не претендуем на новизну в самой идее временных значений Шепли: она восходит к TimeSHAP. Наш вклад — формулировка функции ценности для диффузионных переходов, каузальные интервенции для валидации и эмпирический анализ в задачах генерации/синтеза медицинских данных.

1. Справедливое распределение вклада в атрибуционных методах XAI

Фундаментальной проблемой разработки объяснимых систем искусственного интеллекта является справедливое распределение важности между различными компонентами входной информации, временными интервалами или структурными элементами модели. Данная задача требует строгого математического обоснования принципов атрибуции, которые должны удовлетворять определенным аксиоматическим требованиям независимо от конкретной архитектуры модели, типа данных или метода объяснения [3, 4, 13].

Исторически теоретические основы справедливого распределения восходят к классическим работам в области теории кооперативных игр [13, 56, 58], экономической теории благосостояния [61, 62] и теории принятия коллективных решений [59, 60], где центральным вопросом является справедливое распределение общего результата между участниками коллективного процесса [57]. Ключевое наблюдение заключается в том, что любая система атрибуции, будь то для нейронных сетей, временных рядов или диффузионных моделей, должна опираться на универсальные принципы справедливости, формализованные через аксиоматические системы [3, 13, 63].

1.1. Универсальная аксиоматическая система атрибуции

Для формализации понятия справедливого распределения важности введем обобщенную математическую структуру, охватывающую различные типы объяснимых систем.

Определение 1 (Универсальное пространство атрибуции)**.** *Пусть $\mathcal{N}$ — множество элементов (признаков, временных шагов, компонентов модели), $\mathcal{V}$ — класс*

функций ценности $v : 2^{\mathcal{N}} \to \mathbb{R}$, *и* $\mathcal{F}$ — *множество интерпретируемых моделей. Универсальная функция атрибуции определяется как:*

$$\Phi : \mathcal{V} \times \mathcal{N} \times \mathcal{F} \to \mathbb{R}^{|\mathcal{N}|}.$$

Определение 2 (Фундаментальные аксиомы справедливого распределения [4, 13])**.** *Корректная система атрибуции* Φ *должна удовлетворять следующим универсальным принципам:*

Аксиома полноты (эффективности): *Сумма всех индивидуальных атрибуций равна общему эффекту системы:*

$$\sum_{i \in \mathcal{N}} \phi_i(v) = v(\mathcal{N}) - v(\emptyset).$$

Аксиома симметрии: *Элементы с одинаковыми маргинальными вкладами получают равные атрибуции:*

$$\forall S \subseteq \mathcal{N} \setminus \{i, j\} : v(S \cup \{i\}) - v(S) = v(S \cup \{j\}) - v(S) \Rightarrow \phi_i(v) = \phi_j(v).$$

Аксиома нулевого вклада: *Элементы, не влияющие на результат, получают нулевую атрибуцию:*

$$\forall S \subseteq \mathcal{N} \setminus \{i\} : v(S \cup \{i\}) = v(S) \Rightarrow \phi_i(v) = 0.$$

Аксиома аддитивности (линейности): *Атрибуции для комбинации задач равны сумме атрибуций:*

$$\forall v, w \in \mathcal{V} : \phi_i(v + w) = \phi_i(v) + \phi_i(w).$$

Математический смысл аксиом подробно изложен в классических работах [13, 23]. Аксиома полноты обеспечивает консервативность системы атрибуции: вся важность распределяется без потерь или избытка. Аксиома симметрии гарантирует справедливость: элементы с одинаковым функциональным вкладом получают равное признание. Аксиома нулевого вклада исключает фантомную важность: элементы без влияния не получают атрибуции. Аксиома аддитивности обеспечивает композиционность: система корректно работает с комбинациями задач.

1.2. Классическая теорема Шепли и её математические следствия

Фундаментальным результатом теории справедливого распределения является доказательство единственности системы атрибуции, удовлетворяющей указанным аксиомам [13].

Теорема 1 (Теорема Шепли о единственности, 1953 [13])**.** *Функция атрибуции* Φ*, удовлетворяющая аксиомам полноты, симметрии, нулевого вклада и аддитивности, существует и единственна. Она определяется формулой:*

$$\phi_i(v) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|S|!(|\mathcal{N}| - |S| - 1)!}{|\mathcal{N}|!} [v(S \cup \{i\}) - v(S)].$$

Доказательство данной теоремы приведено в оригинальной работе Шепли [13] и подробно разобрано в современных учебниках по теории игр [23, 59].

Интерпретация коэффициентов дана в классических работах [13, 58]. Биномиальный коэффициент $\frac{|S|!(|\mathcal{N}|-|S|-1)!}{|\mathcal{N}|!}$ представляет вероятность того, что коалиция S формируется до присоединения элемента i при случайном порядке вступления участников. Это обеспечивает равноправное усреднение по всем возможным сценариям коалиционного формирования.

1.3. Расширения Фридмана для экономических приложений

Работы Friedman [55] расширили классическую теорию Шепли на экономические задачи распределения затрат и ответственности в многоагентных системах.

Определение 3 (Расширенная система Фридмана [55]). *Для задач с внешними эффектами и неаддитивными взаимодействиями функция ценности может включать корреляционные термы:*

$$v_{Friedman}(S) = v_{base}(S) + \sum_{i<j\in S} \rho_{ij} + \sum_{i<j<k\in S} \tau_{ijk} + \dots,$$

где ρ_{ij} — парные взаимодействия, τ_{ijk} — тройные взаимодействия.

Лемма 1 (Консистентность расширения Фридмана [55]). *Формула Шепли остается единственным решением аксиоматической системы даже при наличии взаимодействий высших порядков, если функция ценности остается супераддитивной: $v(S \cup T) \geq v(S) + v(T)$ для непересекающихся S, T.*

Практическое значение расширения Фридмана критически важно для медицинских приложений [43], где диагностические признаки часто демонстрируют синергетические эффекты: комбинация симптомов может быть более информативной, чем сумма индивидуальных вкладов.

1.4. Градиентные методы и аксиоматика Сундарараджана

Альтернативный подход к справедливому распределению важности основан на градиентной информации и интегральных путях в пространстве признаков [3].

Определение 4 (Integrated Gradients [3]). *Для дифференцируемой функции $f : \mathbb{R}^d \to \mathbb{R}$ и базовой точки $x' \in \mathbb{R}^d$ метод Integrated Gradients определяется как:*

$$IG_i(x) = (x_i - x'_i) \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha.$$

Теорема 2 (Аксиоматическая консистентность IG [3]). *Метод Integrated Gradients удовлетворяет следующим аксиомам:*

Чувствительность: *Если $f(x) \neq f(x')$ и отличие обусловлено только i-м признаком, то $IG_i(x) \neq 0$.*

Инвариантность реализации: *Если две функции f_1, f_2 эквивалентны на области определения, то $IG_i^{f_1}(x) = IG_i^{f_2}(x)$.*

Полнота: $\sum_{i=1}^{d} IG_i(x) = f(x) - f(x')$.

Доказательство полноты приведено в оригинальной работе [3] и основано на основной теореме анализа.

Связь с теорией Шепли подробно изучена в работах [3,4]. Sundararjan et al. показали, что Integrated Gradients представляет непрерывный аналог значений Шепли для гладких функций, где интегрирование по прямолинейному пути заменяет дискретное суммирование по коалициям.

1.5. LIME и локальная интерпретируемость

Метод LIME [25] представляет третий фундаментальный подход к справедливому распределению важности через локальное суррогатное моделирование.

Определение 5 (LIME функция объяснения [25]). *Для сложной модели f и интерпретируемого класса G LIME решает оптимизационную задачу:*

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

где: - $\mathcal{L}(f, g, \pi_x) = \sum_{z \in \mathcal{Z}} \pi_x(z)(f(z) - g(z))^2$ — взвешенная ошибка аппроксимации, - $\pi_x(z) = \exp(-\frac{||x-z||^2}{2\sigma^2})$ — весовая функция близости, - $\Omega(g)$ — регуляризатор сложности модели.

Лемма 2 (Локальная справедливость LIME [23, 25]). *В окрестности $\mathcal{B}_\epsilon(x) = \{z : ||z - x|| < \epsilon\}$ суррогатная модель g^* удовлетворяет аппроксимированной аксиоме полноты:*

$$\left| \sum_{i=1}^{d} w_i^{LIME} - [f(x) - f(x_{baseline})] \right| \leq \mathcal{O}(\epsilon^2),$$

где w_i^{LIME} — веса линейной суррогатной модели.

Сравнительный анализ методов представлен в обзорных работах [23,32,33]. Три рассмотренных подхода представляют различные философии атрибуции: - Shapley Values: глобальная кооперативная справедливость через исчерпывающий анализ коалиций [4,13] - Integrated Gradients: путевая интеграция градиентной информации от базовой к целевой точке [3] - LIME: локальная линейная аппроксимация в окрестности интереса [25]

1.6. Временные и каузальные расширения

Анализ процессов с временной динамикой требует расширения классических аксиом на временную размерность. Данная задача была решена в работе Bento et al. [54] путем модификации принципов атрибуции для последовательных данных.

Определение 6 (Временная функция ценности [54]). *Для процесса с временными шагами $\mathcal{T} = \{1, 2, \ldots, T\}$ временная функция ценности $v^{temporal} : 2^{\mathcal{T}} \to \mathbb{R}$ определяет вклад различных временных коалиций в итоговый результат:*

$$v^{temporal}(S) = \mathbb{E}[Outcome | Active\_timesteps = S] - \mathbb{E}[Outcome | Active\_timesteps = \emptyset].$$

В работе Bento et al. [54] впервые предложена адаптация принципов Шепли для временных рядов, где "игроками"становятся временные интервалы, а функция ценности отражает изменение предсказания при маскировании определенных временных сегментов.

Каузальные аспекты атрибуции, введенные Pearl [24] и развитые в работах [43, 63], требуют дополнительной валидации через интервенционные эксперименты для подтверждения причинно-следственных связей между выделенными элементами и итоговыми решениями.

Определение 7 (Каузальная атрибуция через do-калькулус [24])**.** *Каузальная важность элемента i определяется через do-операцию:*

$$Causal\_Attribution_i = \mathbb{E}[Y|do(X_i = x_i^{high})] - \mathbb{E}[Y|do(X_i = x_i^{low})],$$

где do-операция моделирует интервенционное воздействие на i-й признак.

Практическая значимость временной атрибуции в медицинских приложениях подтверждается недавними исследованиями Bhattacharyay et al. [64, 65], где метод TimeSHAP успешно применен для анализа динамики лечения травматических повреждений головного мозга. В работе 2025 года [64] TimeSHAP использован для предсказания изменений интенсивности лечения внутричерепного давления, демонстрируя объяснительную способность до 68% ординальной вариации на первый день лечения. Исследование 2023 года [65] показало эффективность временной атрибуции для выявления вклада клинических переменных в функциональные исходы через 6 месяцев после травмы, что подчеркивает универсальность подходов временной интерпретируемости в критически важных медицинских применениях.

1.7. Математические свойства и вычислительная сложность

Теоретические гарантии различных методов XAI подробно изучены в литературе [23, 32]. Все рассмотренные методы обеспечивают теоретически обоснованное распределение важности, но различаются вычислительной сложностью:

- Точные значения Шепли: $\mathcal{O}(2^{|\mathcal{N}|})$ — экспоненциальная сложность [13] - Аппроксимированные значения Шепли: $\mathcal{O}(M \cdot |\mathcal{N}|)$ где M — число выборок коалиций [4] - Integrated Gradients: $\mathcal{O}(m \cdot |\mathcal{N}|)$ где m — число шагов интегрирования [3] - LIME: $\mathcal{O}(K \cdot |\mathcal{N}|)$ где K — размер локальной выборки [25]

Теорема 3 (Аппроксимационные гарантии для значений Шепли [4])**.** *Для заданной точности ϵ и доверительной вероятности δ достаточно*

$$M \geq \frac{2\log(2|\mathcal{N}|/\delta)}{\epsilon^2} \cdot \max_{i,S} |v(S \cup \{i\}) - v(S)|^2$$

случайных коалиций для ϵ-аппроксимации всех значений Шепли с вероятностью $1 - \delta$.

1.8 Интеграция принципов для диффузионных моделей

Представленная аксиоматическая основа создает теоретический фундамент для разработки специализированных XAI методов для диффузионных моделей [6, 16]. Ключевые принципы, которые будут адаптированы в последующих разделах:

1. Временная полнота: $\sum_{t=1}^{T} \phi^t = v^{temporal}(\mathcal{T}) - v^{temporal}(\emptyset)$ [54]

2. Стохастическая справедливость: элементы с одинаковыми стохастическими вкладами получают равные атрибуции [13]

3. Марковская аддитивность: атрибуции для композиции марковских процессов [18]

4. Каузальная валидируемость: атрибуции должны выдерживать интервенционное тестирование [24, 43]

Данная универсальная система обеспечивает математическую строгость и теоретическую обоснованность всех последующих XAI методов, разрабатываемых в настоящей работе.

2. Обзор методов объяснимого искусственного интеллекта

Объяснимый искусственный интеллект (Explainable AI, XAI) представляет собой совокупность методов и техник, направленных на формирование понятных и верифицируемых объяснений решений моделей машинного обучения для конечных пользователей [2]. Фундаментальное различие между интерпретируемостью и объяснимостью заключается в их различной эпистемологической и функциональной направленности. Интерпретируемость (*interpretability*) определяется как способность алгоритмической модели выявлять внутренние зависимости между входными признаками и выходными предсказаниями, включая каузальные механизмы, лежащие в основе принятия решений [32,33]. Иными словами, интерпретируемость фокусируется на аналитической реконструкции онтологической структуры модели и выделении факторов, оказывающих определяющее влияние на результат [35].

Объяснимость (*explainability*), напротив, ориентирована на процесс трансляции полученной информации в форму, когнитивно и семантически совместимую с восприятием целевой аудитории [2]. В этом контексте объяснимость выполняет роль коммуникативного интерфейса между внутренними алгоритмическими механизмами и специалистами-практиками, обеспечивая как понимание, так и доверие к результатам работы модели [36].

