УДК 004.8

МЕТОДЫ ОЦЕНКИ КАЧЕСТВА СИСТЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ: ЭВОЛЮЦИЯ ПОДХОДОВ, ТЕНДЕНЦИИ И ОГРАНИЧЕНИЯ

METHODS FOR EVALUATING THE QUALITY OF COMPUTER VISION SYSTEMS: EVOLUTION OF APPROACHES, TRENDS, AND LIMITATIONS

Шарова Д.Е., начальник управления, Департамент информационных технологий города Москвы, г. Москва

Гарбук С.В., канд. техн. наук, директор, ФГБУН ВИНИТИ РАН, г. Москва

Sharova D.E., Head of department, Department of Information Technologies of the City of Moscow, Moscow, Russia

Garbuk S.V., Ph.D. (Engineering), Director, All-Russian Institute of Scientific and Technical Information of the Russian Academy of Sciences (VINITI RAS), Moscow, Russia

Рост применения систем компьютерного зрения на базе искусственного интеллекта сопровождается расширением спектра методов оценки их качества. При этом существующие подходы существенно различаются по охватываемым аспектам и часто не обеспечивают полноты представления о работе системы в реальных условиях. В статье представлен обзор основных направлений оценки качества: классических статистических метрик (точность, чувствительность, площадь под ROC-кривой), показателей устойчивости и обобщаемости, эксплуатационных характеристик, а также методов анализа доверия, объяснимости и взаимодействия с пользователем.

Особое внимание уделено роли существенных факторов эксплуатации (СФЭ) – изменяемых условий применения систем (тип данных, оборудование, сценарии использования, структура пользовательских действий), влияющих на воспроизводимость метрик и достоверность испытаний. Показано, что даже высокие значения точности или площади под ROC-кривой не гарантируют качества в эксплуатации при изменении распределения данных, условий съемки или параметров подготовки изображений.

Отмечается, что современные метрики и показатели качества охватывают точность, устойчивость к дрейфу данных, интерпретируемость результатов, доверие пользователя, скорость и предсказуемость работы системы. Однако их применение остается фрагментарным, а нормативно обоснованных комплексных процедур, учитывающих вариативность СФЭ, по-прежнему недостаточно.

Обзор позволяет оценить текущее состояние подходов к контролю качества систем компьютерного зрения, выявляет ограничения существующих методов и определяет направления дальнейшего развития – усиление роли эксплуатационных показателей, повышение требований к репрезентативности тестов, интеграция показателей доверия и переход к постоянному мониторингу качества.

Ключевые слова: искусственный интеллект, компьютерное зрение, оценка качества, показатели качества, информационные процессы.

Для цитирования: Шарова Д.Е, Гарбук С.В. Методы оценки качества систем компьютерного зрения: эволюция подходов, тенденции и ограничения // Информационно-экономические аспекты стандартизации и технического регулирования. 2026. № 1(88). С. 35–41.

The growing use of computer vision systems based on artificial intelligence has led to the emergence of diverse approaches for evaluating their quality. However, these methods differ significantly in scope and often fail to reflect the full spectrum of system behavior under real-world conditions. This article provides a structured review of the main groups of quality assessment techniques, including classical statistical metrics (accuracy, precision, recall, ROC-AUC, MCC), robustness and generalization indicators, operational performance characteristics, as well as approaches addressing explainability, user interaction, and trustworthiness.

A particular focus is placed on significant operational factors – variable conditions of system use such as data source, acquisition parameters, equipment, and user workflow – that substantially affect the reproducibility of metrics and the reliability of testing. The analysis shows that high accuracy or AUC scores do not guarantee stable performance when data distributions shift or when operating conditions deviate from those used during model development.

The review highlights that contemporary quality indicators encompass accuracy, robustness to data drift, interpretability of results, user trust, and operational predictability. Nevertheless, their application remains fragmented, and comprehensive, regulation-ready evaluation procedures that account for SOF variability are still lacking. The findings outline current limitations of existing assessment methods and identify key directions for future development, including improved operational metrics, higher requirements for test representativeness, and the transition toward continuous quality monitoring.

