НАУЧНЫЙ ОБЗОР

УДК 004.827:519.816
DOI 10.17513/snt.40785

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ИНТЕРПРЕТИРУЕМОСТИ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ПРОЦЕССАХ ПРИНЯТИЯ РЕШЕНИЙ

Матвеев А. В. ORCID ID 0000-0002-0778-3218

Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский университет Государственной противопожарной службы МЧС России имени Героя Российской Федерации генерала армии Е. Н. Зиничева», Санкт-Петербург, Российская Федерация, e-mail: fcvega_10@mail.ru

Современное внедрение технологий искусственного интеллекта в процессы принятия решений сопровождается проблемой их недостаточной интерпретируемости. Это создает недоверие со стороны пользователей и лиц, принимающих решения, проблему прозрачности и ответственного применения алгоритмов в критически важных областях, где решения алгоритмов напрямую влияют на безопасность людей, экономическую стабильность, соблюдение прав человека и функционирование государственных институтов. Цель исследования – проведение сравнительного анализа известных методов интерпретации моделей искусственного интеллекта для выявления преимуществ и ограничений каждого из методов по ряду критериев, включая теоретическую обоснованность, простоту интерпретации, вычислительную сложность, применимость к различным типам данных и моделей, масштабируемость, удобство для пользователей, поддержку объяснения индивидуальных решений, возможность экспертной верификации решения. Методологией исследования является систематический сравнительный анализ, построенный на комплексном изучении релевантных научных публикаций, входящих в базы Scopus и Web of Science за 2018–2025 гг. В результате исследования были рассмотрены и систематизированы пять широко известных подходов к интерпретируемости моделей искусственного интеллекта. Результаты анализа показали, что не существует универсального метода. Основной вывод заключается в необходимости выбора метода интерпретации в зависимости от конкретной задачи, уровня ответственности решения, доступных вычислительных ресурсов и подготовки пользователей. Подчеркивается важность комбинирования методов и определяются перспективные направления для дальнейшей работы, такие как разработка гибридных подходов и метрик оценки интерпретируемости с точки зрения конечного пользователя.

Ключевые слова: искусственный интеллект, интерпретируемость, принятие решений, обоснованность

THE COMPARATIVE ANALYSIS OF INTERPRETABILITY METHODS OF ARTIFICIAL INTELLIGENCE MODELS IN DECISION-MAKING PROCESSES

Matveev A. V. ORCID ID 0000-0002-0778-3218

Federal State Budgetary Educational Institution of Higher Education "Saint Petersburg University of the State Fire Service of the Ministry of Emergency Situations of Russia named after the Hero of the Russian Federation, General of the Army E.N. Zinichev", Saint Petersburg, Russian Federation, e-mail: fcvega_10@mail.ru

The modern integration of artificial intelligence technologies into decision-making processes is accompanied by the problem of their insufficient interpretability. This creates mistrust among users and decision-makers, as well as challenges in the transparency and responsible application of algorithms in critical areas where algorithmic decisions directly impact human safety, economic stability, human rights, and the functioning of public institutions. The aim of this study is to conduct a comparative analysis of known methods for interpreting artificial intelligence models to identify the advantages and limitations of each method across a number of criteria, including theoretical soundness, ease of interpretation, computational complexity, applicability to various types of data and models, scalability, user friendliness, support for explaining individual decisions, and the possibility of expert verification of the solution. The research methodology is a systematic comparative analysis based on a comprehensive study of relevant scientific publications included in the Scopus and Web of Science databases for the period 2018-2025. The study examined and systematized five widely known approaches to the interpretability of artificial intelligence models. The analysis showed that there is no universal method. The key conclusion is the need to select an interpretation method based on the specific task, the criticality of the solution, available computing resources, and user experience. The importance of combining methods is emphasized, and promising areas for further work are identified, such as the development of hybrid approaches and metrics for assessing interpretability from the end-user perspective.

Keywords: artificial intelligence, interpretability, decision making, reasoning

Введение

Современный этап развития искусственного интеллекта (ИИ) характеризуется его стремительным внедрением в процессы принятия решений в различных сферах человеческой деятельности [1]. Однако по мере возрастания сложности моделей ИИ появляется проблема их непрозрачности, часто обозначаемая в научной литературе термином «черный ящик». Эта непрозрачность создает фундаментальную проблему в контексте практического применения ИИ [2, 3].