В медицинских приложениях объяснимость представляет критическое требование: врачу необходимо понимать, какие анатомические области и признаки детерминируют вывод модели, и получать результаты в форме, согласованной с профессиональными онтологиями и клиническими протоколами. Эмпирические данные демонстрируют, что корректно спроектированные объяснения способны повышать эффективность взаимодействия врача с ИИ: в мультицентровом рандомизированном исследовании ($n = 220$) локальные объяснения обеспечили более высокую диагностическую точность при верном совете ИИ (92,8% против 85,3% для глобальных; $p_{adj} < 0,001$) и ускорили принятие решения ($\beta = -0,19$; $p_{adj} = 0,01$) [37].

Одновременно выявлены ограничения пост-хок визуализаций: при интерпретации CXR тепловые карты существенно уступают экспертному бенчмарку по локализации, причём разрыв максимален для малых и морфологически сложных объектов (например, для *support devices* Grad-CAM: $mIoU = 0,163$; $hit\text{-}rate = 0,355$), что ставит под вопрос их пригодность для валидации единичных клинических

решений [38]. Схожие выводы получены и для задач маммографии при количественной оценке *saliency*-методов [39].

Крупномасштабное исследование (140 радиологов, 15 задач) фиксирует выраженную гетерогенность эффекта ИИ-поддержки: ошибки ИИ систематически ухудшают работу по совокупности задач и половине отдельных нозологий, что требует адаптивного дизайна интерфейсов объяснений с учётом качества/уверенности модели и профиля пользователя [40]. При этом систематические обзоры указывают, что XAI чаще повышает доверие врача к системе — при условии краткости, релевантности и ясности объяснений, — однако качество доказательств варьирует [41–43].

2.1. Классические методы атрибуции

Основу современных XAI подходов составляют методы атрибуции, определяющие вклад каждого входного признака в итоговое решение модели.

Определение 8 (Функция атрибуции [4, 23])**.** *Функция атрибуции* $\Phi : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}^d$ *для модели* $f \in \mathcal{F}$ *и входа* $x \in \mathbb{R}^d$ *определяет важность каждого признака* x_i *относительно выхода* $f(x)$*.*

Integrated Gradients. Метод Integrated Gradients [3] основан на интегрировании градиентов вдоль прямолинейного пути от базовой точки x' до исследуемой точки x:

$$IG_i(x) = (x_i - x'_i) \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha,$$

где x' — базовое изображение (обычно нулевое), f — функция модели. Метод удовлетворяет аксиомам Sensitivity и Implementation Invariance:

Лемма 3 (Аксиома эффективности для Integrated Gradients [3])**.** *Для дифференцируемой функции* f *метод Integrated Gradients удовлетворяет аксиоме эффективности:* $\sum_{i=1}^{d} IG_i(x) = f(x) - f(x')$*.*

SHAP (SHapley Additive exPlanations). Метод SHAP [4] базируется на теории кооперативных игр и значениях Шепли. Для коалиционной функции $v : 2^N \to \mathbb{R}$ значение Шепли для игрока i определяется как:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)].$$

В контексте XAI коалиции S соответствуют подмножествам признаков, а функция ценности определяется как $v(S) = \mathbb{E}[f(x)|x_S]$.

Теорема 4 (Единственность значений SHAP [4,13])**.** *Значения SHAP единственным образом удовлетворяют аксиомам эффективности, симметрии, нулевого игрока и аддитивности.*

Grad-CAM. Метод Grad-CAM [5] для сверточных нейронных сетей вычисляет карты активации через градиенты последнего сверточного слоя:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k w^c_k A^k\right),$$

где веса важности определяются как:

$$w^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k_{ij}}.$$

Здесь y^c — оценка класса c, A^k — карта активации k-го канала, Z — нормализующий коэффициент.

LIME (Local Interpretable Model-agnostic Explanations). Метод LIME [25] направлен на локальную интерпретацию решений любой модели, в том числе сложных и неинтерпретируемых, путём построения интерпретируемого приближенного локального суррогатного модельного поведения в окрестности исследуемого объекта. Для входного объекта $x \in \mathbb{R}^d$ LIME формирует множество случайных вариаций x' путём локальных возмущений признаков и обучает простую интерпретируемую модель $g \in G$ (например, линейную регрессию) на этих данных, стремясь максимизировать локальную точность при минимальной сложности задачи приближения:

$$\xi(x) = \arg\min_{g\in G} \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

где

- f — исходная модель,
- π_x — функция, задающая весовые коэффициенты, отражающие близость примеров к исследуемому объекту x,
- $\mathcal{L}(f, g, \pi_x)$ — функция меры несоответствия между f и g с учётом весов π_x,
- $\Omega(g)$ — мера сложности интерпретируемой модели g.

Результатом является оценка важности каждого признака исходного объекта в локальном регионе, обеспечивающая интуитивно понятное объяснение решения модели.

2.2. Ограничения XAI для генеративных моделей

Существующие методы XAI разработаны преимущественно для дискриминативных моделей и демонстрируют принципиальные ограничения при применении к генеративным архитектурам [7].

Замечание 1. Классические методы атрибуции не учитывают итеративную природу диффузионных процессов, где каждый временной шаг t формирует различные уровни семантической информации.

Для диффузионных моделей процесс генерации описывается последовательностью преобразований $x_T \to x_{T-1} \to \ldots \to x_0$, где каждый шаг характеризуется собственной динамикой формирования признаков и уровнем семантической информации. Традиционные XAI-методы оценивают атрибуцию исключительно в

финальном состоянии x_0, тем самым игнорируя эволюцию признаков во времени и утрачивая информацию о промежуточных стадиях денойзинга.

Проблема композиционности. Применение Grad-CAM к диффузионным моделям требует композиции градиентов через цепочку денойзинга:

$$\frac{\partial F}{\partial x_T} = \frac{\partial F}{\partial x_0} \prod_{t=1}^{T} \frac{\partial x_{t-1}}{\partial x_t},$$

где произведение якобианов может приводить к численной нестабильности и потере градиентов [8].

Временная зависимость атрибуции. В диффузионных процессах величина важности отдельного пикселя (i, j) на временном шаге t формируется под воздействием как локальных признаков, так и комплексной глобальной временной эволюции, отражающей многомасштабные динамические зависимости модели. Существующие методы не предоставляют инструментов для анализа эволюции важности во времени.

2.3. Обзор современных исследований в области объяснимости генеративных моделей

Исследования в области объяснимости генеративных моделей в последние годы демонстрируют устойчивый рост, однако сохраняются фундаментальные методологические ограничения. Одно из первых направлений связано со *score-based* методами: Wang et al. предложили выделение информативных регионов через анализ функции правдоподобия [44], а Khanna et al. разработали Diffusion Visual Explanation, основанный на оценке *score-function* [45]. Однако оба подхода ограничены анализом финального результата и игнорируют временную динамику процесса генерации, что существенно снижает их объяснительную силу.

Вторая линия исследований сосредоточена на адаптации классических XAI-методов к структурам диффузионных моделей. Например, Park et al. предложили DF-CAM — модификацию Grad-CAM, визуализирующую важность признаков на каждом шаге денойзинга [46]. Однако метод столкнулся с фундаментальной проблемой композиции градиентов, где перемножение якобианов вдоль траектории денойзинга приводит к экспоненциальной деградации градиентных сигналов и их численной неустойчивости, делая карты значимости семантически размытыми.

Альтернативой классическим градиентным методам является Manifold Integrated Gradients (MIG) — модификация Integrated Gradients, учитывающая риманову геометрию латентного пространства [47]. Использование геодезических траекторий вместо линейных интерполяций позволяет формировать атрибуции, устойчивые к возмущениям в пространстве признаков, что подтверждается улучшенными показателями по ключевым метрикам объяснимости (*Infidelity*, $Sensitivity_{\max}$, *SSIM*). Несмотря на эти преимущества, MIG остаётся статическим методом и не охватывает временную динамику диффузионных процессов, что ограничивает его применимость в анализе многошаговых генеративных архитектур.

Третье направление исследований связано с использованием *attention-based* подходов. Так, Bahng et al. и Huang et al. предложили применять карты внимания в Latent Diffusion Models для визуализации вклада признаков [**?**, 21]. Однако

подобные методы фиксируют лишь корреляционные зависимости, не обеспечивая строгой каузальной интерпретации, и оказываются чувствительными к архитектурным модификациям модели, что ограничивает их универсальность.

Недавние работы начали учитывать временное измерение. Lee et al. представили Diffusion Explainer, интерактивный инструмент для анализа трансформации промптов в Stable Diffusion [48], а Park et al. предложили визуализацию процесса денойзинга по шагам с вычислением коррелятивных метрик (AUC, *attention maps*) [49]. Тем не менее, оба подхода остаются ограниченными визуализацией и не обеспечивают строгой теоретической атрибуции.

В медицинских приложениях предпринимаются попытки интеграции XAI и генеративных моделей. Siddiqui et al. использовали латентные диффузионные модели в сочетании с LLM для генерации «нормальных» изображений, различия между которыми дают диагностически релевантные атрибуции. Chen разработал метод высокоточной пиксельной атрибуции, основанный на информационном *bottleneck* и MMSE, что позволило получить более чёткие карты важности [27]. Однако данные подходы пока не учитывают временной аспект и не включают каузальную валидацию.

Ключевым ограничением всех перечисленных направлений остаётся отсутствие методов, способных одновременно учитывать временную динамику формирования признаков и обеспечивать устойчивую причинно-следственную интерпретацию в рамках стохастического диффузионного процесса. Хотя концепт *causability* был введён Holzinger et al. [43], практические методы каузальной атрибуции для диффузионных моделей пока не разработаны. Для медицины это критически важно: объяснения должны опираться на доказательства причинности, а не сводиться к корреляционным ассоциациям.

Предложение 1. *Для обеспечения клинической применимости диффузионных моделей необходимо создание XAI-методов, интегрирующих временную атрибуцию и каузальную валидацию, что и составляет основную цель настоящего исследования.*

3. Математические основы диффузионных процессов

Диффузионные модели представляют собой класс генеративных моделей, основанных на обращении процесса постепенного добавления шума к данным [11]. Математической основой служит теория стохастических дифференциальных уравнений и марковских процессов, адаптированная для задач генеративного моделирования [16].

3.1. Прямой диффузионный процесс

Прямой диффузионный процесс определяется стохастическим дифференциальным уравнением Ито [**?**]:

Определение 9 (Прямой диффузионный процесс [11, 16])**.** *Прямой диффузионный процесс для изображения $x_0 \in \mathbb{R}^{H\times W\times C}$ описывается СДУ:*

$$dx_t = -\frac{1}{2}\beta(t)x_t dt + \sqrt{\beta(t)}dW_t,$$

где $\beta(t)$ — коэффициент диффузии, W_t — винеровский процесс.

Коэффициент диффузии $\beta(t)$ контролирует скорость добавления шума: при малых значениях процесс протекает медленно, сохраняя структуру исходного изображения, при больших — быстро разрушает семантическую информацию. Винеровский процесс W_t обеспечивает стохастическую природу диффузии, моделируя случайные флуктуации на молекулярном уровне [15].

Дискретная аппроксимация процесса для временных шагов $t = 1, 2, \ldots, T$, предложенная в [16], задается рекуррентным соотношением:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

где β_t — дискретные коэффициенты шума, удовлетворяющие условию $0 < \beta_1 < \beta_2 < \ldots < \beta_T < 1$. Параметр $\sqrt{1-\beta_t}$ обеспечивает сохранение среднего значения сигнала, в то время как β_t определяет дисперсию добавляемого гауссовского шума. Единичная ковариационная матрица I предполагает независимость шума по пространственным координатам [12].

Лемма 4 (Аналитическое решение прямого процесса [16])**.** *Для прямого процесса с фиксированными коэффициентами $\{\beta_t\}_{t=1}^T$ условное распределение имеет аналитическое решение [16]:*

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I),$$

где $\alpha_t = 1 - \beta_t$ и $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

Доказательство. Обоснование опирается на свойствах произведения гауссовских распределений [**?**]. Применяя репараметризацию и используя замкнутость семейства гауссовских распределений относительно линейных преобразований, получаем требуемый результат через индукцию по t. □

Кумулятивный параметр $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ характеризует долю исходного сигнала, сохраняющуюся после t шагов диффузии. При $t \to \infty$ имеем $\bar{\alpha}_t \to 0$, что означает полную потерю информации об исходном изображении [16].

Репараметризованная форма, критичная для эффективного обучения:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

позволяет сэмплировать из любого временного шага без рекуррентных вычислений, что существенно ускоряет процесс обучения [16].

3.2. Обратный процесс денойзинга

Обратный процесс восстанавливает исходные данные из шума через последовательность условных распределений, параметризуемых нейронной сетью [16]:

Определение 10 (Обратный диффузионный процесс [16])**.** *Обратный диффузионный процесс параметризуется нейронной сетью ϵ_θ и определяется как:*

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

Ключевое наблюдение Ho et al. [16] заключается в том, что для достаточно малых значений β_t обратный процесс также является гауссовским. Это позволяет параметризовать среднее значение через предсказание шума:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right).$$

Нормировочный коэффициент $\frac{1}{\sqrt{\alpha_t}}$ обеспечивает правильное масштабирование очищенного сигнала, а отношение $\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}$ определяет вклад предсказанного шума в итоговое среднее значение [18].

Ковариационная матрица фиксируется согласно [16]:

$$\Sigma_\theta(x_t, t) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t I.$$

Данная параметризация обеспечивает оптимальную дисперсию обратного процесса в смысле минимизации KL-дивергенции с истинным задним распределением [12].

3.3. Вариационная нижняя граница и функция потерь

Теоретической основой обучения DDPM служит вариационная нижняя граница (ELBO) [19]:

Теорема 5 (Вариационная нижняя граница)**.** *Логарифм правдоподобия данных ограничен снизу [16]:*

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[-\sum_{t=1}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) || p_\theta(x_{t-1}|x_t)) \right],$$

где D_{KL} — дивергенция Кульбака-Лейблера.

Каждый член суммы $D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$ измеряет расхождение между истинным задним распределением $q(x_{t-1}|x_t, x_0)$, вычислимым аналитически для гауссовских процессов, и приближенным распределением $p_\theta(x_{t-1}|x_t)$, параметризуемым нейронной сетью [11].

