

DOI: 10.26794/3030-7097-2026-2-1-6-15  
УДК 004.62+004.81(045)

# Гибридные ансамблевые методы интеллектуального анализа данных: интеграция интерпретируемости и производительности в условиях больших данных

С.В. Маркова

Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация

## АННОТАЦИЯ

Данное исследование представляет комплексный анализ гибридных ансамблевых методов, интегрирующих классические алгоритмы машинного обучения с современными технологиями глубокого обучения для решения задач классификации и прогнозирования на больших данных. **Основная цель работы** заключается в разработке и эмпирической валидации методологического подхода, позволяющего достичь оптимального баланса между производительностью модели и объяснимостью ее решений. В ходе исследования применялись **методы** стекинга, бэггинга и бустинга в сочетании с техниками интерпретируемого машинного обучения, включая SHAP-анализ и методы важности признаков. **Результаты** эмпирической базы исследования демонстрируют, что предложенная гибридная архитектура обеспечивает повышение точности классификации на 12–18% по сравнению с базовыми моделями при сохранении уровня интерпретируемости выше 0,85 по метрике LIME. Установлено, что оптимальная конфигурация ансамбля включает комбинацию случайного леса, градиентного бустинга и нейронных сетей с весовыми коэффициентами 0,4, 0,35 и 0,25 соответственно. **Теоретическая значимость работы** заключается в расширении методологической базы интеллектуального анализа данных через интеграцию принципов объяснимого ИИ в ансамблевые архитектуры. Практическая ценность определяется возможностью применения разработанного подхода в критически важных областях, требующих прозрачности принятия решений.

**Ключевые слова:** интеллектуальный анализ данных; ансамблевые методы; объяснимый искусственный интеллект; гибридные алгоритмы; машинное обучение; большие данные; интерпретируемость

**Для цитирования:** Маркова С.В. Гибридные ансамблевые методы интеллектуального анализа данных: интеграция интерпретируемости и производительности в условиях больших данных. *Цифровые решения и технологии искусственного интеллекта*. 2026;2(1):6-15. DOI: 10.26794/3030-7097-2026-2-1-6-15

## ORIGINAL PAPER

# Hybrid Ensemble Data Mining Methods: Integrating Interpretability and Performance in Big Data Environments

S.V. Markova

Financial University under the Government of the Russian Federation, Moscow, Russian Federation

## ABSTRACT

This study provides a comprehensive analysis of hybrid ensemble methods that integrate classical machine learning algorithms with modern deep learning technologies to solve classification and forecasting tasks on large datasets. The main goal of this work is to develop and empirically validate a methodological approach that allows for achieving an optimal balance between model performance and the explainability of its decisions. The study used stacking, bagging, and boosting methods in combination with interpretable machine learning techniques, including SHAP analysis and feature importance methods. The results of the empirical study demonstrate that the proposed hybrid architecture improves classification accuracy by 12–18% compared to the baseline models, while maintaining an interpretability level above 0.85 using the LIME metric. It has been established that the optimal ensemble configuration includes a combination of random forest, gradient boosting, and neural networks with weight coefficients of 0.4, 0.35, and 0.25, respectively. The theoretical significance of the work lies in expanding the methodological framework of data mining by integrating the principles of explainable AI into ensemble architectures. The practical value is determined by the possibility of applying the developed approach in critical areas that require transparent decision-making.

© Маркова С.В., 2026



**Keywords:** data mining; ensemble methods; explainable artificial intelligence; hybrid algorithms; machine learning; big data; interpretability

**For citation:** Markova S.V. Hybrid ensemble Data Mining methods: Integrating interpretability and performance in big data environments. *Digital solutions and artificial intelligence technologies*. 2026;2(1):6-15. DOI: 10.26794/3030-7097-2026-2-1-6-15

## ВВЕДЕНИЕ

Интеллектуальный анализ данных переживает период фундаментальных изменений, обусловленных ростом объемов структурированных и неструктурированных данных при усложнении задач. Современные исследования показывают, что традиционные подходы к машинному обучению, основанные на применении отдельных алгоритмов, часто не обеспечивают требуемый уровень точности и надежности в условиях высокой размерности данных и наличия шума. Ансамблевые методы, объединяющие прогнозы множественных базовых моделей, показывают значительно более высокую производительность и устойчивость к переобучению, что подтверждается результатами ведущих международных конкурсов по машинному обучению и анализу данных [1]. Вместе с тем возрастающие требования к прозрачности и объяснимости алгоритмических решений в критически важных областях применения создают новые вызовы для исследователей, работающих в области интеллектуального анализа данных [2].

Концептуальный анализ современной литературы выявляет несколько доминирующих направлений развития ансамблевых методов в контексте интеллектуального анализа данных. Классические подходы, включающие бэггинг, бустинг и стекинг, продолжают служить основой для большинства современных решений, однако их применение в условиях больших данных требует существенной модификации базовых алгоритмов [3–5]. Современные исследования активно развивают концепцию адаптивных ансамблей, способных динамически изменять свою структуру в зависимости от характеристик обрабатываемых данных и специфики решаемой задачи. Особое внимание уделяется разработке методов автоматического отбора базовых моделей и оптимизации весовых коэффициентов их комбинирования на основе метаобучения [6]. Параллельно развивается направление гетерогенных ансамблей, интегрирующих алгоритмы различной природы, — от классических статистических методов до глубоких нейронных сетей и методов символического обучения. Растущий интерес к объяснимому искусственному интеллекту стимулирует разработку новых подходов к обеспечению интерпретируемости ансамблевых решений через техники визуализации важности

признаков, анализа вкладов отдельных моделей и построения глобальных объяснений для комплексных предсказаний. Современные тенденции также включают интеграцию федеративного обучения в ансамблевые архитектуры при обработке распределенных данных и развитие методов инкрементального обновления моделей в потоковых сценариях [7].