Интерпретируемость ИИ можно определить как способность модели объяснить, как она пришла к тому или иному выводу. Интерпретируемость ИИ важна по нескольким причинам. Во-первых, она обеспечивает доверие к системам ИИ со стороны лиц, принимающих решения (ЛПР) или других заинтересованных сторон. Во-вторых, интерпретируемость позволяет выявлять потенциальные ошибки, предвзятости и возможности дискриминации в работе алгоритмов, использующих технологии ИИ. Кроме того, интерпретируемость раскрывает механизмы принятия решений, которые используют сложные модели, что обеспечивает повышение их обоснованности и адекватности.

Необходимо отметить, что обеспечение высокой степени интерпретируемости моделей ИИ представляет собой достаточно сложную задачу. Особенно это проявляется при применении ансамблевых методов или глубоких нейронных сетей, отличающихся высокой степенью сложности. В ряде научных исследований выявлено существующее противоречие между точностью модели и ее интерпретируемостью, в которых показано, что рост точности сопровождается снижением прозрачности и объяснимости результатов [4]. Простые линейные модели характеризуются высокой степенью интерпретируемости, однако их прогностическая способность ограничена в случае сложных нелинейных зависимостей. Глубокие нейронные сети, напротив, при решении сложных задач обладают высокой прогностической способностью, при этом их внутренняя архитектура и процессы преобразования входных данных в выходные слабо поддаются интерпретации со стороны пользователей, что ограничивает прозрачность и объяснимость таких моделей и остается достаточно серьезной проблемой для ЛПР [5].

В последние годы в научной литературе было предложено множество подходов и методов к оценке интерпретируемости существующих моделей ИИ [6]. Однако практический интерес имеет сравнительный анализ их сильных и слабых сторон в разрезе решения различного класса задач принятия решений и различных требований к интерпретируемости моделей ИИ. Существующие исследования, как правило, базируются на отдельных особенностях различных методов или демонстрируют их практическое применение в специфических предметных областях. Комплексная оценка преимуществ и ограничений методов интерпретируемости моделей ИИ в целом не представлена. Данная совокупность факторов в целом препятствует обоснованному выбору конкретных методов интерпретации для различных практических задач.

При возрастающей сложности моделей ИИ и расширения сфер и направлений их применения в практической деятельности становится важным понимание не только характеристик существующих методов интерпретации, но и их соответствия различным требованиям заинтересованных сторон и ограничениям соответствующих предметных областей. Различные области применения ИИ предъявляют специфические требования к интерпретируемости [7]. Так, в частности, в медицинской диагностике необходима точность и полнота объяснений для обоснования клинических решений, в финансовом секторе важно соответствие нормативным требованиям, а в автономных системах необходима скорость интерпретации для обеспечения безопасности. Эти разнообразные требования делают особенно актуальным систематическое сравнение методов интерпретации по множеству критериев.

Цель исследования – проведение сравнительного анализа известных методов интерпретации моделей искусственного интеллекта для выявления преимуществ и ограничений каждого из методов по ряду критериев, включая теоретическую обоснованность, простоту интерпретации, вычислительную сложность, применимость к различным типам данных и моделей, масштабируемость, удобство для пользователей, поддержку объяснения индивидуальных решений, возможность экспертной верификации решения.

Материалы и методы исследования

Исследование носит аналитический характер и основывается на анализе современных научных работ в области интерпретируемости ИИ. В исследовании использовались научные труды, входящие в базы Scopus и Web of Science. Временной промежуток поиска источников – с 2018 г. по настоящее время. В целом было проанализи-

ровано более 40 трудов, из них в ходе анализа было отобрано 25 работ, наиболее релевантных цели проводимого исследования.

Данное исследование направлено на выявление сравнительных преимуществ и ограничений каждого метода по ряду критериев, включая теоретическую обоснованность, простоту интерпретации, вычислительную сложность, применимость к различным типам данных и моделей, масштабируемость, удобство для пользователей. Ожидается, что результаты исследования внесут вклад в развитие теории интерпретируемости моделей ИИ и предоставят практические рекомендации по выбору адекватных методов интерпретации для таких предметных областей, как здравоохранение, финансовый сектор, сфера безопасности и др.