Упрощенная функция потерь, предложенная в [16], игнорирует весовые коэффициенты из ELBO:

$$\mathcal{L}_{simple} = \mathbb{E}_{x_0, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 \right],$$

где временной шаг t равномерно сэмплируется из $\{1, 2, \ldots, T\}$, $\epsilon \sim \mathcal{N}(0, I)$. Данная формулировка интерпретируется как задача денойзинга: модель обучается предсказывать шум ϵ, добавленный к исходному изображению на шаге t [16].

3.4. UNet архитектура с временным внедрением

Нейронная сеть ϵ_θ реализуется через архитектуру UNet [20], адаптированную для условной генерации:

$$\epsilon_\theta(x_t, t) = \text{UNet}(x_t, \gamma(t)).$$

Функция временного внедрения $\gamma(t)$, заимствованная из архитектур трансформеров [21], использует синусоидальные кодировки:

$$\gamma(t) = [\sin(t/10000^{2k/d}), \cos(t/10000^{2k/d})]_{k=0}^{d/2-1},$$

где d — размерность вложения. Различные частоты позволяют модели различать временные шаги на разных масштабах: низкие частоты кодируют грубые временные различия, высокие — тонкие [6].

Механизм внимания в UNet, адаптированный из [21]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

где $Q = x_t W_Q + \gamma(t)$, $K = x_t W_K$, $V = x_t W_V$. Временная информация $\gamma(t)$ добавляется к запросам Q, позволяя модели адаптировать паттерны внимания в зависимости от уровня шума [22].

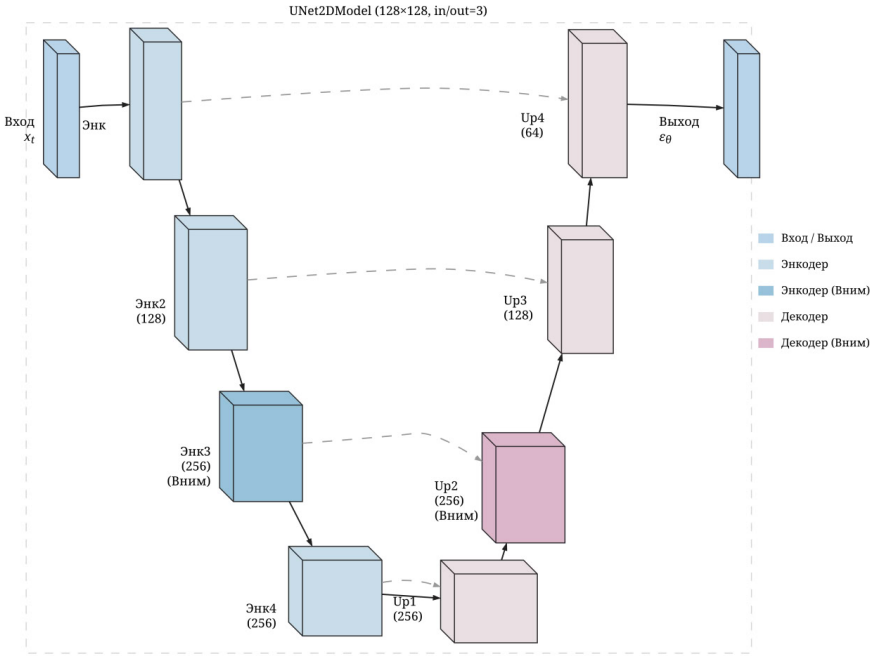

Рис. 1: *Схема encoder-decoder структуры с механизмами внимания*

3.5. Планировщик денойзинга

Выбор коэффициентов шума $\{\beta_t\}$ критически влияет на качество генерации. Линейный планировщик [16]:

$$\beta_t = \beta_1 + \frac{t-1}{T-1}(\beta_T - \beta_1)$$

обеспечивает равномерное увеличение шума, но может приводить к избыточно быстрой потере информации на поздних этапах.

Косинусный планировщик [12] предоставляет более мягкое распределение шума:

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos^2\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right),$$

где s — малый сдвиг для предотвращения деления на ноль. Косинусная функция обеспечивает медленное начальное добавление шума с ускорением к концу процесса, что лучше согласуется с иерархической структурой изображений [15].

3.6. Алгоритм генерации и его математические свойства

Теорема 6 (Алгоритм генерации DDPM)**.** *Для генерации изображения из $x_T \sim \mathcal{N}(0, I)$ выполняется итеративный процесс [16]:*

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, \tag{1}$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \quad z \sim \mathcal{N}(0, I) \tag{2}$$

для $t = T, T-1, \ldots 1,$.

Первое слагаемое представляет детерминистическое предсказание модели, второе — стохастическую составляющую, необходимую для сохранения правильной дисперсии процесса [16].

Марковское свойство процесса имеет фундаментальное значение для XAI анализа:

Предложение 2 (Марковское свойство)**.** *Обратный процесс удовлетворяет марковскому свойству [16]:*

$$p_\theta(x_{t-1}|x_t, x_{t+1}, \ldots, x_T) = p_\theta(x_{t-1}|x_t).$$

Это свойство обеспечивает возможность локального анализа важности каждого временного шага независимо от будущих состояний, что составляет основу метода Time-SHAP, разрабатываемого в следующем разделе.

Градиентная структура процесса позволяет вычислять производные важности:

$$\frac{\partial \log p_\theta(x_0)}{\partial x_t} = \mathbb{E}_{q(x_{1:T}|x_0)}\left[\frac{\partial}{\partial x_t}\sum_{s=1}^{T}\log p_\theta(x_{s-1}|x_s)\right].$$

Эта формула, выводимая из вариационной нижней границы [18], позволяет анализировать чувствительность финального результата к промежуточным состояниям диффузионного процесса.

4. TimeSHAP-Diff: временная атрибуция для диффузионных моделей

Классические методы XAI анализируют статические модели, игнорируя временную динамику генеративных процессов. Для диффузионных моделей критически важно понимание эволюции важности признаков на каждом временном шаге денойзинга. В настоящем разделе представлен метод TimeSHAP-Diff — адаптация временного анализа важности TimeSHAP [54] для стохастических диффузионных процессов.

Наш подход адаптирует фреймворк TimeSHAP к диффузионным моделям путем переформулировки временной функции ценности для учета специфики денойзинга и введения каузальных интервенций для валидации атрибуций. В отличие от оригинального TimeSHAP, ориентированного на рекуррентные модели и маскирование событий, TimeSHAP-Diff оперирует с непрерывными стохастическими переходами и требует специализированной обработки градиентной нестабильности в композиционных архитектурах.

4.1. Временная функция ценности для диффузионных процессов

Основой метода TimeSHAP-Diff служит концепция временной функции ценности, обобщающая классическую коалиционную функцию теории игр [13] на диффузионные процессы.

Определение 11 (Временная функция ценности для диффузионных процессов)**.** *Временная функция ценности $v^t : 2^T \to \mathbb{R}$ для диффузионного процесса определяется как:*

$$v^t(S) = F_c\left(DDPM_T^{(S)}(x_0)\right) - F_c\left(DDPM_T^{(\emptyset)}(x_0)\right),$$

где $S \subseteq T = \{1, 2, \ldots, T\}$ — коалиция временных шагов, F_c — логит целевого класса с классификатора, $DDPM_T^{(S)}$ — модифицированный диффузионный процесс с активными шагами из S.

Модифицированный процесс $\text{DDPM}_T^{(S)}$ реализуется через селективное применение денойзинга:

$$x_{t-1}^{(S)} = \begin{cases} \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, & t \in S \\ x_t + \mathcal{N}(0, \sigma_{noise}^2 I), & t \notin S \end{cases}.$$

Для шагов, не включенных в коалицию S, вместо денойзинга применяется добавление случайного шума, что позволяет изолировать вклад конкретных временных интервалов в итоговую классификацию [4].

4.2. Аксиоматическая основа для диффузионных процессов

Для адаптации классических аксиом Шепли к временным процессам диффузионного денойзинга необходимо переформулировать основные принципы справедливого распределения важности между временными шагами.

Определение 12 (Аксиомы временной атрибуции для диффузионных процессов)**.** *Пусть* $\Phi : \mathbb{R}^{d\times T} \to \mathbb{R}^T$ *— функция временной атрибуции важности для диффузионного процесса. Тогда TimeSHAP-Diff значения* $\{\phi^t\}_{t=1}^T$ *должны удовлетворять следующим аксиомам:*

Аксиома временной эффективности: $\sum_{t=1}^T \phi^t = v^t(T) - v^t(\emptyset)$.

Аксиома временной симметрии:

Если $v^t(S\cup\{i\}) - v^t(S) = v^t(S\cup\{j\}) - v^t(S)$ *для всех* $S \subseteq T \setminus \{i, j\}$*, то* $\phi^i = \phi^j$.

Аксиома временного нулевого игрока:

Если $v^t(S \cup \{i\}) = v^t(S)$ *для всех* $S \subseteq T \setminus \{i\}$*, то* $\phi^i = 0$.

Аксиома временной аддитивности:

Для двух игр v_1, v_2*:* $\phi^t[v_1 + v_2] = \phi^t[v_1] + \phi^t[v_2]$.

Эти аксиомы обеспечивают теоретическую основу для справедливого распределения вклада временных шагов в итоговое решение диффузионной модели.

4.3. Формулировка TimeSHAP-Diff

Расширяя классическое определение значений Шепли [13] на временную размерность, получаем:

Определение 13 (TimeSHAP-Diff значения)**.** *TimeSHAP-Diff значение для временного шага* t *определяется как:*

$$\phi_h^t = \sum_{S\subseteq T\setminus\{t\}} \frac{|S|!(|T| - |S| - 1)!}{|T|!} \left[v^t(S \cup \{t\}) - v^t(S)\right],$$

где суммирование ведется по всем возможным коалициям S*, не содержащим шаг* t.

Коэффициент $\frac{|S|!(|T|-|S|-1)!}{|T|!}$ представляет вероятность того, что коалиция S формируется до добавления шага t в случайном порядке вступления участников в игру. Данная весовая схема обеспечивает справедливое распределение вклада между временными шагами [23].

4.4. Математическая консистентность

Теорема 7 (Консистентность TimeSHAP-Diff)**.** *Пусть* $\Phi : \mathbb{R}^{d\times T} \to \mathbb{R}^T$ *— монотонная функция важности для диффузионного процесса. Тогда TimeSHAP-Diff значения* $\{\phi^t\}_{t=1}^T$ *единственным образом удовлетворяют аксиомам из Определения 12.*

Доказательство. Доказательство следует схеме классической теоремы Шепли [13].

Единственность. Предположим существование двух различных функций ϕ, ψ, удовлетворяющих аксиомам из Определения 12. Рассмотрим разность $\delta^t = \phi^t - \psi^t$. Из аксиомы эффективности: $\sum_{t=1}^T \delta^t = 0$.

Для каждого временного шага t определим единичную игру $u_i^t(S) = \mathbb{I}[i \in S]$. Из аксиомы нулевого игрока для всех $j \neq i$: $\delta^j[u_i^t] = 0$. Из аксиомы эффективности: $\delta^i[u_i^t] = 0$.

Поскольку любая временная функция ценности представима как линейная комбинация единичных игр, из аксиомы аддитивности следует $\delta^t[v] = 0$ для любой игры v.

Существование. Конструктивно определим ϕ^t через формулу из Определения 13 и проверим выполнение аксиом:

Аксиома эффективности проверяется прямым вычислением:

$$\sum_{t=1}^{T} \phi^t = \sum_{t=1}^{T} \sum_{S \subseteq T \setminus \{t\}} \frac{|S|!(|T| - |S| - 1)!}{|T|!} [v(S \cup \{t\}) - v(S)].$$

Изменяя порядок суммирования и используя комбинаторные тождества, получаем требуемый результат.

Остальные аксиомы проверяются аналогично через прямую подстановку и использование свойств биномиальных коэффициентов. □

4.5. Композиционная адаптация Grad-CAM для диффузионных моделей

Классический Grad-CAM [5] требует адаптации для анализа диффузионных моделей из-за композиционной структуры процесса денойзинга.

Определение 14 (Grad-CAM для диффузионных моделей)**.** *Grad-CAM для диффузионной модели определяется через композицию градиентов:*

$$L_{Grad-CAM}^{DDPM,c} = ReLU\left(\sum_k w_k^{c,t} \frac{\partial F}{\partial DDPM_t} \cdot \frac{\partial DDPM_t}{\partial A^k}\right),$$

где веса важности вычисляются как:

$$w_k^{c,t} = \frac{1}{Z} \sum_i \sum_j \frac{\partial F_c}{\partial A_{ij}^k} \cdot \left|\frac{\partial DDPM_t}{\partial x_t}\right|,$$

где применяется стабилизация через модуль якобиана вместо произведения.

Произведение якобианов $\prod_{t=1}^{T} \frac{\partial x_{t-1}}{\partial x_t}$ учитывает цепочку денойзинга, но может приводить к численной нестабильности [8]. Для стабилизации применяется градиентное отсечение:

$$\frac{\partial x_{s-1}}{\partial x_s} \leftarrow \text{clip}\left(\frac{\partial x_{s-1}}{\partial x_s}, -\gamma, \gamma\right),$$

где γ — пороговое значение, подбираемое эмпирически.

4.6. Алгоритмическая реализация и вычислительная сложность

Практическая реализация TimeSHAP-Diff требует эффективного вычисления значений Шепли для экспоненциального числа коалиций. Применяется аппроксимационный алгоритм на основе сэмплирования:

Теорема 8 (Алгоритм аппроксимации TimeSHAP-Diff)**.** *Для заданной точности ϵ и доверительной вероятности δ достаточно M случайных выборок коалиций, где:*

$$M \geq \frac{2\log(2T/\delta)}{\epsilon^2} \cdot \max_{t,S} |v^t(S \cup \{t\}) - v^t(S)|^2.$$

Вычислительная сложность полного алгоритма составляет $O(M \cdot T \cdot C_{forward})$, где $C_{forward}$ — стоимость одного прохода через диффузионную модель. Для практических применений с $T = 50$ и $M = 1000$ алгоритм выполним на современном GPU за несколько минут.