Keywords: computer vision; quality assessment; significant operational factors; metrics; quality indicators.

For citation: Sharova D.E., Garbuk S.V. Methods For Evaluating The Quality Of Computer Vision Systems: Evolution Of Approaches, Trends, And Limitations. Information and Economic Aspects of Standardization and Technical Regulation. 2026; 1(88): 35–41. (In Russ.).

ВВЕДЕНИЕ

Системы с технологией искусственного интеллекта (ИИ), основанные на методах компьютерного зрения, в последние годы получили широкое распространение в самых разных областях – от промышленности и транспорта до медицины и безопасности. Их внедрение сопровождается ростом требований к объективной и воспроизводимой оценке качества. Для того чтобы обеспечить доверие пользователей и возможность нормативного регулирования, необходимы методы оценки, позволяющие не только фиксировать показатели точности, но и учитывать устойчивость, интерпретируемость, эксплуатационные характеристики и изменения качества в динамике.

Существующие подходы к оценке качества систем с компьютерным зрением можно условно разделить на несколько групп. Классические метрики (точность, чувствительность, специфичность и др.) широко применяются для сопоставления алгоритмов, однако зависят от выборки и не отражают эксплуатационных условий. Комплексные индексы, такие как сбалансированная точность (Balanced Accuracy), коэффициент корреляции Мэттьюса (Matthews Correlation Coefficient), коэффициент согласия Коэна (Cohen's Kappa) и др., позволяют интегрировать несколько характеристик в единый показатель, но содержат элемент субъективности и могут скрывать слабые стороны моделей. Методы устойчивости и обобщаемости направлены на проверку качества в условиях изменяющихся данных, включая стресс-тестирование и анализ дрейфа распределений. Эксплуатационно-ориентированные методы оценивают влияние систем ИИ на пользователей и процессы, включая показатели совместной работы «человек–машина» и экономическую эффективность. Методы объяснимости и доверия повышают прозрачность работы алгоритмов и влияют на принятие решений пользователями. Наконец, динамические подходы позволяют отслеживать деградацию качества во времени, но, как правило, ограничиваются анализом распределений входных данных и стандартных метрик точности.

Таким образом, несмотря на значительный прогресс в развитии методов оценки, существующие подходы остаются фрагментарными и охватывают преимущественно отдельные аспекты функционирования систем компьютерного зрения. Различия в методологических основаниях, зависимости от характеристик тестовых данных, отсутствие единых требований к репрезентативности, учету существенных факторов эксплуатации и интерпретации результатов оценки приводят к тому, что выводы о качестве систем нередко оказываются несопоставимыми и не отражают реальных условий применения. Это затрудняет как практическое внедрение ИИ-технологий, так и формирование нормативно обоснованных процедур контроля качества.

В этой связи возникает необходимость в систематизации существующих методов и формировании целостного представления об их принципах, ограничениях и взаимосвязях. Целью настоящей работы является проведение структурированного обзора подходов к оценке качества систем компьютерного зрения, выявление их сильных и слабых сторон, а также определение ключевых тенденций развития области. Сделанное обобщение позволит определить, какие аспекты оценки остаются недостаточно проработанными, и обозначить направления, требующие дальнейших исследований и методологического развития.

ОБЗОР ПОДХОДОВ К ОЦЕНКЕ КАЧЕСТВА СИСТЕМ С ИИ

Существующие подходы к количественной оценке качества систем компьютерного зрения разнообразны и отражают разные аспекты их функционирования. Для удобства анализа их можно классифицировать по шести основным направлениям, начиная с наиболее традиционных – классических статистических метрик, и заканчивая современными методами динамического мониторинга.

1. Классические метрики

Наиболее широко применяемым подходом к количественной оценке качества систем с технологией искусственного интеллекта являются классические статистические метрики точности.