Методы интерпретации моделей ИИ классифицируются по различным критериям [8]: по уровню применения делятся на глобальные, направленные на понимание общей логики работы модели, и локальные, объясняющие конкретные предсказания для отдельных примеров; по степени зависимости от конкретной архитектуры модели выделяют специальные методы, разработанные для определенных типов моделей, и универсальные методы, применимые к любым моделям машинного обучения независимо от их внутренней структуры [9]. Специальным методам часто не хватает гибкости, поскольку они не применимы к моделям с разной архитектурой. Универсальные методы, напротив, предназначены для использования с любой моделью, обеспечивая гибкость и широкую применимость, что делает их особенно ценными в областях, где используются различные модели ИИ.

В данном исследовании рассмотрены пять известных методов интерпретации моделей ИИ (LIME, SHAP, DPD, ICE, Anchors), наиболее часто встречающихся в научной литературе, исследуются их механизмы, преимущества и недостатки, а также практические аспекты их применения.

Результаты исследования и их обсуждение

1. Метод LIME (Local Interpretable Model-agnostic Explanations)

Метод LIME основывается на построении локально точной интерпретируемой аппроксимации сложной модели в окрестности объясняемого предсказания [10]. Фундаментальная идея заключается в том, что, даже если глобальное поведение модели чрезвычайно сложно, локально в небольшой области пространства признаков модель может быть адекватно аппроксимирована простой интерпретируемой моделью, например линейной [10].

Алгоритм метода локальной интерпретации состоит из нескольких ключевых этапов [11]. На первом этапе генерируется набор возмущенных вариантов исходных данных путем случайного изменения значений признаков исходного наблюдения, для которого необходимо получить объяснение. Эти возмущенные данные формируют выборку из окрестности исходной точки в пространстве признаков. Количество генерируемых возмущенных примеров является гиперпараметром метода и влияет на точность аппроксимации и вычислительную сложность. На втором этапе для каждого возмущенного примера получается предсказание от объясняемой модели, что позволяет собрать данные о локальном поведении модели. На третьем этапе каждому возмущенному примеру присваивается вес, отражающий его близость к исходному объясняемому примеру (чем новый пример ближе к исходному, тем выше его вес).

На заключительном этапе производится обучение интерпретируемой модели, обычно линейной регрессии или логистической регрессии для задач классификации, на возмущенных вариантах исходных данных с использованием функции ошибки, в которой каждому возмущенному наблюдению назначается вес, отражающий его близость к объясняемому объекту, что обеспечивает приоритетное влияние локальных наблюдений при обучении интерпретируемой аппроксимирующей модели. Целевой переменной для обучения являются предсказания объясняемой модели, полученные на втором этапе. Коэффициенты обученной линейной модели интерпретируются как показатели важности признаков для данного конкретного предсказания: признаки с большими по абсолютному значению коэффициентами оказывают более сильное влияние на предсказание в локальной окрестности [12].

Метод LIME не требует доступа к архитектуре исследуемой модели и может быть применен к любой модели машинного обучения.

2. Метод SHAP (SHapley Additive exPlanations)

Метод SHAP (в некоторых источниках носит название «Метод значений Шепли» или Shapley values) основан на теории игр, где значения Шепли представляют собой способ количественной оценки вклада каждого признака в предсказание модели для конкретного наблюдения [13]. Фундаментальная идея заключается в рассмотре-

нии признаков как игроков в кооперативной игре, где выигрышем является предсказание модели. При этом вклад в общий выигрыш должен удовлетворять набору аксиом: эффективности (сумма вкладов равна общему выигрышу), симметрии (равный вклад – равные значения), аксиоме нулевого игрока (игрок, не влияющий на выигрыш, имеет нулевой вклад) и аддитивности (вклады аддитивны по играм). Значение Шепли для каждого признака определяется как средний вклад этого признака в предсказания [13].