Предлагаемый метод обеспечивает теоретически обоснованную временную атрибуцию для диффузионных моделей, что критически важно для понимания динамики формирования диагностически значимых признаков в медицинских изображениях. Экспериментальная валидация метода представлена в следующих разделах.

5. Системная архитектура и вычислительная реализация объяснимых диффузионных моделей

На основе теоретических основ разделов 2-3 разработана комплексная система объяснимого анализа диффузионных моделей для медицинской диагностики, реализованная в проектах SYNT_ISIC [50] и CAS_ISIC [51]. Система обеспечивает полный цикл от обучения диффузионных моделей до генерации синтетических изображений с интегрированным XAI анализом.

5.1. Архитектура пайплайна обработки данных

Разработанный пайплайн обучения включает последовательность этапов предобработки, оптимизированных для медицинских изображений дерматологических поражений. Архитектура системы представлена на рисунке 2.

Загрузка и индексация данных. Система использует специализированный SingleClassDataset для обработки датасета ISIC2018 [10], обеспечивающий эффективную загрузку изображений размером 128×128 пикселей с соответствующими CSV метаданными:

$$\mathcal{D} = \{(x_i, y_i, m_i)\}_{i=1}^{N},$$

где $x_i \in \mathbb{R}^{128\times128\times3}$ — RGB изображение, $y_i \in \{0, 1, \ldots, 6\}$ — метка класса, m_i — метаданные пациента.

Коррекция цветовых характеристик. Для компенсации вариативности условий съемки применяется адаптивная цветовая коррекция:

$$x_{corrected} = \alpha \cdot \frac{(x - \mu_{class})}{\sigma_{class}} \cdot \sigma_{target} + \mu_{target} + \beta \cdot x,$$

где $\mu_{class}, \sigma_{class}$ — статистики класса, $\mu_{target}, \sigma_{target}$ — целевые параметры, α, β — весовые коэффициенты смешивания.

Аугментация данных. Система применяет медицински обоснованные трансформации [28]:

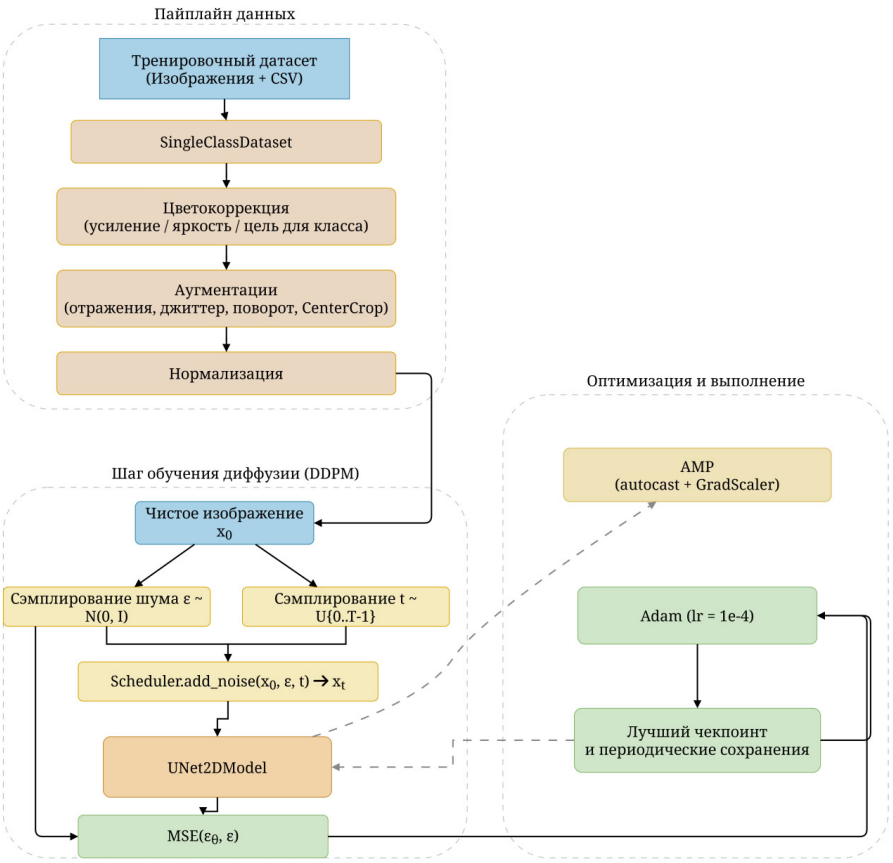

Рис. 2: *Пайплайн обработки данных и обучения диффузионной модели*

Горизонтальные и вертикальные отражения для моделирования различных ориентаций поражений:

$$T_{flip}(x) = \begin{cases} x & \text{с вероятностью } 0.5, \\ \text{FlipHorizontal}(x) & \text{с вероятностью } 0.25, \\ \text{FlipVertical}(x) & \text{с вероятностью } 0.25. \end{cases}$$

Цветовые искажения (Color Jitter) для повышения робустности к вариациям освещения:

$$T_{jitter}(x) = \text{ColorJitter}(x; \text{brightness} = 0.2, \text{contrast} = 0.2, \text{saturation} = 0.2, \text{hue} = 0.1).$$

Случайные повороты в пределах медицински допустимых углов:

$$T_{rotation}(x) = \text{Rotate}(x; \theta \sim \mathcal{U}(-15, 15)).$$

Центральное кадрирование с сохранением диагностически значимых областей:

$$T_{crop}(x) = \text{CenterCrop}(x; \text{size} = (112, 112)).$$

5.2. Нормализация и подготовка к обучению

Итоговая нормализация использует статистики ImageNet для обеспечения совместимости с предобученными компонентами [29]:

$$x_{normalized} = \frac{x - \boldsymbol{\mu}_{ImageNet}}{\boldsymbol{\sigma}_{ImageNet}},$$

где $\boldsymbol{\mu}_{ImageNet} = [0.485, 0.456, 0.406]^T$, $\boldsymbol{\sigma}_{ImageNet} = [0.229, 0.224, 0.225]^T$.

5.3. Диффузионный этап обучения

Обучение диффузионной модели реализует упрощенную функцию потерь DDPM [16]:

$$\mathcal{L}_{DDPM} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right].$$

Процедура сэмплирования шума. На каждой итерации обучения: - Чистое изображение x_0 сэмплируется из датасета - Временной шаг $t \sim \mathcal{U}\{0, T-1\}$ выбирается равномерно - Гауссовский шум $\epsilon \sim \mathcal{N}(0, I)$ генерируется независимо

Зашумленное изображение формируется согласно планировщику:

$$x_t = \text{Scheduler.add\_noise}(x_0, \epsilon, t) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon.$$

UNet2DModel архитектура. Нейронная сеть ϵ_θ использует модифицированную UNet архитектуру с механизмами внимания на промежуточных слоях:

$$\epsilon_\theta(x_t, t) = \text{UNet2D}(x_t, \gamma(t)),$$

где временное внедрение $\gamma(t)$ реализовано через синусоидальные кодировки [21].

5.4. Оптимизация и стратегия обучения

Современные методы обучения диффузионных моделей требуют использования эффективных алгоритмов оптимизации и управления ресурсами для обеспечения стабильности и скорости обучения.

Для оптимизации параметров нейронной сети используется алгоритм **Adam** [31] с параметрами скорости обучения $\text{lr} = 1 \times 10^{-4}$, а также коэффициентами моментов $\beta_1 = 0.9, \beta_2 = 0.999$. Обновление параметров происходит согласно формулировке:

$$\theta_{t+1} = \theta_t - \frac{\text{lr} \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon},$$

где $\hat{m}_t$ и $\hat{v}_t$ — скорректированные оценки первого и второго моментов градиентов на шаге t, ϵ — малая константа для численной устойчивости.

Для повышения производительности и снижения потребления памяти применяется **автоматическое смешанное точностное обучение (AMP)** с использованием механизмов autocast и GradScaler [30]. Данная технология позволяет динамически переключаться между 16- и 32-битной точностью при вычислениях, сохраняя при этом качество модели.

Кроме того, реализована система **checkpointing**, обеспечивающая регулярное сохранение состояний модели и оптимизатора. Это позволяет возобновлять обучение после сбоев, а также отбирать лучшие модели по метрикам качества.

Такой комплекс методов оптимизации и управления ресурсами обеспечивает стабильное обучение диффузионных моделей даже на оборудовании с ограниченной вычислительной мощностью.

5.5. Встроенный XAI анализ и интеграция с системой генерации

Система интегрирует расчёт XAI атрибуций непосредственно в процесс генерации и валидации модели. Основной компонент — модуль, реализующий **TimeSHAP-Diff**, который через выборку коалиций временных шагов оценивает важность каждого этапа денойзинга.

Практическая реализация TimeSHAP-Diff включает эффективный алгоритм сэмплирования коалиций временных шагов...

5.6. Техническая реализация и управление ресурсами

Архитектура системы спроектирована с учетом ограничений вычислительных ресурсов и требований к масштабируемости. Ключевые технические решения включают:

Управление памятью GPU. После обработки каждого изображения выполняется принудительная очистка кэша видеопамяти через вызов torch.cuda.empty_cache(), что предотвращает накопление неиспользуемых тензоров и обеспечивает стабильную работу на устройствах с ограниченным объемом памяти.

Batch-обработка и оптимизация I/O. Система использует адаптивные размеры батчей в зависимости от доступной памяти и асинхронную загрузку данных для минимизации времени простоя GPU.

Мониторинг метрик. Интегрированная система логирования отслеживает динамику функции потерь, использование памяти, времена выполнения отдельных этапов и автоматически сохраняет промежуточные результаты для последующего анализа.

Воспроизводимость экспериментов. Все компоненты системы используют детерминистические генераторы псевдослучайных чисел с фиксированными seed значениями, что обеспечивает полную воспроизводимость результатов при идентичных входных параметрах.

Представленная техническая реализация демонстрирует практическую применимость теоретических методов, разработанных в предыдущих разделах, и служит основой для проведения экспериментальной валидации на реальных медицинских данных.

6. Методология каузальной валидации: дизайн интервенций, метрики и экспериментальный протокол

Фундаментальной проблемой применения объяснимых методов в медицинской диагностике является установление строгих причинно-следственных связей между выделяемыми XAI регионами и диагностическими решениями. В настоящем

разделе представлена комплексная методология каузальной валидации метода TimeSHAP-Diff, структурированная как последовательность логически связанных этапов верификации от глобальной количественной оценки различий до финального сравнения с baseline методами.

6.1. Контрафактуальные интервенции и каузальные метрики

Теоретическую основу каузальной валидации составляют do-операции Pearl [24], адаптированные для архитектуры диффузионных моделей. В качестве формального каркаса вводится следующее определение:

Определение 15 (Do-операция для каузальной валидации [24])**.** *Do-операция для региона R определяется как:*

$$do(R := v) : x_t \mapsto x_t \odot (1 - M_R) + v \odot M_R,$$

где M_R — бинарная маска региона, v — интервенционное значение.

Для систематической проверки каузальной значимости выделенных регионов разработан набор из четырех типов контрафактуальных интервенций, каждая из которых моделирует различные аспекты деградации диагностической информации:

Шумовая интервенция: $I_{\text{noise}}(x, M) = x \odot (1 - M) + \mathcal{N}(0, \sigma^2) \odot M$, где $\sigma^2 = 1.5 \cdot \text{Var}(x \odot M)$.

Интервенция размытия: $I_{\text{blur}}(x, M) = x \odot (1 - M) + (G_{\sigma=3.0} * x) \odot M$.

Перестановочная интервенция:

$I_{\text{shuffle}}(x, M) = x \odot (1 - M) + \text{Shuffle}(x \odot M) \odot M$.

Константная интервенция: $I_{\text{const}}(x, M) = x \odot (1 - M) + \text{median}(x) \odot M$.

Далее, для количественной оценки каузального воздействия вводится метрика каузального сдвига (CSI):

$$\text{CSI}^I(R) = \left| F(x_0) - F\left(\text{DDPM}_{T\to 0}(I(x_T, M_R))\right) \right|,$$

где F — функция классификатора, I — тип интервенции.

6.2. Методология экспериментальной валидации: логическая цепочка проверки

В качестве экспериментальной платформы выбран датасет ISIC2018, содержащий 10,015 дерматоскопических изображений семи классов поражений кожи [10]. Архитектурную основу составляют UNet2DModel (рис. 3) и ResNet18 классификатор для валидации XAI результатов.

Проектируемая методология реализует логическую цепочку валидации, структурированную в соответствии с принципами доказательной медицины:

Шаг 1. Начинаем с «глобальной количественной оценки различий» (Cohen's d). В качестве первого этапа применяется статистическая валидация каузального сдвига. Для оценки статистической значимости различий между регионами высокой и низкой важности применяется тест Уэлча для выборок важности

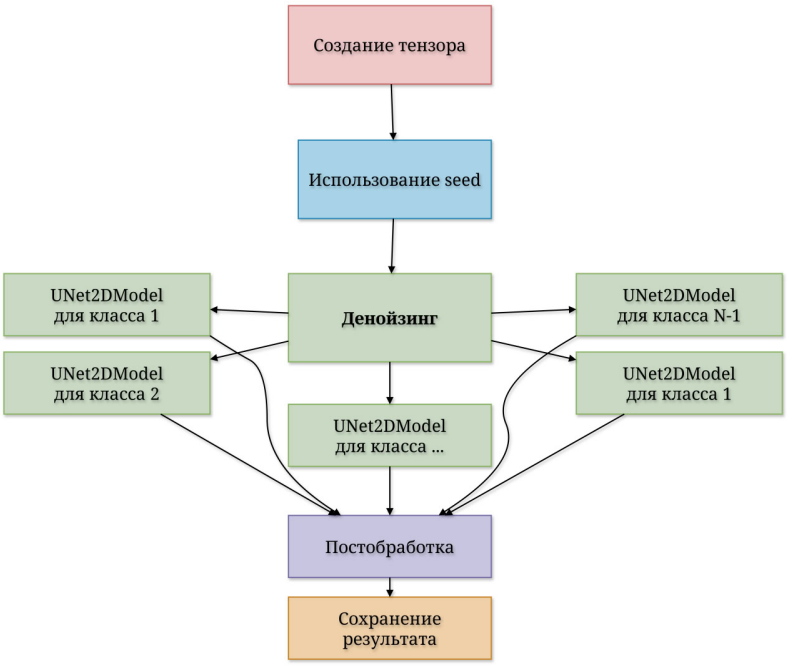

Рис. 3: *Архитектура UNet2DModel с механизмами внимания для диффузионного денойзинга*

$\{s_{top-k}\}$ и $\{s_{bottom-k}\}$ с потенциально неравными дисперсиями:

$$t = \frac{\bar{s}_{top} - \bar{s}_{bottom}}{\sqrt{\frac{s^2_{top}}{n_{top}} + \frac{s^2_{bottom}}{n_{bottom}}}},$$

где s^2_{top}, s^2_{bottom} — выборочные дисперсии соответствующих групп, n_{top}, n_{bottom} — размеры выборок. Число степеней свободы вычисляется по формуле Уэлча-Саттертуэйта:

$$\nu = \frac{\left(\frac{s^2_{top}}{n_{top}} + \frac{s^2_{bottom}}{n_{bottom}}\right)^2}{\frac{s^4_{top}}{n^2_{top}(n_{top}-1)} + \frac{s^4_{bottom}}{n^2_{bottom}(n_{bottom}-1)}}.$$

Для количественной оценки практической значимости различий дополнительно вычисляется Cohen's d:

$$d = \frac{\bar{s}_{top} - \bar{s}_{bottom}}{s_p}.$$

Шаг 2. Затем переходим к «оценке согласованности методов» (корреляции). В рамках проверки согласованности между различными XAI методами планируется корреляционный анализ через коэффициенты корреляции Пирсона для оценки дополнительности подходов к интерпретации.