Ключевыми метриками считаются точность (accuracy, доля правильных предсказаний среди всех случаев), чувствительность (sensitivity/recall – способность корректно выявлять положительные объекты), специфичность (specificity – способность корректно определять отрицательные объекты), точность положительных классификаций (precision), прогностическая ценность положительного результата (positive predictive value) и их производные, такие как F1-мера (F1-score), объединяющие точность положительных предсказаний и чусвтительность в единый показатель [1, 2]. Для задач бинарной классификации метрики удобно представлять через матрицу ошибок (confusion matrix), которая позволяет детализировать баланс между ложноположительными и ложноотрицательными результатами.

В задачах компьютерного зрения площадь под ROC-кривой и площадь под кривой точности-полноты (PR-AUC) также получили широкое распространение как интегральные показатели качества классификаторов. ROC-AUC измеряет вероятность того, что модель присвоит более высокий балл случайному положительному примеру по сравнению со случайным отрицательным, тогда как PR-AUC более чувствителен к дисбалансу классов и полезен при работе с редкими событиями [3].

Несмотря на широкое применение, классические метрики имеют ограниченную информативность. Во-первых, они зависят от используемого тестового набора данных и не гарантируют воспроизводимость результатов на других выборках. Во-вторых, агрегированные показатели могут маскировать слабые места модели: например, высокая точность при крайне несбалансированных классах не отражает реального качества. Наконец, они не учитывают эксплуатационные аспекты, такие как скорость работы, устойчивость к шумам или удобство использования в конкретной прикладной задаче.

Таким образом, классические диагностические метрики являются необходимым, но не достаточным инструментом для оценки систем компьютерного зрения. Они позволяют проводить сравнение алгоритмов на единых тестовых наборах и служат базовой основой для сертификации и академических публикаций, однако требуют дополнения более комплексными методами, учитывающими устойчивость, интерпретируемость и динамику работы моделей.

2. Комплексные индексы качества

Классические метрики обладают высокой наглядностью, но нередко демонстрируют ограниченную информативность при анализе сложных или несбалансированных наборов данных. В ответ на эти ограничения были разработаны комплексные индексы, которые интегрируют несколько характеристик модели в единый показатель. Такие индексы позволяют более адекватно отражать общую эффективность алгоритма и упрощают процедуру сравнения между различными системами компьютерного зрения [2, 4, 5].

Одним из наиболее часто используемых показателей является сбалансированная точность (Balanced Accuracy), определяемая как среднее арифметическое чувствительности и специфичности. Он особенно полезен в задачах с сильным дисбалансом классов, где общая точность может вводить в заблуждение. Другой популярный индекс – коэффициент корреляции Мэттьюса (MCC), который вычисляется на основе всех элементов матрицы ошибок. MCC принимает значения от –1 до +1, где +1 соответствует идеальному предсказанию, 0 – случайной классификации, а –1 – полной обратной зависимости. В ряде исследований показано, что MCC является более надежным интегральным индикатором по сравнению с F1-мерой или точностью при работе с асимметричными классами [6].

К числу современных обобщенных индексов можно отнести коэффициент согласия Коэна и его модификации, которые учитывают вероятность случайного совпадения предсказаний и истинных меток. Данный показатель активно используется при сравнении согласованности между несколькими алгоритмами или при сопоставлении модели с экспертной оценкой. Развитие этого направления привело к появлению многомерных коэффициентов согласия (например, каппа Флейса – Fleiss' Kappa), применимых в задачах многоклассовой классификации [7].

В последние годы также активно обсуждаются новые способы построения композитных метрик, включая взвешенные индексы, где вес каждого критерия определяется исходя из важности конкретного аспекта задачи. Например, для промышленных систем компьютерного зрения на производстве может быть важнее минимизация ложноположительных случаев, тогда как в системах безопасности – снижение числа ложных пропусков. В таких условиях классические метрики объединяются в единый показатель с заранее заданными или экспертно определенными весами [8].