Значение Шепли для признака определяется через рассмотрение всех возможных перестановок признаков и вычисление разности между предсказанием модели с включением данного признака и без него для каждой перестановки. Формально для признака в наборе из общего количества признаков значение Шепли вычисляется как средневзвешенная сумма по всем подмножествам признаков, не содержащим данный признак. Весовой коэффициент для каждого подмножества зависит от его размера и отражает вероятность формирования данной коалиции при случайном упорядочивании признаков. Это обеспечивает справедливое распределение вклада между признаками в соответствии с их истинным влиянием на предсказание модели.

Метод SHAP дает локальные объяснения (вклад признаков для конкретного примера) и при этом накапливает информацию по объектам для построения глобального объяснения (средние вклады признаков по всем данным). Метод гарантирует согласованность: если в модели вклад признака увеличивается, его оценка не уменьшится, а сумма вкладов всегда точно соответствует результату модели.

Практическое вычисление точных значений Шепли для реальных приложений машинного обучения сталкивается с серьезными вычислительными трудностями. Количество возможных коалиций признаков растет экспоненциально с числом признаков, что делает точное вычисление значений Шепли неразрешимым для задач с большим количеством признаков [14]. Для преодоления проблемы вычислительной сложности предложены различные подходы к вычислению значений Шепли [15]. Один из наиболее известных подходов основывается на методе Монте-Карло, который заключается в случайной выборке подмножества всех возможных перестановок признаков и усреднении маргинальных вкладов по этой выборке [16]. Этот метод обеспечивает несмещенную оценку значений Шепли, а точность оценки может быть обеспечена путем увеличения числа выборок. При этом даже при использовании подобных аппроксимационных методов вычислительные затраты все еще остаются значительными.

Еще одним подходом к аппроксимации значений Шепли является метод, известный как KernelSHAP (обобщенный подход с генерацией искусственных образцов), который для вычисления значений Шепли использует взвешенную линейную регрессию [17, 18]. Этот метод использует специальным образом подобранные веса для различных коалиций признаков, что позволяет эффективно аппроксимировать значения Шепли без необходимости перебора всех возможных коалиций.

Таким образом, использование метода SHAP позволяет определить вклад каждого признака в результаты предсказания модели и обеспечивает ряд важных свойств объяснений (аддитивность, симметрию и согласованность) [13].

3. Метод PDP (Partial Dependence Plots)

Метод PDP (графиков частичной зависимости) предназначен для анализа влияния отдельных признаков (или их комбинаций) на результаты, которые дают модели ИИ при усреднении вклада остальных признаков [19]. Данный метод может быть применен к любой модели машинного обучения (от простой линейной регрессии до нейронных сетей или градиентного бустинга) без необходимости знания ее внутренней структуры, позволяет визуализировать влияние изменения значения одного или нескольких признаков на результаты предсказания модели в среднем по всему набору данных, что позволяет его использовать для оценки интерпретируемости сложных моделей.

Алгоритм применения метода PDP может быть представлен в следующей последовательности шагов. Выбирается признак (или пара признаков) для анализа. Для каждого уникального значения признака формируется множество так называемых синтетических объектов (исходные данные, но с фиксированным значением для анализируемого признака, при этом все остальные признаки остаются неизменными). Модель многократно вычисляет предсказания для всех этих объектов, после чего полученные значения усредняются. Итоговый график частичной зависимости отражает усредненное влияние признака на результат, который дает модель, сглаживая индивидуальные вариации, обусловленные взаимодействием с другими признаками [20]. Таким образом, метод позволяет понять, как конкретный признак влияет на пред-

сказание модели для отдельных объектов, а не в среднем по выборке. Для двух признаков процедура аналогична, но результатом становится уже двумерная поверхность отклика.

С точки зрения интерпретируемости PDP позволяет выявить направление и характер влияния признака на выход модели (линейное, нелинейное, монотонное, пороговое), что делает метод особенно полезным для анализа глобальных закономерностей, заложенных в модель [21], а также для верификации соответствия поведения модели экспертным ожиданиям.