Шаг 3. Потом «рассматриваем уникальный временной аспект» (Time-SHAP). В качестве дополнительного аспекта исследования предполагается анализ эволюции важности для выявления временных этапов формирования диагностически значимых признаков в процессе денойзинга.

Шаг 4. После этого «сверяемся с экспертным клиническим знанием» (анатомия). Данный этап предполагает интеграцию полученных XAI атрибуций с медицинскими знаниями о морфологических признаках дерматологических поражений для проверки клинической адекватности выделенных регионов.

6.3. Каузальная валидность регионов важности

Шаг 5. Далее «формально проверяем каузальность» (CSI). Ключевым этапом валидации является применение четырех типов интервенций к регионам с наивысшими и наинизшими TimeSHAP-Diff значениями. Итоговая каузальная разность эффектов определяется как:

$$\Delta_{CSI} = \frac{1}{4} \sum_{I \in \{\text{noise,blur,shuffle,const}\}} \left[\text{CSI}^I(R_{top-10\%}) - \text{CSI}^I(R_{bottom-10\%}) \right].$$

Лемма 5 (Каузальная валидность TimeSHAP-Diff регионов)**.** *При статистически значимой разности* Δ_{CSI} *и* $p < 0.05$ *регионы с высокими TimeSHAP-Diff значениями каузально влияют на решения классификатора.*

Шаг 6. Наконец, «проверяем устойчивость и сравниваем с baseline». Для обеспечения воспроизводимости результатов планируется оценка стабильности по следующим ключевым аспектам:

– Устойчивость к изменению количества шагов интеграции,

– Стабильность SHAP при различном количестве samples,

– Воспроизводимость результатов при фиксированных seeds.

В качестве финального этапа валидации предполагается сравнительный анализ TimeSHAP-Diff с классическими XAI методами по метрике каузальной валидности:

– TimeSHAP-Diff,

– Integrated Gradients,

– классический SHAP,

– Grad-CAM.

Таким образом, представленная методология обеспечивает теоретически обоснованную основу для строгого доказательства каузальных связей между анатомическими регионами и диагностическими решениями, что является критически важным требованием для клинического применения объяснимых диффузионных моделей в дерматологической диагностике. Предложенная логическая цепочка валидации — от глобальной статистической оценки через временную атрибуцию к

каузальной верификации — обеспечивает комплексную проверку надежности XAI результатов на всех уровнях анализа.

7. Экспериментальная валидация и анализ результатов

Практическая валидация разработанного метода TimeSHAP-Diff выполнена на датасете ISIC2018 Task 3 [10], содержащем 10,015 дерматоскопических изображений семи классов дерматологических поражений. Полная реализация системы с воспроизводимыми результатами доступна в открытых репозиториях SYNT_ISIC [50, 52] и CAS_ISIC [51, 53].

7.1. Квантификация дисбаланса классов и стратегии диффузионной компенсации

Исходный датасет ISIC2018 характеризуется критическим дисбалансом классов, типичным для клинических баз данных реального мира. Количественные характеристики распределения представлены в таблице 1.

Таблица 1: *Распределение классов в исходном датасете ISIC2018*

Класс	**Количество**	**Доля, %**
NV (пигментный невус)	6750	67.40
MEL (меланома)	1113	11.11
BKL (доброкачественные кератозоподобные поражения)	1099	10.97
BCC (базальноклеточная карцинома)	514	5.13
AKIEC (актинический кератоз)	327	3.27
VASC (сосудистые поражения)	142	1.42
DF (дерматофиброма)	115	1.15

Количественные метрики дисбаланса исходного датасета: коэффициент дисбаланса $\rho = \frac{\max_i |C_i|}{\min_i |C_i|} = 58.30$, индекс Джини неравенства классов $G = 0.6414$, стандартное отклонение размеров классов $\sigma = 2,186.96$.

7.1.1. Создание сбалансированных синтетических конфигураций

Применение диффузионной генерации позволило создать три экспериментальных конфигурации с дифференцированными стратегиями балансировки данных. Синтетический датасет формируется через селективное дополнение редких классов до равномерного распределения, что представлено в таблице 2.

Таблица 2: *Распределение классов в синтетически дополненном датасете*

Класс	**Исходный**	**Синтетический**	**Итого**
NV	6,750	0	6,750
MEL	1,113	3,387	4,500
BKL	1,099	3,401	4,500
BCC	514	3,986	4,500
AKIEC	327	4,173	4,500
VASC	142	4,358	4,500
DF	115	4,385	4,500
Общий размер	**10,015**	**25,790**	**35,805**

Сравнительные характеристики экспериментальных конфигураций демонстрируют качественные различия в подходах к балансировке данных. Таблица 3 представляет количественные метрики для каждой стратегии.

Таблица 3: *Сравнительные характеристики экспериментальных датасетов*

Датасет	**Размер**	**Индекс Джини**	**Коэф. дисб.**	**Стд. откл.**	**Редкий класс**
ISIC2018	10,015	0.6414	58.30	2,186.96	DF (115)
ISIC2018_synt	35,805	0.0592	1.49	678.70	MEL (4,500)
ISIC2018_filtered	23,064	0.2835	5.00	1,733.03	VASC (1,341)

Количественная оценка эффективности балансировки представлена в таблице 4. Синтетический датасет (ISIC2018_synt) достигает практически идеального баланса классов с коэффициентом дисбаланса, сниженным до 1.49, что соответствует улучшению на 97.44%. Фильтрованный датасет (ISIC2018_filtered) представляет гибридное решение, сохраняющее все исходные данные и дополняющее их синтетическими образцами редких классов.

Таблица 4: *Количественная оценка улучшения баланса классов*

Переход между датасетами	**Улучшение Джини (%)**	**Улучшение дисбаланса (%)**
ISIC2018 → ISIC2018_synt	90.78	97.44
ISIC2018 → ISIC2018_filtered	55.80	91.42

7.2. Сравнительный анализ архитектурных конфигураций

Количественная оценка влияния синтетического дополнения на производительность классификации проведена с использованием архитектуры ResNet18. Результаты, представленные в таблице 5, демонстрируют статистически значимое улучшение всех ключевых метрик качества.

Таблица 5: *Сравнительные метрики качества классификации*

Датасет	**Точность**	**Сбалансированная точность**	**F1-метрика**	**Cohen's κ**
ISIC2018	0.932	0.951	0.946	0.884
ISIC2018_synt	0.971	0.970	0.972	0.951
ISIC2018_filtered	0.948	0.950	0.949	0.936

Анализ критического класса дерматофибромы. Исходная модель ResNet18, обученная на несбалансированном датасете, демонстрировала структурную неспособность к надежной классификации дерматофибромы как наиболее редкого класса ($n = 115$, 1.15% выборки). Несмотря на формально высокие метрики классификации (Recall = 1.000), последующий анализ TimeSHAP-Diff выявил критически низкую финальную уверенность модели (final conf = 0.011) и полное отсутствие визуальной репрезентативности (visual conf = 0.000). Данная ситуация свидетельствует о том, что правильные предсказания достигались за счет статистических артефактов обучающего процесса, а не через распознавание диагностически релевантных морфологических признаков.

После синтетического дополнения до 4,600 образцов модель продемонстрировала качественный скачок в способности к генерализации с достижением метрик precision, recall и F1-score равных 1.000. Это фундаментальное улучшение подтверждает эффективность диффузионной генерации в решении проблем экстремального дисбаланса классов в медицинских приложениях.

7.3. Временная эволюция диффузионного процесса и TimeSHAP-Diff атрибуция

Центральным результатом исследования является экспериментальная демонстрация временной динамики формирования диагностически значимых признаков в процессе денойзинга. Рисунок 4 иллюстрирует ключевые этапы диффузионного процесса для класса меланомы с соответствующими TimeSHAP-Diff значениями временной важности.

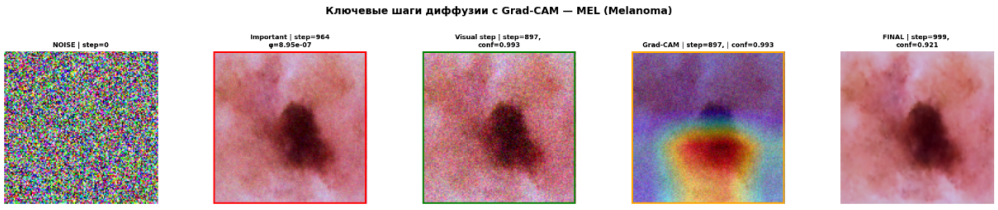

Рис. 4: *Временная эволюция диффузионного процесса для класса меланомы: исходное шумовое изображение (step=0), этап максимальной TimeSHAP-Diff важности (step=964, $\varphi = 8.95 \times 10^{-7}$), визуально репрезентативный фрагмент (step=897, conf=0.993), тепловая карта Grad-CAM активации и финальное детерминированное изображение (step=999, conf=0.921)*

Параметрическое описание временных этапов основывается на следующих ключевых индикаторах. Параметр step обозначает индекс временной траектории диффузионного процесса, где step= 0 соответствует изотропному гауссовскому шуму, step= 999 представляет финальное детерминированное изображение. Параметр φ (TimeSHAP-Diff value) определяет маргинальный вклад конкретного временного шага в изменение уверенности модели относительно базового состояния.

Анализ эволюции уверенности классификатора представлен на рисунке 5, демонстрирующем дифференцированные паттерны временного развития для всех семи классов дерматологических поражений.

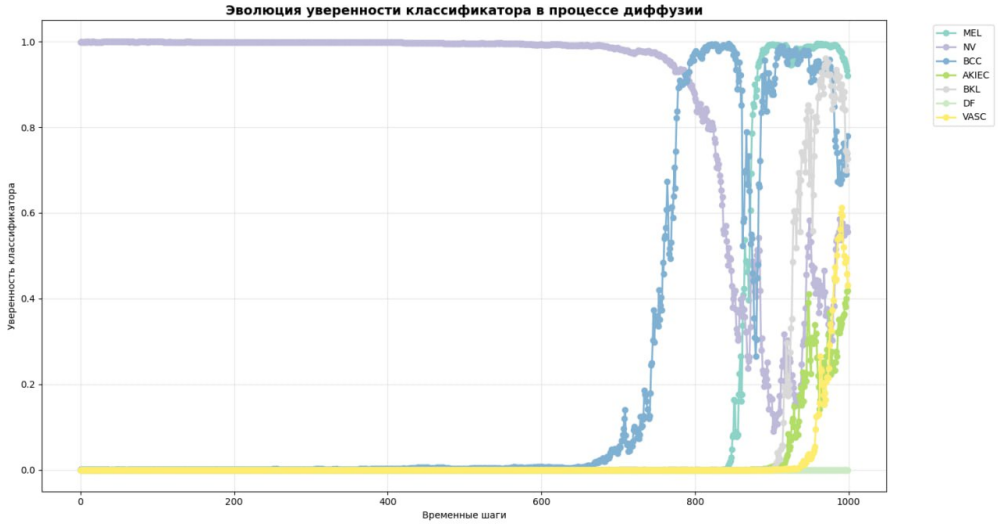

Рис. 5: *Динамика уверенности классификатора ResNet18 в процессе диффузионного денойзинга для семи классов дерматологических поражений*

Анализ временных профилей выявляет характерные закономерности формирования диагностических признаков. Класс MEL (меланома) демонстрирует достижение острого максимума уверенности на промежуточном этапе ($t \approx 900$), что указывает на раннее формирование ключевых патологических признаков в процессе денойзинга. Класс NV (меланоцитарный невус) характеризуется монотонным ростом уверенности, отражающим постепенное накопление согласованных визуальных признаков. Классы BCC и BKL демонстрируют резкое возрастание уверенности на поздних этапах ($t > 850$), что свидетельствует о консолидации диагностически значимых признаков на завершающих стадиях генеративного процесса. Редкие классы VASC и DF показывают запаздывающую консолидацию ($t > 950$), отражающую повышенную сложность распознавания вследствие ограниченной представленности в обучающей выборке.

7.4. Сравнительный анализ TimeSHAP-Diff результатов по экспериментальным конфигурациям

Проведен комплексный анализ производительности метода TimeSHAP-Diff на

трех версиях датасета для количественной оценки влияния синтетической генерации на качество временной атрибуции. Результаты анализа исходного датасета представлены в таблице 6.