Следует отметить, что комплексные индексы существенно повышают удобство сравнения моделей и их ранжирования, особенно в условиях конкурирующих требований. Однако они сохраняют определенную долю субъективности при выборе весов и приоритизации критериев. Кроме того, интегральные показатели могут скрывать важные детали работы модели и ограничивать возможность глубокого анализа слабых мест алгоритма [4].

В целом комплексные индексы качества можно рассматривать как следующий шаг после базовых метрик, обеспечивающий более сбалансированное и справедливое сравнение алгоритмов. Тем не менее для получения полноценной картины их необходимо дополнять другими подходами, ориентированными на устойчивость, эксплуатационные факторы и динамику работы систем.

3. Методы устойчивости и обобщаемости

Одним из ключевых направлений оценки качества систем компьютерного зрения является анализ их устойчивости и обобщающей способности. Под устойчивостью понимается способность алгоритма сохранять корректность предсказаний при изменении условий или внесении возмущений в данные, а под обобщаемостью – способность демонстрировать стабильное качество на новых, ранее не встречавшихся выборках [9, 10]. Эти свойства особенно важны в прикладных задачах, где условия эксплуатации могут существенно отличаться от лабораторных.

Классическим методом оценки обобщаемости является внешняя валидация (external validation) – тестирование алгоритма на данных, полностью отличных от обучающего набора. В отличие от обычной кросс-валидации этот подход позволяет выявить «переобучение на домен» и проверить переносимость модели на новые источники данных. Практикой в области компьютерного зрения стало проведение мультицентровых или мультидоменных тестов, включающих изображения, полученные на различ-

ных устройствах, в разных условиях освещенности или при изменении ракурса [11].

Дополнительным инструментом служат стресс-тесты, которые моделируют неблагоприятные условия эксплуатации. К ним относятся: добавление гауссовского шума, сжатие изображений, снижение разрешения, варьирование контрастности и яркости, а также имитация артефактов (например, размытий или перекрытий объектов). Эффективность алгоритма при таких искажениях позволяет судить о его робастности. В задачах безопасности все чаще применяются также адверсариальные атаки (adversarial attacks), проверяющие устойчивость к специально созданным возмущениям [12].

Важным направлением является мониторинг дрейфа данных (data drift), который возникает при изменении распределения входных данных в процессе эксплуатации. Для его количественной оценки применяются статистические метрики, в частности:

– Индекс стабильности популяции (Population Stability Index, PSI) – показатель, сравнивающий распределение признаков в текущем и базовом датасете.

– Дивергенция Кульбака–Лейблера (Kullback–Leibler (KL) divergence) – мера расхождения между двумя вероятностными распределениями.

– Дивергенция Дженсена–Шеннона (Jensen–Shannon (JS) divergence) – симметричная и более устойчивая модификация KL-дивергенции.

– Расстояние Васерштейна (Wasserstein distance) – метрика расстояния между распределениями, учитывающая «стоимость переноса массы».

Эти методы позволяют количественно фиксировать сдвиги в данных и связывать их с возможной деградацией качества алгоритма [13]. При этом важно, что сама по себе фиксация дрейфа не всегда означает ухудшение точности модели: требуется сопоставление с динамикой метрик качества.

При суммировании методы устойчивости и обобщаемости обеспечивают более глубокое понимание ограничений систем компьютерного зрения по сравнению с базовыми метриками точности. Их применение позволяет не только выявлять переобучение и скрытые уязвимости, но и организовывать долгосрочный мониторинг моделей в условиях изменяющихся данных. Однако эти подходы в большинстве случаев представляют собой точечные процедуры, требующие регулярного обновления, и не заменяют собой динамического анализа качества в реальной эксплуатации.