4. Метод ICE (Individual Conditional Expectation)

Метод ICE (метод индивидуальных условных ожиданий) относится к классу глобально-локальных методов оценки интерпретации моделей ИИ и предназначен для анализа влияния отдельных признаков на предсказание модели на уровне отдельных наблюдений [22]. Данный метод является в целом развитием метода PDP и позволяет преодолеть одно из его ключевых ограничений, а именно потерю информации об индивидуальной гетерогенности данных вследствие усреднения.

Суть метода ICE состоит в следующем. Для каждого наблюдения из исходной выборки последовательно изменяются значения одного выбранного признака в заданном диапазоне, после чего вычисляются соответствующие предсказания модели [23]. Затем полученные значения визуализируются в виде набора кривых (ICE-линий), каждая из которых соответствует одному объекту. Вся совокупность таких кривых отражает индивидуальные траектории влияния признака на результат модели.

С точки зрения интерпретируемости ICE занимает промежуточное положение между локальными и глобальными методами. С одной стороны, ICE предоставляет локальную информацию, поскольку объясняет влияние признака на конкретные предсказания. С другой стороны, анализ совокупности ICE-кривых позволяет получить представление о глобальном поведении модели и выявить общие паттерны и аномалии. Часто на практике ICE-графики используются совместно с PDP, при этом PDP интерпретируется как среднее значение набора ICE-кривых.

В контексте интеграции ИИ в процесс принятия решений метод ICE представляет значимую ценность, поскольку позволяет выявлять индивидуальные различия в логике работы модели, что особенно важно в задачах с персонализированными решениями (например, оценка рисков, рекомендации, диагностика). Использование ICE способствует повышению прозрачности моделей, выявлению потенциальных смещений и обеспечению более обоснованного и ответственного применения интеллектуальных систем.

5. Метод Anchors

Метод Anchors относится к классу локальных методов оценки интерпретации моделей ИИ, направленных на получение четких, высокоточных и человеко-интерпретируемых правил, объясняющих отдельные предсказания сложных моделей. Метод Anchors был предложен в развитие подхода LIME и ориентирован на формирование объяснений в виде логических правил вида «если – то», которые обладают высокой степенью надежности в локальной области пространства признаков [24].

Основная идея метода Anchors заключается в поиске набора условий, которые «якорят» предсказание модели, то есть обеспечивают его устойчивость при изменении остальных признаков [25]. Метод представляет собой конъюнкцию условий на значения признаков, при выполнении которых предсказание модели с высокой вероятностью остается неизменным. Таким образом, объяснение в методе Anchors формулируется не как взвешенный вклад признаков, а как интерпретируемое правило, гарантирующее сохранение результата модели.

Результаты исследования и их обсуждение

Сравнительный анализ рассмотренных методов интерпретации моделей был проведен по ряду критериев, по которым данные подходы демонстрируют как сходства, так и существенные различия. Комплексное понимание этих сходств и различий важно для обоснованного выбора метода интерпретации в конкретных практических ситуациях. В таблице представлено детальное сравнение методов по ряду критериев, выявляющих особенности их практического применения в системах поддержки принятия решений.

Так как основная идея метода LIME заключается в аппроксимации поведения сложной модели в окрестности конкретного наблюдения с помощью простой и интерпретируемой суррогатной модели (как правило, линейной регрессии), то это обеспечивает высокую наглядность объяснений и позволяет быстро получить интуитивное понимание факторов, повлиявших на отдельное предсказание.

Результаты анализа методов интерпретации моделей ИИ при их интеграции в процессы принятия решений

Метод / Критерий	LIME	SHAP	PDP	ICE	Anchors
Локальная/глобальная интерпретируемость	Локальная	Локальная и глобальная	Глобальная	Локальная и глобальная	Локальная
Теоретическая обоснованность	Средняя	Высокая	Средняя	Средняя	Высокая
Простота интерпретации	Высокая	Средняя	Очень высокая	Средняя	Очень высокая
Вычислительная сложность	Средняя	Высокая	Низкая	Средняя	Высокая
Применимость к различным типам данных и моделей	Высокая	Высокая	Средняя	Средняя	Высокая
Масштабируемость	Высокая	Низкая	Очень высокая	Высокая	Низкая
Удобство для пользователя	Высокое	Среднее	Высокое	Среднее	Высокое
Поддержка объяснения индивидуальных решений	Высокая	Высокая	Низкая	Средняя	Очень высокая
Возможность экспертной верификации решения	Средняя	Высокая	Высокая	Средняя	Высокая

Примечание: составлена автором на основе полученных данных в ходе исследования.