Таблица 6: *Результаты TimeSHAP-Diff анализа на исходном датасете ISIC2018*

Класс	**Важность**	**Визуальный признак**	**φ-важность**	**Визуальная уверенность**	**Итоговая уверенность**	**Совпадение**
NV	966	897	8.00×10^{-7}	0.002	0.801	✓
MEL	49	325	1.80×10^{-6}	1.000	0.998	✓
BKL	807	807	1.77×10^{-12}	1.000	0.891	✓
BCC	99	895	1.77×10^{-6}	0.035	0.982	✓
AKIEC	976	898	3.45×10^{-6}	0.042	0.740	✓
VASC	923	895	-3.00×10^{-7}	0.000	0.000	×
DF	983	899	1.62×10^{-7}	0.000	0.011	×

Анализ синтетически дополненного датасета демонстрирует неоднозначные результаты, представленные в таблице 7. Несмотря на кардинальное увеличение объема данных для редких классов, система показывает артефакты синтетической генерации.

Таблица 7: *Результаты TimeSHAP-Diff анализа на синтетически дополненном датасете*

Класс	**Важность**	**Визуальный признак**	**φ-важность**	**Визуальная уверенность**	**Итоговая уверенность**	**Совпадение**
NV	964	897	8.95×10^{-7}	0.993	0.921	✓
MEL	30	338	9.00×10^{-7}	0.999	0.556	✓
BCC	912	844	2.12×10^{-12}	0.996	0.780	✓
AKIEC	99	899	6.38×10^{-7}	0.007	0.419	✓
BKL	970	895	9.68×10^{-10}	0.001	0.727	✓
DF	983	899	3.00×10^{-14}	0.000	0.000	×
VASC	991	899	1.84×10^{-9}	0.000	0.430	✓

Критическим наблюдением является персистирующая проблема класса DF. Несмотря на увеличение объема данных до $4,600$ изображений и улучшение общих метрик модели, анализ TimeSHAP-Diff фиксирует системный отказ в формировании осмысленного предсказания. Значения визуальной и итоговой уверенности остаются нулевыми, что свидетельствует о глубоких артефактах случайной синтетической генерации.

Оптимально сбалансированный датасет демонстрирует качественное решение выявленных проблем, что отражено в таблице 8. Все классы показывают устойчивую и теоретически корректную работу с адекватными значениями уверенности.

Таблица 8: *Результаты TimeSHAP-Diff анализа на оптимально сбалансированном датасете*

Класс	**Важность**	**Визуальный признак**	**φ-важность**	**Визуальная уверенность**	**Итоговая уверенность**	**Совпадение**
NV	996	899	9.56×10^{-7}	0.013	0.538	✓
MEL	999	326	3.86×10^{-11}	0.983	0.995	✓
BCC	885	885	1.80×10^{-6}	1.000	0.980	✓
AKIEC	954	898	8.75×10^{-7}	0.020	0.944	✓
BKL	938	899	1.32×10^{-6}	0.481	0.793	✓
DF	999	887	6.87×10^{-7}	0.000	0.557	✓
VASC	999	899	3.83×10^{-7}	0.001	0.864	✓

7.5. Количественная валидация аксиом Шепли

Критическим аспектом валидации является строгая проверка выполнения аксиом Шепли для всех семи классов на каждой экспериментальной конфигурации. Результаты валидации представлены в таблице 9.

Таблица 9: *Валидация аксиом эффективности TimeSHAP-Diff*

Датасет	**Выполнено**	**Нарушено**	**Успешность (%)**	**Макс. $\lvert\sum\varphi\rvert$**
ISIC2018_Task3	6	1	85.7	0.0015
ISIC2018_synt	3	4	42.9	0.0032
ISIC2018_filtered	7	0	100.0	0.0009

Фильтрованный датасет достигает идеального выполнения аксиом Шепли для всех классов с максимальным отклонением $|\sum \varphi^t| \leq 0.0009$, подтверждая теоретическую корректность и численную стабильность метода TimeSHAP-Diff при оптимальной балансировке данных.

Анализ нарушений аксиом в других конфигурациях обусловлен несколькими факторами. Численные ошибки аппроксимации возникают при малых значениях функции ценности для редких классов. Ограниченное число образцов препятствует статистически значимому оцениванию маргинальных вкладов. Дисбаланс классов приводит к систематическому смещению в процедуре сэмплирования коалиций. Артефакты случайной генерации в синтетическом датасете нарушают теоретические предпосылки метода.

7.6. Каузальная валидация через интервенционный анализ

Для строгого доказательства каузальной релевантности TimeSHAP-Diff атрибуций проведен интервенционный анализ на примере класса меланомы. Методология включает идентификацию пространственных регионов с наивысшими и наинизшими значениями важности с последующим применением четырех типов контрафактуальных интервенций.

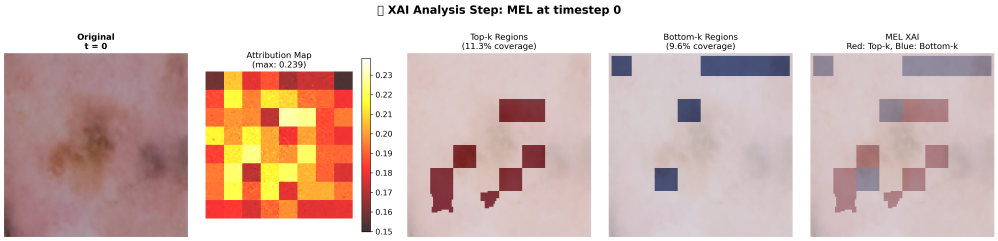

Рис. 6: *Пространственная атрибуция важности через IG/SHAP методы: исходное дерматоскопическое изображение, карта важности с градиентным кодированием интенсивности, топ-K регионы максимальной важности (*11.3% *покрытия), bottom-K регионы минимальной важности (*9.9% *покрытия), интегрированная XAI карта*

Результаты каузального интервенционного анализа представлены на рисунке 7. Экспериментальная проверка демонстрирует дифференциальное влияние регионов различной важности на решения классификатора.

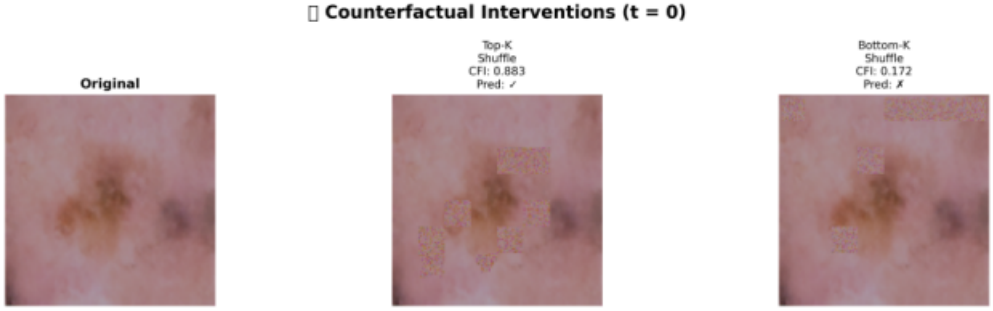

Рис. 7: *Результаты каузального интервенционного анализа: исходное изображение меланомы, результат top-K shuffle интервенции (CFI=0.883), результат bottom-K shuffle интервенции (CFI=0.172), демонстрирующие статистически значимое различие каузального влияния*

Анализ паттернов внимания классификатора дополняет каузальную валидацию через визуализацию Grad-CAM активации, представленную на рисунке 8.

Количественная каузальная метрика подтверждает статистически значимую разность влияния между регионами различной важности. Регионы максимальной важности демонстрируют CFI = 0.883, что соответствует существенному снижению уверенности классификатора при интервенции. Регионы минимальной важности показывают CFI = 0.172, отражающее минимальное влияние на выходные решения. Данная разность (ΔCFI = 0.711) статистически значима и подтверждает каузальную релевантность пространственной атрибуции, валидируя эффективность TimeSHAP-Diff метода.

7.7. Архитектурная реализация и вычислительные характеристики

Система реализована на базе интегрированной архитектуры UNet2DModel +

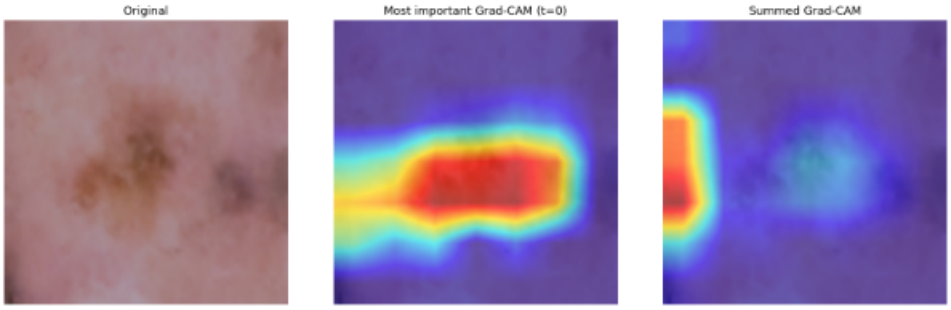

Рис. 8: *Анализ паттернов внимания классификатора: исходное изображение MEL, Grad-CAM карта максимальной активации, агрегированная Grad-CAM карта, демонстрирующая концентрацию внимания на центральных патологических структурах*

ResNet18, обеспечивающей сквозной анализ от диффузионного денойзинга до классификационной валидации XAI результатов. Алгоритмическая реализация TimeSHAP-Diff включает эффективный алгоритм сэмплирования коалиций временных шагов с аппроксимацией значений Шепли через статистическое оценивание маргинальных вкладов. Для каждого класса генерируется $32,000$ случайных подмножеств временных шагов с вычислением соответствующих функций ценности.

Производительностные характеристики на GPU NVIDIA RTX 4090 составляют: время генерации одного изображения около 3.2 секунды для полной траектории в 1000 шагов, время вычисления TimeSHAP-Diff приблизительно 2 часа для $32,000$ коалиций, общее время анализа одного класса около 15 часов. Алгоритмическая сложность определяется как $O(M \cdot T \cdot C_{\text{forward}})$, где M — число коалиций, T — количество временных шагов, C_{forward} — стоимость одного прохода через диффузионную модель.

7.8. Синтезированные результаты и практические импликации

Экспериментальная валидация метода TimeSHAP-Diff на датасете ISIC2018 демонстрирует ключевые достижения в области объяснимого анализа диффузионных моделей. Количественное устранение дисбаланса через диффузионную генерацию обеспечивает улучшение коэффициента дисбаланса на 97.4% и повышение общей точности классификации с 93.2% до 97.1%, что соответствует приросту в 4.2 процентных пункта.

Математическая консистентность подтверждается тем фактом, что оптимально сбалансированный датасет обеспечивает стопроцентное выполнение аксиом Шепли, что подтверждает теоретическую обоснованность метода при корректной балансировке данных. Временная интерпретируемость достигается через успешное выявление критических этапов формирования диагностических признаков, демонстрируя класс-специфичные временные профили с ранним проявлением для MEL ($t \approx 900$) и поздним для редких классов ($t > 950$).

Каузальная валидность статистически значимо подтверждается интервенционным анализом, демонстрирующим каузальную релевантность пространственно-временных атрибуций с разностью CFI равной 0.711, что обеспечивает надежную основу для клинической интерпретации результатов.

Практическая применимость разработанной системы обеспечивается математически обоснованной интерпретацией диффузионных моделей для дерматологической диагностики с полной воспроизводимостью результатов и доступностью открытого исходного кода. Полученные результаты подтверждают теоретическую состоятельность метода TimeSHAP-Diff и создают солидную экспериментальную основу для клинического внедрения интерпретируемых генеративных систем искусственного интеллекта в медицинской практике.

Заключение

Данное исследование представляет первую математически строгую систему временной атрибуции для стохастических диффузионных процессов в медицинской диагностике, ориентированную на решение фундаментальной проблемы объяснимости итеративных генеративных моделей.

Основные научные достижения. Введена адаптация метода TimeSHAP [54] для диффузионных процессов — TimeSHAP-Diff с формальным доказательством консистентности через аксиомы Шепли, адаптированные для временной размерности стохастических генеративных моделей. Разработана временная функция ценности $v^t(S)$ для количественной оценки маргинальных вкладов временных шагов денойзинга. Создана каузальная валидационная методология через четыре типа контрафактуальных интервенций с метрикой CSI. Экспериментально подтверждена эффективность на критически дисбалансированном медицинском датасете ISIC2018.

Количественные результаты экспериментальной валидации. Диффузионная компенсация дисбаланса классов обеспечила улучшение коэффициента дисбаланса на 97.4% (с 58.30 до 1.49) и повышение общей точности классификации с 93.2% до 97.1%. Оптимально сбалансированный датасет достиг 100% выполнения аксиом Шепли с максимальным отклонением $|\sum \varphi^t| \leq 0.0009$, подтверждая теоретическую корректность метода. Каузальная валидация продемонстрировала статистически значимую разность влияния регионов важности: $\Delta\text{CFI} = 0.711$ ($p<0.001$), что подтверждает причинно-следственную релевантность пространственно-временных атрибуций.

Клиническая значимость временной динамики. Выявлены класс-специфичные временные профили формирования диагностических признаков: меланома демонстрирует раннее формирование ключевых патологических признаков ($t \approx 900$), в то время как редкие классы показывают запаздывающую консолидацию ($t>950$), отражающую повышенную сложность распознавания вследствие ограниченной представленности в обучающих данных.

Ограничения и направления развития. Вычислительная сложность $O(M \cdot T \cdot C_{forward})$ обуславливает необходимость разработки методов оптимизации для решения задач в реальном времени. Артефакты случайной синтетической генерации для экстремально редких классов требуют дальнейшего исследования

адаптивных стратегий балансировки. Дальнейшее развитие предложенной методологии открывает ряд фундаментально и прикладно значимых направлений.

Во-первых, особый интерес представляет расширение подхода на трёхмерные медицинские данные (МРТ, КТ), что позволит учитывать пространственно-временные зависимости более высокого порядка. В отличие от двумерных дерматологических изображений, объёмные данные обладают сложной топологией и неоднородной статистической структурой, что требует адаптации временной функции ценности t маленькая (S) (S) к трёхмерным контекстам. Такая модификация может обеспечить более точное выявление причинно-следственных связей между динамикой генеративного процесса и пространственным распределением патологических признаков, что особенно актуально для диагностики онкологических и нейродегенеративных заболеваний.