4. Эксплуатационно-ориентированные методы

В отличие от традиционных метрик точности, эксплуатационно-ориентированные методы направлены на количественную оценку того, как система искусственного интеллекта влияет на работу пользователей и эффективность процессов после внедрения. Они позволяют оценивать не только алгоритмическое качество, но и фактическую полезность в условиях реальной эксплуатации.

Метрики взаимодействия человек–машина. Важное направление составляет анализ работы системы в условиях совместного использования с человеком (human-in-the-loop). Для оценки применяются показатели: время до принятия решения, доля исправленных прогнозов, частота принятия или отклонения рекомендаций модели, изменение итогового качества решений при разных режимах («только человек», «только система», «человек+система»). Эти показатели позволяют количественно измерять вклад алгоритма в совместный результат [14].

Показатели практической полезности. Для оценки того, как система влияет на конечные результаты после развертывания, используются специальные метрики: частота ошибок при эксплуатации, количество предотвращенных ложных решений, изменение маршрутизации объектов или задач, распределение нагрузки между уровнями системы. Такой анализ позволяет сопоставить формальные показатели точности с реальной результативностью в прикладной задаче [15,16].

Нагрузка и удобство использования. Существенным элементом оценки является измерение влияния системы на рабочую нагрузку пользователей. Для этого фиксируют время взаимодействия с интерфейсом, количество действий, необходимых для получения результата, субъективную когнитивную нагрузку, частоту доверия и отказа от рекомендаций. Важно, чтобы система снижала когнитивную и временную нагрузку, а не создавалась новая форма «перегрузки» пользователя [17, 18].

Экономическая эффективность. Для управленческих решений применяются методы анализа «затраты – эффективность» (Cost-Effectiveness Analysis, CEA), «затраты – полезность» (Cost-Utility Analysis, CUA) и анализа влияния на бюджет (Budget Impact Analysis, BIA). Рассчитываются показатели стоимости корректного решения, стоимости предотвращенной ошибки, влияние внедрения на общий бюджет и рентабельность инвестиций. Такие методы позволяют сопоставлять качество алгоритма с его экономическим эффектом [19, 20].

Инфраструктура мониторинга. Для обеспечения объективной оценки эксплуатационных характеристик требуется постпроектный мониторинг: фиксация решений системы

и пользователя, учет версий модели, сбор информации о динамике ошибок, а также регулярный пересчет метрик качества и устойчивости. Все это обеспечивает возможность сопоставлять формальные показатели с эксплуатационными эффектами и своевременно выявлять деградацию качества [21].

Таким образом, эксплуатационно-ориентированные методы позволяют оценивать не только алгоритм как таковой, но и его влияние на процессы и пользователей. Они обеспечивают количественную фиксацию совместной эффективности человека и системы, полезности для конечной задачи, удобства использования и экономического эффекта. Их сильной стороной является прямая связь с практическими целями, а ограничением – необходимость организации сбора данных в условиях эксплуатации и проведения сравнительных экспериментов.

5. Методы объяснимости и доверия

Для широкого внедрения систем компьютерного зрения важны не только показатели точности, но и прозрачность их работы. Методы объяснимости и доверия направлены на то, чтобы сделать выводы модели понятными для человека и снизить риск некорректной интерпретации или переоценки возможностей алгоритма.

Формальные показатели объяснимости. Наиболее распространенным направлением является оценка согласованности между объяснением и реальным механизмом работы модели. Для этого используются согласованность (fidelity) и устойчивость (stability) объяснения работы модели. Применяются как локальные методы интерпретации (например, LIME, SHAP), так и глобальные, позволяющие выявить важность признаков во всей выборке [22,23].

Метрики понятности для пользователя. Отдельный класс методов связан с измерением того, насколько объяснение воспринимается человеком как понятное и полезное. Для оценки применяются опросники, измерение времени принятия решения с использованием объяснения, частота ошибок и количество отклоненных рекомендаций. Такой подход позволяет связать интерпретируемость с конечным результатом взаимодействия человек–система [24].