К достоинству метода LIME относится возможность его применения к широкому спектру алгоритмов машинного обучения без необходимости доступа к их внутренней архитектуре. Также LIME отличается сравнительно низкой вычислительной сложностью и высокой гибкостью. Данная совокупность факторов делает его удобным инструментом для оперативного анализа решений модели, включая использование в интерактивных системах поддержки принятия решений.

К ограничениям метода LIME в первую очередь относится то, что он носит эвристический характер. Результаты интерпретируемости модели зависят от результатов генерации возмущенных данных и выбора метрики близости, а это потенциально может снижать устойчивость и воспроизводимость объяснений. Еще одним ограничением метода LIME является то, что он формирует исключительно локальные объяснения (для каждого конкретного наблюдения он аппроксимирует поведение исходной модели в малой окрестности этой точки с помощью простой интерпретируемой модели) и не позволяет делать выводы о глобальном поведении модели. В случаях существующих нелинейных зависимостей локальная линейная аппроксимация может искажать реальную логику работы модели, что может ограничивать возможности применения метода в критически значимых предметных областях.

Метод SHAP, который основывается на аддитивной структуре, а также аксиомах согласованности и локальной точности, в целом позволяет обеспечить высокое качество и строгость объяснений. Основным достоинством метода является его способность обеспечивать как локальные, так и глобальные интерпретации. Оценки значений Шепли, показывающих вклады признаков для отдельных наблюдений, могут быть агрегированы, что позволяет анализировать общее поведение модели на уровне всего набора данных. Наличие специализированных реализаций метода для деревьев решений и нейронных сетей повышает его практическую эффективность. Высокая устойчивость объяснений делает метод SHAP особенно востребованным в критически важных областях, например финансах или медицине. Практическое применение метода ограничивается в условиях большого количества признаков ввиду высокой вычислительной сложности, а также при их коррелированности (зависимости).

Метод PDP характеризуется высокой наглядностью и относительной простотой интерпретации. Данный метод является удобным средством для выявления общих тенденций, нелинейностей и пороговых эффектов в поведении модели. Метод хорошо подходит для верификации логики модели и сопоставления ее результатов с экспертными ожиданиями, отличается относительно невысокой вычислительной сложностью и может быть применен к различным типам моделей, включая ансамбли и нейронные сети.

При этом метод PDP имеет ряд ограничений. В первую очередь это усреднение эффектов по всему набору данных, что при-

водит к потере информации об индивидуальных различиях между наблюдениями. Кроме того, метод предполагает условную независимость признаков, а в случае наличия сильной их корреляции может искажать результаты интерпретации. Метод PDP не предназначен для объяснения конкретных предсказаний и поэтому ограничивается возможность его использования в задачах, требующих персонализированных объяснений решений модели.

Метод ICE является особенно полезным для выявления гетерогенности эффектов. Главным его достоинством является способность выявлять различия в реакции модели на изменение признаков для различных объектов, что позволяет обнаруживать скрытые нелинейные зависимости, взаимодействия признаков и подгруппы данных с отличающимся поведением, что невозможно при использовании усредненных методов. Метод ICE сочетает в себе элементы локальной и глобальной интерпретации, предоставляя возможности для глубокого понимания структуры модели.

К достоинствам метода Anchors можно отнести его высокую ясность и надежность объяснений. Формулировка в виде правил «если – то» хорошо соответствует логике человеческого мышления и легко интегрируется в процессы принятия решений, что делает его человеко-ориентированным.

Ограничением метода Anchors является его высокая вычислительная сложность, обусловленная поиском оптимальных правил, а также ограниченная масштабируемость в данных с большим признаковым пространством. Метод предоставляет исключительно локальные объяснения и не предназначен для анализа глобальной структуры модели.

Заключение

В исследовании проведен комплексный сравнительный анализ пяти известных методов интерпретации моделей ИИ: LIME, SHAP, PDP, ICE и Anchors. Каждый метод характеризуется уникальными преимуществами и определенными проблемами в обеспечении интерпретируемости моделей ИИ в различных предметных областях. Результаты исследования позволили сформулировать ряд важных выводов относительно сильных и слабых сторон каждого метода.