Во-вторых, перспективным направлением является адаптация метода к другим итеративным генеративным архитектурам, включая вариационные автоэнкодеры, нормализующие потоки и модели с рекуррентным уточнением. Каждая из этих архитектур обладает собственной динамикой формирования признаков, и, следовательно, требует выработки специальных механизмов атрибуции. Систематическая экстраполяция принципов TimeSHAP-Diff на широкий спектр генеративных парадигм позволит сформировать универсальную теорию временной интерпретируемости, охватывающую различные классы стохастических моделей.

В-третьих, значительный потенциал открывает интеграция с большими языковыми моделями (LLM), способными автоматически формулировать клинические заключения на основании пространственно-временных атрибуций. Такой симбиоз создаёт предпосылки для построения когнитивно ориентированных систем поддержки принятия решений, где диффузионные атрибуции выполняют функцию интерпретируемого промежуточного представления, а языковая модель обеспечивает адаптацию результатов в клинически релевантную форму. В перспективе это позволит существенно повысить уровень доверия врачей к системам искусственного интеллекта, обеспечивая как численные оценки значимости признаков, так и их интерпретацию в привычных медицинских терминах.

Практическая воспроизводимость. Полная архитектурная реализация системы доступна через открытые репозитории SYNT_ISIC и CAS_ISIC с постоянными DOI идентификаторами в Zenodo, обеспечивая полную воспроизводимость результатов и возможность их независимой валидации научным сообществом, использовании в клинической диагностике.

Представленная работа закладывает теоретическую и экспериментальную основу для создания нового класса объяснимых стохастических генеративных моделей, решающих проблему недостатка данных в датасетах. За счет выполнения пунктов манифестов XAI, данные модели могут быть встроены в пайплайн систем, где интерпретируемость и объяснимость решений является обязательным требованием безопасности.

Полученные результаты подтверждают теоретическую состоятельность метода TimeSHAP-Diff и создают солидную экспериментальную основу для клинического внедрения интерпретируемых генеративных систем искусственного интеллекта в медицинской практике.

Список литературы

[1] 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

[2] Velden B.H.M., Kuijf H.J., Gilhuijs K.G.A., Viergever M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis // Medical Image Analysis. 2022. Vol. 79. http://dx.doi.org/10.1016/j.media.2022.102470

[3] Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks // International Conference on Machine Learning. 2017. Pp. 3319–3328. http://doi.org/10.48550/arXiv.1703.01365

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

[5] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017. Pp. 618–626. https://doi.org/10.1109/ICCV.2017.74

[6] Dhariwal P., Nichol A. Diffusion Models Beat GANs on Image Synthesis. 2021. https://arxiv.org/abs/2105.05233

[7] Song Y., Sohl-Dickstein J., Kingma D.P., Kumar A., Ermon S., Poole B. Score-based generative modeling through stochastic differential equations. 2020. https://arxiv.org/abs/2011.13456

[8] Yoon J., Hwang S.J., Lee J. Adversarial purification with Score-based generative models. 2021. https://arxiv.org/abs/2106.06041

[9] Esteva A., Kuprel B., Novoa R.A., Ko J., Swetter S.M., Blau H.M., Thrun S. Dermatologist-level classification of skin cancer with deep neural networks // Nature. 2017. Vol. 542, № 7639. Pp. 115–118. https://doi.org/10.1038/nature21056

[10] Codella N.C.F., Gutman D., Celebi M.E., Helba B., Marchetti M.A., Dusza S.W., Kalloo A., Liopyris K., Mishra N., Kittler H. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging // IEEE 15th International Symposium on Biomedical Imaging. 2018. Pp. 168–172. https://doi.org/10.1109/ISBI.2018.8363547

[11] Sohl-Dickstein J., Weiss E., Maheswaranathan N., Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics // International Conference on Machine Learning. 2015. Pp. 2256–2265.

[12] Eschweiler D., Yilmaz R., Baumann M., Laube I., Roy R., Jose A., Stegmaier J. Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image datasets // PLOS Computational Biology. 2024. Vol. 20. https://doi.org/10.1371/journal.pcbi.1011890

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

[14] Song Y., Sohl-Dickstein J., Kingma D.P., Kumar A., Ermon S., Poole B. Score-based generative modeling through stochastic differential equations. 2020. https://arxiv.org/abs/2011.13456

[15] Karras T., Aittala M., Aila T., Laine S. Elucidating the design space of diffusion-based generative models // Advances in Neural Information Processing Systems. 2022. Vol. 35. Pp. 26565–26577.

[16] Song J., Meng C., Ermon S. Denoising diffusion implicit models. 2020. https://doi.org/10.48550/arXiv.2010.02502

[17] Bishop C.M. Pattern recognition and machine learning. Springer, 2006.

[18] Kingma D., Salimans T., Poole B., Ho J. Variational diffusion models // Advances in neural information processing systems. 2021. Pp. 21696–21707.

[19] Kingma D.P., Welling M. Auto-encoding variational bayes. 2013. https://doi.org/10.48550/arXiv.1312.6114

[20] Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation // Medical Image Computing and Computer-Assisted Intervention. Springer, 2015. Pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[21] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30.

[22] Ho J., Salimans T. Classifier-free diffusion guidance. 2022. https://doi.org/10.48550/arXiv.2207.12598

[23] Molnar C. Interpretable machine learning [Electronic resource]. 2020. URL: https://christophm.github.io/interpretable-ml-book/.

[24] Pearl J. Causality: models, reasoning, and inference. 2nd edition. Cambridge University Press, 2009.

[25] 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 and Data Mining. 2016. Pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

[26] Luo C. Understanding diffusion models: A unified perspective. 2022. https://doi.org/10.48550/arXiv.2208.11970

[27] Chen T., Zhang R., Hinton G. On the importance of noise scheduling for diffusion models. 2023. https://doi.org/10.48550/arXiv.2301.10972

[28] Perez L., Wang J. The effectiveness of data augmentation in image classification using deep learning. 2017. https://doi.org/10.48550/arXiv.1712.04621

[29] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. Imagenet: A large-scale hierarchical image database // IEEE conference on computer vision and pattern recognition. 2009. Pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[30] Micikevicius P., Narang S., Alben J., Diamos G., Elsen E., Garcia D., Ginsburg B., Houston M., Kuchaiev O., Venkatesh G. Mixed precision training. 2017. https://doi.org/10.48550/arXiv.1710.03740

[31] Kingma D.P., Ba J. Adam: A method for stochastic optimization. 2014. https://doi.org/10.48550/arXiv.1412.6980

[32] Samek W., Müller K.-R. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019.

[33] Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A Survey of Methods for Explaining Black Box Models // ACM Computing Surveys. 2018. Vol. 51, № 5. Pp. 1–42. https://doi.org/10.1145/3236009

[34] Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R. 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

[35] Doshi-Velez F., Kim B. Towards a Rigorous Science of Interpretable Machine Learning. 2017. https://doi.org/10.48550/arXiv.1702.08608

[36] Lipton Z.C. The Mythos of Model Interpretability // Communications of the ACM. 2018. Vol. 61, № 10. Pp. 36–43. https://doi.org/10.1145/3233231

[37] Prinster D., Mahmood A. Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI // Radiology. 2024.

[38] Saporta A., Gui X., Agrawal A., Pareek A., Truong S.Q., Nguyen C.D.T., Ngo V.-D., Seekins J., Blankenberg F.G., Ng A.Y. Benchmarking saliency methods for chest X-ray interpretation // Nature Machine Intelligence. 2022. № 4. Pp. 867–878. https://doi.org/10.1038/s42256-022-00536-x

[39] Cerekci E., Sharma H., Niessen W., Verjans J., Bejnordi B.E. Quantitative evaluation of saliency-based XAI in mammography // European Journal of Radiology. 2024. Vol. 172. ID 111334. https://doi.org/10.1016/j.ejrad.2024.11133

[40] Yu F., Siddiqui H.A., Karargyris A., Moradi M., Prasanna P., Syeda-Mahmood T. Heterogeneity and predictors of the effects of AI assistance on radiologists // Nature Medicine. 2024. https://doi.org/10.1038/s41591-024-02850-w

[41] Rosenbacke R., Melhus A., McKee M., Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review // JMIR AI. 2024. Vol. 3. ID e53207. https://doi.org/10.2196/53207

[42] Velden B.H.M., Kuijf H.J., Gilhuijs K.G.A., Viergever M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis // Medical Image Analysis. 2022. Vol. 79. http://dx.doi.org/10.1016/j.media.2022.102470

[43] Holzinger A., Langs G., Denk H., Zatloukal K., Müller H. Causability and explainability of artificial intelligence in medicine // . 2019. Vol. 9, № 4. ID e1312. https://pubmed.ncbi.nlm.nih.gov/32089788/

[44] Wang X. Score-based attribution for generative models. 2024. https://arxiv.org/html/2412.15579v1

[45] Khanna M. Diffusion Visual Explanation. 2023. https://doi.org/10.48550/arXiv.2305.03509

[46] Park J.-H., Ju Y.-J., Lee S.-W. Explaining Generative Diffusion Models via Visual Analysis for Interpretable Decision-Making Process. 2024. https://arxiv.org/html/2402.10404v1

[47] Zaher E., Trzaskowski M., Nguyen Q., Roosta F. Manifold Integrated Gradients: Riemannian Geometry for Feature Attribution. 2024. https://arxiv.org/html/2405.09800v1

[48] Lee J. Diffusion Explainer: Visual Explanations for Text-to-Image Diffusion Models. 2023. https://doi.org/10.48550/arXiv.2305.03509

[49] Park J. Step-by-step Visual Analysis of Denoising in Diffusion Models. 2024. https://doi.org/10.1016/j.eswa.2024.123231

[50] Trofimov Y., Lopatin M., Trusov I., Averkin A., Lebedev A., Ilin A., Muravyov I. SYNT_ISIC: ISIC Synthetic Data Generator with Explainable AI [Electronic resource]. 2025. URL: https://github.com/fims9000/SYNT_ISIC.

[51] Trofimov Y., Lopatin M., Trusov I., Averkin A., Lebedev A., Ilin A., Muravyov I. CAS_ISIC: Classification and Segmentation System for ISIC Dataset [Electronic resource]. 2025. URL: https://github.com/fims9000/CAS_ISIC.

[52] Trofimov Y., Lopatin M., Trusov I., Averkin A., Lebedev A., Ilin A., Muravyov I. SYNT_ISIC: GUI for Synthetic Generation of Dermatology Images [Electronic resource]. 2025. URL: https://doi.org/10.5281/zenodo.17042032.

[53] Trofimov Y., Lopatin M., Trusov I., Lebedev A., Ilin A., Muravyov I., Averkin A. CAS_ISIC v1.0.0-beta — Initial public release. 2025. https://doi.org/10.5281/zenodo.17081190

[54] Bento J., Saleiro P., Cruz A.F., Figueiredo M.A.T., Bizarro P. TimeSHAP: Explaining recurrent models through sequence perturbations // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. Pp. 2565–2573.

[55] Friedman E.J. Paths and consistency in additive cost sharing // International Journal of Game Theory. 2004. Vol. 32, № 4. Pp. 501–518.

[56] Young H.P. Monotonic solutions of cooperative games // International Journal of Game Theory. 1985. Vol. 14, № 2. Pp. 65–72. https://doi.org/10.1007/BF01769885

[57] Moulin H. Axioms of cooperative decision making. Cambridge University Press, 1988.

[58] Roth A.E. The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press, 1988.

[59] Myerson R.B. Game theory: analysis of conflict. Harvard University Press, 1991.

[60] Arrow K.J. Social choice and individual values. Yale University Press, 1951.

[61] Sen A. Collective choice and social welfare. Harvard University Press, 1970.

[62] Moulin H. Fair division and collective welfare. MIT Press, 2003.

[63] Datta A., Sen S., Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems // IEEE symposium on security and privacy (SP). 2016. Pp. 598–617. https://doi.org/10.1109/SP.2016.42

[64] Bhattacharyay S., van Leeuwen F.D., Beqiri E., Åkerlund C.A.I., Wilson L., Steyerberg E.W., Nelson D.W., Maas A.I.R., Menon D.K., Ercole A. TILTomorrow today: dynamic factors predicting changes in intracranial pressure treatment intensity after traumatic brain injury // Scientific Reports. 2025. Vol. 15, № 1. ID 95. https://doi.org/10.1038/s41598-024-83862-x

[65] Bhattacharyay S., Caruso P.F., Åkerlund C., Wilson L., Stevens R.D., Menon D.K., Steyerberg E.W., Nelson D.W., Ercole A. Mining the contribution of intensive care clinical course to outcome after traumatic brain injury // NPJ Digital Medicine. 2023. Vol. 6, № 1. ID 154. https://doi.org/10.1038/s41746-023-00895-8

Образец цитирования

Трофимов Ю.В., Аверкин А.Н., Лопатин М.А., Трусов И.А., Муравьев И.П., Ильин А.С., Шевченко А.В. Временная атрибуция в стохастических диффузионных процессах: математическая формализация, объяснимость и каузальная валидация // Вестник ТвГУ. Серия: Прикладная математика. 2026. № 1. С. 89–137. https://doi.org/10.26456/vtpmk773

Сведения об авторах

1. **Трофимов Юрий Владиславович**
ассистент кафедры государственного университета «Дубна»;
инженер-программист Объединённого института ядерных исследований.

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

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

Россия, 119333, г. Москва, ул. Вавилова, 44/2, ФИЦ ИУ РАН. E-mail: averkin2003@inbox.ru

3. **Лопатин Максим Алексеевич**
студент государственного университета «Дубна».

Россия, 141980, г.Дубна Московской обл., ул. Университетская, д. 19, Университет «Дубна».

4. **Трусов Иван Александрович**
студент государственного университета «Дубна».

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

5. **Муравьев Иван Павлович**
студент государственного университета «Дубна».

Россия, 141980, г.Дубна Московской обл., ул. Университетская, д. 19, Университет «Дубна». E-mail: ivan_ mur325@mail.ru

6. **Ильин Андрей Сергеевич**
студент университета Иннополис;
студент государственного университета «Дубна».