Оценка доверия и риска переиспользования. Методы доверия включают измерение уровня согласия пользователя с рекомендациями системы, частоты слепого следования выводам модели, а также случаев отказа от ее использования. При этом важно фиксировать не только общий уровень доверия, но и его баланс: чрезмерное доверие может быть столь же опасным, как и избыточное недоверие. Для количественной оценки предлагается использовать частоту принятия решений, согласованных с системой, а также анализ корреляции между уверенностью модели и доверием человека [17].

Интеграция в процедуры оценки. Методы объяснимости и доверия применяются в комплексе с другими группами метрик. Они позволяют выявить, насколько система понятна конечному пользователю и как именно интерпретируемость влияет на готовность применять ее в реальной задаче. В ряде работ отмечается, что даже высокая точность алгоритма не гарантирует его принятие пользователями, если объяснения остаются непонятными или избыточными [25].

В результате методы объяснимости и доверия обеспечивают дополнительный уровень контроля качества систем компьютерного зрения. Они позволяют количественно фиксировать прозрачность работы алгоритмов, устойчивость объяснений, восприятие их человеком и уровень доверия. Эти методы не заменяют метрик точности или эксплуатационных показателей, но усиливают их, обеспечивая основу для безопасного и обоснованного внедрения.

6. Динамические методы оценки качества

Большинство традиционных подходов к оценке качества систем компьютерного зрения дают статичную картину, ограничиваясь тестированием на фиксированном наборе данных. Однако в реальной эксплуатации распределения входных данных и условия применения меняются во времени, что приводит к необходимости разработки динамических методов оценки.

Оценка качества во времени. Одним из направлений является анализ изменений диагностических метрик в процессе эксплуатации. Показатели чувствительности, специфичности или точности рассчитываются не только по итогам одного теста, но и регулярно на новых потоках данных, что позволяет строить кривые деградации или улучшения качества (performance over time). Такой мониторинг дает возможность своевременно выявлять снижение эффективности модели и планировать ее обновление [21].

Рассмотренные ранее *методы обнаружения дрейфа* (PSI, KL-дивергенция, JS-дивергенция, расстояние Васерштейна) по своей сути являются динамической оценкой и используются для регулярного сравнения текущих данных с эталонным распределением и количественной фиксации степени дрейфа, что позволяет прогнозировать снижение качества модели [13, 26].

Непрерывное обучение и катастрофическое забывание. При обновлении моделей возникает риск «катастрофического забывания», когда точность на новых данных растет, но снижается на старых. Для его контроля применяются метрики, сопоставляющие качество на исторических и новых выборках, а также коэффициенты сохранения знаний (retention

metrics). Таким образом, динамические методы позволяют оценивать не только текущее качество, но и способность системы сохранять накопленный опыт [27].

Динамические методы создают основу для долгосрочного мониторинга систем компьютерного зрения. В отличие от статических показателей они учитывают изменчивость входных данных и эксплуатационных условий. Их сильной стороной является способность фиксировать деградацию или улучшение качества во времени, однако реализация требует развертывания инфраструктуры для сбора и анализа данных.

ВЫВОДЫ И ПОСТАНОВКА ЗАДАЧИ РАЗРАБОТКИ МЕТОДА ОЦЕНКИ КАЧЕСТВА

Проведенный обзор показывает, что существующие подходы к оценке систем с технологией компьютерного зрения охватывают широкий спектр задач: от базовых диагностических метрик до эксплуатационных показателей и методов динамического мониторинга. Классические точность, чувствительность и специфичность обеспечивают базовый уровень сопоставимости алгоритмов, но не отражают особенностей их работы в реальной среде. Комплексные индексы позволяют интегрировать несколько характеристик в единый показатель, однако сохраняют субъективность при выборе весов и часто скрывают слабые стороны моделей. Методы устойчивости и обобщаемости выявляют зависимость качества от условий и изменений данных, но в основном дают точечную картину. Эксплуатационные показатели фокусируются на взаимодействии человека и системы, а также на экономической и организационной эффективности, однако зависят от конкретных сценариев применения и трудны для стандартизации. Методы объяснимости и доверия добавляют прозрачность и повышают приемлемость систем, но также не могут быть использованы как единственный критерий. Наконец, динамические подходы позволяют фиксировать деградацию качества во времени, но обычно ограничены анализом распределений и метрик точности, без учета комплексных факторов эксплуатации.