Выводом исследования является отсутствие универсального метода интерпретации, в полной мере удовлетворяющего всем критериям. Методы существенно различаются по типу предоставляемых объяснений и по своему назначению.

Научная новизна исследования заключается в комплексной систематизации методов интерпретации ИИ на основе единого набора критериев, охватывающих как теоретические, так и практические аспекты интерпретируемости.

Результаты анализа подтверждают, что интерпретируемость не может быть сведена исключительно к локальным или глобальным объяснениям, а должна рассматриваться как совокупность свойств, включающих устойчивость, воспроизводимость и когнитивную доступность объяснений для ЛПР.

Практическая значимость работы определяется возможностью использования полученных выводов при проектировании и внедрении ИИ-систем в реальных процессах принятия решений. Результаты проведенного в статье сравнительного анализа методов могут служить основой для специалистов, выбирающих инструменты интерпретации моделей ИИ в различных предметных областях.

Перспективным направлением дальнейших исследований является применение гибридных подходов, сочетающих преимущества различных методов, исследование влияния интерпретации на фактическое качество принимаемых решений, адаптация методов интерпретируемости к высокоразмерным и мультимодальным данным.

Обеспечение интерпретируемости моделей и алгоритмов ИИ является необходимым условием их практического применения в различных сферах деятельности и интеграции в процессы принятия решений, способствует повышению прозрачности алгоритмов, укреплению доверия со стороны пользователей и экспертов, а также снижению рисков, связанных с некорректными или предвзятыми решениями.

Список литературы

1. Бородушко И. В., Матвеев А. В. Современные тенденции и стратегические цели развития искусственного интеллекта в Российской Федерации // Национальная безопасность и стратегическое планирование. 2024. № 2 (46). С. 66–74. DOI: 10.37468/2307-1400-2024-2-66-74. EDN: EFWHYT.

2. Минуллин Д. А., Гафаров Ф. М. Анализ моделей машинного обучения на основе методов объяснимого искусственного интеллекта в образовательной аналитике // Электронные библиотеки. 2024. Т. 27. № 3. С. 294–315. DOI: 10.26907/1562-5419-2024-27-3-294-315. EDN: XFIGCX.

3. Путихин Ю. Е., Буняк В. Л., Матвеев В. В. Внедрение искусственного интеллекта в экономику России // Национальная безопасность и стратегическое планирование. 2025. № 1 (49). С. 31–46. DOI: 10.37468/2307-1400-2025-1-31-46. EDN: MCCNPG.

4. Бирюков Д. Н., Дудкин А. С. Объяснимость и интерпретируемость – важные аспекты безопасности решений, принимаемых интеллектуальными системами (обзорная статья) // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25. № 3. С. 373–386. DOI: 10.17586/2226-1494-2025-25-3-373-386. EDN: NHHVUJ.

5. Макаренко А. В. Глубокие нейронные сети: зарождение, становление, современное состояние // Проблемы управления. 2020. № 2. С. 3–19. DOI: 10.25728/pu.2020.2.1. EDN: DBWHHL.

6. Alangari N., El Bachir Menai M., Mathkour H., Almosallam I. Exploring evaluation methods for interpretable machine learning: A survey // Information. 2023. Vol. 14. Is. 8. P. 469. DOI: 10.3390/info14080469.

7. Carvalho D. V., Pereira E. M., Cardoso J. S. Machine learning interpretability: A survey on methods and metrics // Electronics. 2019. Vol. 8. Is. 8. P. 832. DOI: 10.3390/electronics8080832.

8. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A review of machine learning interpretability methods // Entropy. 2021. Vol. 23. Is. 1. P. 18. DOI: 10.3390/e23010018.

9. Шевская Н. В. Объяснимый искусственный интеллект и методы интерпретации результатов // Моделирование, оптимизация и информационные технологии. 2021. Т. 9. № 2 (33). DOI: 10.26102/2310-6018/2021.33.2.024. EDN: VRKUIL.