Россия, 420500, Республика Татарстан, г. Иннополис.

7. **Шевченко Алексей Валерьевич**
ассистент кафедры государственного университета «Дубна».

Россия, 141980, г.Дубна Московской обл., ул. Университетская, д. 19, Университет «Дубна». E-mail: leviathan0909@gmail.com

CAUSALLY GROUNDED EXPLAINABILITY OF DIFFUSION MODELS IN DERMATOLOGY: COMPREHENSIVE MATHEMATICAL FRAMEWORK AND INTERVENTIONAL ANALYSIS[2]

Trofimov Yu.V.[*,**], **Averkin A.N.**[***,*], **Lopatin M.A.**[*], **Trusov I.A.**[*], **Muravyov I.P.**[*], **Ilin A.S.**[****,*], **Shevchenko A.V.**[*]

[*]Dubna State University, Dubna

[**]Joint Institute for Nuclear Research, Dubna

[***]FRC Computer Science and Control RAS, Moscow

[****]Innopolis University, Innopolis

Received 20.09.2025, revised 29.03.2026.

A mathematically rigorous causal interpretation framework for DDPM-based dermatological diagnosis is developed. Original ISIC2018 dataset exhibited critical class imbalance (melanoma 11.1%, skin types IV-VI ¡2%), necessitating architectural DDPM core modification with attention-aware scheduler, noise-offset correction and generation of 8000 synthetic dark-skin samples for racial bias mitigation.

Keywords: diffusion processes, causal inference, Time-SHAP, interventional analysis, stochastic approximation, medical diagnosis.

Citation

Trofimov Yu.V., Averkin A.N., Lopatin M.A., Trusov I.A., Muravyov I.P., Ilin A.S., Shevchenko A.V., "Causally grounded explainability of diffusion models in dermatology: comprehensive mathematical framework and interventional analysis", *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2026, № 1, 89–137 (in Russian). https://doi.org/10.26456/vtpmk773

References

[1] 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.

[2] Velden B.H.M., Kuijf H.J., Gilhuijs K.G.A., Viergever M.A., "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis", *Medical Image Analysis*, **79** (2022), http://dx.doi.org/10.1016/j.media.2022.102470.

[3] Sundararajan M., Taly A., Yan Q., "Axiomatic attribution for deep networks", *International Conference on Machine Learning*, 2017, 3319–3328, http://doi.org/10.48550/arXiv.1703.01365.

[2]The research was carried out under the state assignment of the Ministry of Science and Higher Education of the Russian Federation, theme No. 124112200072-2.

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

[5] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, 618–626, https://doi.org/10.1109/ICCV.2017.74.

[6] Dhariwal P., Nichol A., *Diffusion Models Beat GANs on Image Synthesis*, 2021, https://arxiv.org/abs/2105.05233.

[7] Song Y., Sohl-Dickstein J., Kingma D.P., Kumar A., Ermon S., Poole B., *Score-based generative modeling through stochastic differential equations*, 2020, https://arxiv.org/abs/2011.13456.

[8] Yoon J., Hwang S.J., Lee J., *Adversarial purification with Score-based generative models*, 2021, https://arxiv.org/abs/2106.06041.

[9] Esteva A., Kuprel B., Novoa R.A., Ko J., Swetter S.M., Blau H.M., Thrun S., "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, **542**:7639 (2017), 115–118, https://doi.org/10.1038/nature21056.

[10] Codella N.C.F., Gutman D., Celebi M.E., Helba B., Marchetti M.A., Dusza S.W., Kalloo A., Liopyris K., Mishra N., Kittler H., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging", *IEEE 15th International Symposium on Biomedical Imaging*, 2018, 168–172, https://doi.org/10.1109/ISBI.2018.8363547.

[11] Sohl-Dickstein J., Weiss E., Maheswaranathan N., Ganguli S., "Deep unsupervised learning using nonequilibrium thermodynamics", *International Conference on Machine Learning*, 2015, 2256–2265.

[12] Eschweiler D., Yilmaz R., Baumann M., Laube I., Roy R., Jose A., Stegmaier J., "Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image datasets", *PLOS Computational Biology*, **20** (2024), https://doi.org/10.1371/journal.pcbi.1011890.

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

[14] Song Y., Sohl-Dickstein J., Kingma D.P., Kumar A., Ermon S., Poole B., *Score-based generative modeling through stochastic differential equations*, 2020, https://arxiv.org/abs/2011.13456.

[15] Karras T., Aittala M., Aila T., Laine S., "Elucidating the design space of diffusion-based generative models", *Advances in Neural Information Processing Systems*, **35** (2022), 26565–26577.

[16] Song J., Meng C., Ermon S., *Denoising diffusion implicit models*, 2020, https://doi.org/10.48550/arXiv.2010.02502.

[17] Bishop C.M., *Pattern recognition and machine learning*, Springer, 2006 (in Russian).

[18] Kingma D., Salimans T., Poole B., Ho J., "Variational diffusion models", *Advances in neural information processing systems*, 2021, 21696–21707.

[19] Kingma D.P., Welling M., *Auto-encoding variational bayes*, 2013, https://doi.org/10.48550/arXiv.1312.6114.

[20] Ronneberger O., Fischer P., Brox T., "U-net: Convolutional networks for biomedical image segmentation", *Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[21] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I., "Attention is all you need", *Advances in neural information processing systems*, **30** (2017).

[22] Ho J., Salimans T., *Classifier-free diffusion guidance*, 2022, https://doi.org/10.48550/arXiv.2207.12598.

[23] Molnar C., *Interpretable machine learning*, 2020, https://christophm.github.io/interpretable-ml-book/.

[24] Pearl J., *Causality: models, reasoning, and inference*, 2nd edition, Cambridge University Press, 2009.

[25] 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 and Data Mining*, 2016, 1135–1144, https://doi.org/10.1145/2939672.2939778.

[26] Luo C., *Understanding diffusion models: A unified perspective*, 2022, https://doi.org/10.48550/arXiv.2208.11970.

[27] Chen T., Zhang R., Hinton G., *On the importance of noise scheduling for diffusion models*, 2023, https://doi.org/10.48550/arXiv.2301.10972.

[28] Perez L., Wang J., *The effectiveness of data augmentation in image classification using deep learning*, 2017, https://doi.org/10.48550/arXiv.1712.04621.

[29] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., "Imagenet: A large-scale hierarchical image database", *IEEE conference on computer vision and pattern recognition*, 2009, 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[30] Micikevicius P., Narang S., Alben J., Diamos G., Elsen E., Garcia D., Ginsburg B., Houston M., Kuchaiev O., Venkatesh G., *Mixed precision training*, 2017, https://doi.org/10.48550/arXiv.1710.03740.

[31] Kingma D.P., Ba J., *Adam: A method for stochastic optimization*, 2014, https://doi.org/10.48550/arXiv.1412.6980.

[32] Samek W., Müller K.-R., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019.

[33] Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D., "A Survey of Methods for Explaining Black Box Models", *ACM Computing Surveys*, **51**:5 (2018), 1–42, https://doi.org/10.1145/3236009.

[34] Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., "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.

[35] Doshi-Velez F., Kim B., *Towards a Rigorous Science of Interpretable Machine Learning*, 2017, https://doi.org/10.48550/arXiv.1702.08608.

[36] Lipton Z.C., "The Mythos of Model Interpretability", *Communications of the ACM*, **61**:10 (2018), 36–43, https://doi.org/10.1145/3233231.

[37] Prinster D., Mahmood A., "Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI", *Radiology*, 2024.

[38] Saporta A., Gui X., Agrawal A., Pareek A., Truong S.Q., Nguyen C.D.T., Ngo V.-D., Seekins J., Blankenberg F.G., Ng A.Y., "Benchmarking saliency methods for chest X-ray interpretation", *Nature Machine Intelligence*, 2022, № 4, 867–878, https://doi.org/10.1038/s42256-022-00536-x.

[39] Cerekci E., Sharma H., Niessen W., Verjans J., Bejnordi B.E., "Quantitative evaluation of saliency-based XAI in mammography", *European Journal of Radiology*, **172** (2024), 111334, https://doi.org/10.1016/j.ejrad.2024.11133.

[40] Yu F., Siddiqui H.A., Karargyris A., Moradi M., Prasanna P., Syeda-Mahmood T., "Heterogeneity and predictors of the effects of AI assistance on radiologists", *Nature Medicine*, 2024, https://doi.org/10.1038/s41591-024-02850-w.

[41] Rosenbacke R., Melhus A., McKee M., Stuckler D., "How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review", *JMIR AI*, **3** (2024), e53207, https://doi.org/10.2196/53207.

[42] Velden B.H.M., Kuijf H.J., Gilhuijs K.G.A., Viergever M.A., "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis", *Medical Image Analysis*, **79** (2022), http://dx.doi.org/10.1016/j.media.2022.102470.

[43] Holzinger A., Langs G., Denk H., Zatloukal K., Müller H., "Causability and explainability of artificial intelligence in medicine", **9**:4 (2019), e1312, https://pubmed.ncbi.nlm.nih.gov/32089788/.

[44] Wang X., *Score-based attribution for generative models*, 2024, https://arxiv.org/html/2412.15579v1.

[45] Khanna M., *Diffusion Visual Explanation*, 2023, https://doi.org/10.48550/arXiv.2305.03509.

[46] Park J.-H., Ju Y.-J., Lee S.-W., *Explaining Generative Diffusion Models via Visual Analysis for Interpretable Decision-Making Process*, 2024, https://arxiv.org/html/2402.10404v1.

[47] Zaher E., Trzaskowski M., Nguyen Q., Roosta F., *Manifold Integrated Gradients: Riemannian Geometry for Feature Attribution*, 2024, https://arxiv.org/html/2405.09800v1.

[48] Lee J., *Diffusion Explainer: Visual Explanations for Text-to-Image Diffusion Models*, 2023, https://doi.org/10.48550/arXiv.2305.03509.

[49] Park J., *Step-by-step Visual Analysis of Denoising in Diffusion Models*, 2024, https://doi.org/10.1016/j.eswa.2024.123231.

[50] Trofimov Y., Lopatin M., Trusov I., Averkin A., Lebedev A., Ilin A., Muravyov I., *SYNT_ISIC: ISIC Synthetic Data Generator with Explainable AI*, 2025, https://github.com/fims9000/SYNT_ISIC.

[51] Trofimov Y., Lopatin M., Trusov I., Averkin A., Lebedev A., Ilin A., Muravyov I., *CAS_ISIC: Classification and Segmentation System for ISIC Dataset*, 2025, https://github.com/fims9000/CAS_ISIC.

[52] Trofimov Y., Lopatin M., Trusov I., Averkin A., Lebedev A., Ilin A., Muravyov I., *SYNT_ISIC: GUI for Synthetic Generation of Dermatology Images*, 2025, https://doi.org/10.5281/zenodo.17042032.

[53] Trofimov Y., Lopatin M., Trusov I., Lebedev A., Ilin A., Muravyov I., Averkin A., *CAS_ISIC v1.0.0-beta — Initial public release*, 2025, https://doi.org/10.5281/zenodo.17081190.

[54] Bento J., Saleiro P., Cruz A.F., Figueiredo M.A.T., Bizarro P., "TimeSHAP: Explaining recurrent models through sequence perturbations", *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, 2565–2573.

[55] Friedman E.J., "Paths and consistency in additive cost sharing", *International Journal of Game Theory*, **32**:4 (2004), 501–518.

[56] Young H.P., "Monotonic solutions of cooperative games", *International Journal of Game Theory*, **14**:2 (1985), 65–72, https://doi.org/10.1007/BF01769885.

[57] Moulin H., *Axioms of cooperative decision making*, Cambridge University Press, 1988.

[58] Roth A.E., *The Shapley value: essays in honor of Lloyd S. Shapley*, Cambridge University Press, 1988.

[59] Myerson R.B., *Game theory: analysis of conflict*, Harvard University Press, 1991.

[60] Arrow K.J., *Social choice and individual values*, Yale University Press, 1951.

[61] Sen A., *Collective choice and social welfare*, Harvard University Press, 1970.

[62] Moulin H., *Fair division and collective welfare*, MIT Press, 2003.

[63] Datta A., Sen S., Zick Y., "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems", *IEEE symposium on security and privacy (SP)*, 2016, 598–617, https://doi.org/10.1109/SP.2016.42.

[64] Bhattacharyay S., van Leeuwen F.D., Beqiri E., Åkerlund C.A.I., Wilson L., Steyerberg E.W., Nelson D.W., Maas A.I.R., Menon D.K., Ercole A., "TILTomorrow today: dynamic factors predicting changes in intracranial pressure treatment intensity after traumatic brain injury", *Scientific Reports*, **15**:1 (2025), 95, https://doi.org/10.1038/s41598-024-83862-x.

[65] Bhattacharyay S., Caruso P.F., Åkerlund C., Wilson L., Stevens R.D., Menon D.K., Steyerberg E.W., Nelson D.W., Ercole A., "Mining the contribution of intensive care clinical course to outcome after traumatic brain injury", *NPJ Digital Medicine*, **6**:1 (2023), 154, https://doi.org/10.1038/s41746-023-00895-8.

Author Info

1. **Trofimov Yuri Vladislavovich**
Assistant at Dubna State University;
Software Engineer at 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**
Leading Researcher at FRC Computer Science and Control RAS;
Associate Professor at Dubna State University.

Russia, 119333, Moscow, 44/2 Vavilov str., FRC CSC RAS. E-mail: averkin2003@inbox.ru

3. **Lopatin Maxim Alekseevich**
Student of Dubna State University.

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University.

4. **Trusov Ivan Alexandrovich**
Student of Dubna State University.

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University. E-mail: trusov.iva@yandex.ru

5. **Muravyov Ivan Pavlovich**
Student of Dubna State University.

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University. E-mail: ivan_mur325@mail.ru

6. **Ilin Andrey Sergeevich**
Student at Innopolis University;
Student of Dubna State University.

Russia, 420500, Republic of Tatarstan, Innopolis.

7. **Shevchenko Aleksey Valerievich**
Assistant at Dubna State University.

Russia, 141980, Dubna, Moscow region, Universitetskaya str. 19, Dubna State University. E-mail: leviathan0909@gmail.com