ЗАКЛЮЧЕНИЕ

Таким образом, в современной литературе отсутствует универсальный метод, который позволял бы одновременно:

– учитывать многомерность качества систем компьютерного зрения (точность, устойчивость, интерпретируемость, влияние на пользователя, экономический эффект);

– фиксировать изменения этих характеристик в динамике при реальной эксплуатации;

– обеспечивать количественное сопоставление различных систем и их версий на основе интегрального показателя.

Отдельно стоит отметить, что лишь ограниченное число исследований, в том числе отечественные работы, посвященные рискам жизненного цикла и эксплуатационным характеристикам интеллектуальных систем [28, 29], подчеркивают необходимость явного учета существенных факторов эксплуатации (СФЭ). Эти факторы определяют вариативность условий применения систем: тип и качество данных, особенности оборудования, параметры среды, специфику пользовательских действий и организационный контекст. Именно они формируют реальную «зону применимости» моделей и определяют репрезентативность испытаний. Игнорирование СФЭ приводит к тому, что формальные показатели качества, полученные в лабораторной среде, плохо коррелируют с эксплуатационными результатами, а тестовые наборы оказываются недостаточно показательными.

Анализ литературы демонстрирует, что отсутствие формализованного подхода к описанию и использованию СФЭ является одним из ключевых пробелов существующих методик оценки. Это препятствует стандартизации процедур контроля качества [29], затрудняет сопоставимость результатов различных испытаний и снижает предсказуемость поведения систем в реальных условиях.

Суммарно это определяет необходимость разработки методики, которая позволила бы интегрировать диагностические метрики, показатели устойчивости, эксплуатационные характеристики, факторы доверия и динамические аспекты в единую структуру оценки качества с обязательным учетом существенных факторов эксплуатации как основы репрезентативности тестирования и интерпретации результатов. Именно такая методика позволит обеспечить как научную сопоставимость, так и практическую применимость оценки качества систем компьютерного зрения.

Список литературы / References

1. Powers D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation // Journal of Machine Learning Technologies. 2011. Vol. 2, no. 1. P. 37–63.
2. Chicco D., Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation // BMC Genomics. 2020. Vol. 21, no. 1. P. 6.
3. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. 2006. Vol. 27, no. 8. P. 861–874.
4. Garbuk S. V. Intellimetry as a Way to Ensure AI Trustworthiness // 2018 International Conference on Artificial Intelligence Applications and Innovations. IC-AIAI. 2018. P. 27–30.