10. Wang Y. A comparative analysis of model agnostic techniques for explainable artificial intelligence // Research Reports on Computer Science. 2024. P. 25–33. DOI: 10.37256/rrcs.3220244750.

11. Zafar M. R., Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability // Machine Learning and Knowledge Extraction. 2021. Vol. 3. Is. 3. P. 525–541. DOI: 10.3390/make3030027.

12. Visani G., Bagli E., Chesani F., Poluzzi A., Capuzzo D. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models //Journal of the Operational Research Society. 2022. Vol. 73. Is. 1. P. 91–101. DOI: 10.1080/01605682.2020.1865846.

13. Nohara Y., Matsumoto K., Soejima H., Nakashima N. Explanation of machine learning models using improved Shapley additive explanation // Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. 2019. P. 546–546. DOI: 10.1145/3307339.3343255.

14. Pelegrina G. D., Couceiro M., Duarte L. T. A preprocessing Shapley value-based approach to detect relevant and disparity prone features in machine learning // Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024. P. 279–289. DOI: 10.1145/3630106.3658905.

15. Zhang J., Sun Q., Liu J., Xiong L., Pei J., Ren K. Efficient sampling approaches to shapley value approximation // Proceedings of the ACM on Management of Data. 2023. Vol. 1. Is. 1. P. 1–24. DOI: 10.1145/3588728.

16. Witter R. T., Liu Y., Musco C. Regression-adjusted Monte Carlo Estimators for Shapley Values and Probabilistic Values. 2025. DOI: 10.48550/arXiv.2506.11849.

17. Covert I., Lee S. I. Improving kernelshap: Practical Shapley value estimation using linear regression // International conference on artificial intelligence and statistics. PMLR, 2021. P. 3457–3465. URL: http://proceedings.mlr.press/v130/covert21a/covert21a.pdf (дата обращения: 10.01.2026).

18. Aydoğan B., Aytekin T. An in-depth analysis of KernelSHAP and SamplingSHAP: assessing robustness, error, and efficiency // Knowledge and Information Systems. 2025. Vol. 67. Is. 11. P. 10545–10579. DOI: 10.1007/s10115-025-02541-z.

19. Molnar C., Freiesleben T., König G., Herbinger J., Reisinger T., Casalicchio G., Wright M. N., Bischl B. Relating the partial dependence plot and permutation feature importance to the data generating process // World Conference on Explainable Artificial Intelligence. Cham: Springer Nature Switzerland, 2023. P. 456–479. DOI: 10.1007/978-3-031-44064-9_24.

20. Moosbauer J., Herbinger J., Casalicchio G., Lindauer M., Bischl B. Explaining hyperparameter optimization via partial dependence plots // Advances in neural information processing systems. 2021. Vol. 34. P. 2280–2291. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/12ced2db6f0193dda91ba86224ea1cd8-Paper.pdf (дата обращения: 10.01.2026).

21. Kerrigan D., Barr B., Bertini E. PDPilot: Exploring Partial Dependence Plots Through Ranking, Filtering, and Clustering // IEEE Transactions on Visualization and Computer Graphics. 2025. Vol. 31. Is. 10. P. 7377–7390. DOI: 10.1109/TVCG.2025.3545025.

22. Wright R. Interpreting black-box machine learning models using partial dependence and individual conditional expectation plots // Exploring SAS® Enterprise Miner Special Collection. 2018. Vol. 1950. URL: https://sites.dartmouth.edu/dasug/files/2018/12/RayWright1950-2018.pdf (дата обращения: 10.01.2026).

23. Yeh A., Ngo A. Bringing a ruler into the black box: uncovering feature impact from individual conditional expectation plots // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2021. P. 34–48. DOI: 10.1007/978-3-030-93736-2_4.

24. Ribeiro M. T., Singh S., Guestrin C. Anchors: High-precision model-agnostic explanations // Proceedings of the AAAI conference on artificial intelligence. 2018. Vol. 32. Is. 1. DOI: 10.1609/aaai.v32i1.11491.

25. Ratul Q. E. A., Serra E., Cuzzocrea A. Evaluating attribution methods in machine learning interpretability // 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021. P. 5239–5245. DOI: 10.1109/BigData52589.2021.9671501.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.