5. Brodersen K.H., Ong C.S., Stephan K.E., Buhmann J.M. The balanced accuracy and its posterior distribution // In 2010 20th international conference on pattern recognition. IEEE. 2010. P. 3121–3124.
6. Chicco D., Tötsch N., Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation // BioData Mining. 2021. Vol. 14, no. 1. P. 13.
7. Warrens M. J. Cohen's kappa is a weighted average // Statistical Methodology. 2010. Vol. 8, no. 6. P. 473–484.
8. Grandini M., Bagli E., Visani G. Metrics for Multi-Class Classification: an Overview. 2020. arXiv: 2008.05756.
9. Recht B., Roelofs R., Schmidt L., Shankar V. Do ImageNet Classifiers Generalize to ImageNet // Proceedings of the 36th International Conference on Machine Learning. 2019. Vol. 97. P. 5389–5400.
10. Taori R., Dave A., Shankar V. et al. Measuring Robustness to Natural Distribution Shifts in Image Classification // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 18583–18599.
11. Gulrajani I., Lopez-Paz D. In Search of Lost Domain Generalization. 2020. arXiv:2007.01434v1
12. Goodfellow I.J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. 2014. arXiv:1412.6572v3
13. Lu J., Liu A., Dong F. et. al. Learning under Concept Drift: A Review // Transactions on Knowledge and Data Engineering. IEEE. 2019. Vol. 31, no. 12. P. 2346–2363.
14. Amershi S., Weld D., Vorvoreanu M. et al. Guidelines for Human-AI Interaction // CHI Conference on Human Factors in Computing Systems. 2019. No. 3. P. 1–13.
15. Sendak M., D'Arcy J., Kashyap S. et al. A Path for Translation of Machine Learning Products into Healthcare Delivery // EMJ Innovations. 2020.
16. Holstein K., Wortman Vaughan J., Daumé H. et al. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? // CHI Conference on Human Factors in Computing Systems. 2019. No. 600. P. 1–16.
17. Buçinca Z., Malaya M., Glassman E. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems // Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 2020. P. 454–464.
18. Lai V., Tan C. On Human Predictions with Explanations and Predictions without Explanations // Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. P. 29–38.
19. El Arab R.A., Al Moosa O.A. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare // NPJ Digital Medicine. 2025. Vol. 8, no.1. P. 548.
20. Enholm I.M., Papagiannidis E., Mikalef P. et al. Artificial Intelligence and Business Value: a Literature Review // Information Systems Frontires. 2022. Vol. 24. P. 1709–1734.
21. Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters // Annals of Translational Medicine. 2020. Vol. 8, no. 7. P. 501.
22. Ribeiro M., Singh S., Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier // KDD Conference Proceedings. 2016. P. 1135–1144.
23. Lundberg S., Lee S. A Unified Approach to Interpreting Model Predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. P. 4768–4777.
24. Doshi-Velez F., Kim B. Towards A Rigorous Science of Interpretable Machine Learning. 2017. arXiv: 1702.08608.
25. Poursabzi-Sangdeh F., Goldstein D., Hofman J. et.al. Manipulating and Measuring Model Interpretability // CHI Conference on Human Factors in Computing Systems. 2018. Art. no. 237. P. 1–52.
26. Webb G.I., Hyde R., Cao H. et.al. Characterizing concept drift // Data Mining and Knowledge Discovery. 2016. Vol. 30, no. 4. P. 964–994.
27. De Lange M., Aljundi R., Masana M. et.al. A continual learning survey: Defying forgetting in classification tasks // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021. Vol. 44, no. 7. P. 3366–3385.
28. Гарбук С.В. Качественный анализ рисков в жизненном цикле систем искусственного интеллекта // В кн. Безопасность России. Правовые, социально-экономические и научно-технические аспекты. Тематический блок «Национальная безопасность». Системная инженерия в проблемах национальной безопасности. Научн. рук. чл.-корр. РАН Н.А. Махутов. С. 643–659. М.: МГОФ «Знание». 2025. 898 с. / Garbuk S.V. Qualitative risk analysis in the life cycle of artificial intelligence systems. In: Security of Russia. Legal, socio-economic and scientific-technical aspects. Thematic section "National Security". Systems engineering in national security problems. Scientific editor: N.A. Makhutov. Moscow: MGOF "Znanie". 2025. PP. 643–659. 898 p. (in Russian)
29. Гарбук С.В. Метод оценки влияния параметров стандартизации на эффективность создания и применения систем искусственного интеллекта // Информационно-экономические аспекты стандартизации и технического регулирования. 2022. № 3(67). С. 4–14. / Garbuk S.V. A method for assessing the impact of standardization parameters on the effectiveness of creating and applying artificial intelligence systems // Information and economic aspects of standardization and technical regulation. 2022. Vol. 3, no. 67. P. 4–14.