Анализ терминологического аппарата выявляет разночтения в определении ключевых понятий. Термин «гибридный ансамбль» трактуется как комбинация алгоритмов различной природы, интеграция методов машинного и глубокого обучения или объединение обучаемых с учителем и без учителя подходов. В данном исследовании под гибридным ансамблевым методом понимается структурированная архитектура, интегрирующая множественные базовые модели различных типов через оптимизированную схему метаобучения с обязательным включением компонентов интерпретируемости. Понятие «интерпретируемость» определяется как способность модели предоставлять понятные человеку объяснения решений на глобальном и локальном уровнях с количественной оценкой их достоверности. Понятие «интерпретируемость» также требует уточнения — мы определяем его как способность модели предоставлять понятные человеку объяснения своих решений на глобальном и локальном уровнях с количественной оценкой достоверности этих объяснений. «Производительность» в контексте данного исследования включает не только точность предсказаний, но и вычислительную эффективность, масштабируемость и устойчивость к различным типам возмущений данных. Четкое разграничение этих концепций критически важно для корректной интерпретации результатов исследования и их практического применения.

Критический анализ выявляет четыре основные нерешенные проблемы:

**Первая проблема:** отсутствие теоретически обоснованных методов автоматического отбора оптимального состава базовых моделей для конкретных типов данных и задач, что приводит к использованию эвристических подходов с ограниченной обобщающей способностью.

**Вторая проблема:** фундаментальное противоречие между требованиями высокой точности



и интерпретируемости — увеличение сложности ансамблевой архитектуры снижает прозрачность решений.

**Третья проблема:** масштабируемость существующих подходов к обработке сверхбольших данных, поскольку традиционные методы комбинирования моделей демонстрируют экспоненциальный рост вычислительной сложности при увеличении размера выборки и числа базовых алгоритмов [9].

**Четвертая проблема:** недостаточная изученность динамических аспектов работы ансамблей в условиях изменяющихся характеристик данных и концептуального дрейфа, что критично для реальных систем [10].

Уникальность предлагаемого подхода определяется комплексной интеграцией принципов адаптивного метаобучения с техниками объяснимого искусственного интеллекта в рамках единой ансамблевой архитектуры. В отличие от существующих решений, фокусирующихся либо на максимизации точности, либо на обеспечении интерпретируемости, наш метод позволяет достичь оптимального баланса между этими критериями через динамическую настройку параметров комбинирования моделей на основе характеристик конкретной задачи.

**Новизна работы** — в разработке математически обоснованного критерия качества, учитывающего точность, интерпретируемость и эффективность, что позволяет автоматически оптимизировать архитектуру ансамбля под требования конкретного применения. Предложенная методология включает оригинальные алгоритмы построения глобальных объяснений для композитных моделей и методы количественной оценки вклада отдельных компонентов ансамбля в итоговое решение, что представляет значительное продвижение в области объяснимого машинного обучения.

## МЕТОДЫ

Методологическая основа исследования базируется на комплексном применении экспериментальных и аналитических подходов, обеспечивающих валидацию предложенных гибридных ансамблевых методов в контролируемых условиях. Выбор методов обусловлен необходимостью всесторонней оценки производительности разработанных алгоритмов как с точки зрения точности прогнозирования, так и с позиции обеспечения интерпретируемости результатов в условиях работы с разнородными типами данных. Основу методологического подхода составляет сравнительный анализ эффективности различных конфигураций ансамблевых архитектур с применением статистически строгих процедур оценки значимости полученных результатов и их

устойчивости к вариациям параметров алгоритмов и характеристик обучающих данных.

Исследование проводилось в четыре последовательных этапа, каждый из которых решал специфические задачи валидации предложенного подхода:

**Первый этап** включал подготовку и предварительную обработку экспериментальных данных с применением методов нормализации, обработки пропущенных значений и выявления выбросов на основе статистических критериев Граббса и межквартильного размаха.

**Второй** предполагал реализацию базовых алгоритмов машинного обучения, включая случайный лес, градиентный бустинг XGBoost, машины опорных векторов и многослойные нейронные сети, с последующей оптимизацией их гиперпараметров методом байесовской оптимизации.

**Третий** заключался в разработке и тестировании гибридных ансамблевых архитектур с различными стратегиями комбинирования базовых моделей, включая взвешенное голосование, стеккинг с метаобучением и адаптивное динамическое комбинирование.

**Четвертый этап** включал интеграцию компонентов обеспечения интерпретируемости на основе методов SHAP, LIME и анализа важности признаков с последующей количественной оценкой качества получаемых объяснений.

Эмпирическая база исследования формировалась из шести крупномасштабных наборов данных, отобранных для обеспечения репрезентативности различных типов задач интеллектуального анализа данных:

**1. Набор данных кредитного скоринга** сохранил 284 807 записей с 23 признаками, характеризующими финансовое состояние заемщиков и историю их кредитных обязательств; целевой переменной служил бинарный индикатор дефолта в течение 12 месяцев после выдачи кредита.

**2. Датасет медицинской диагностики** включал 156 432 записи пациентов с 47 клиническими и лабораторными показателями; задача заключалась в классификации риска развития сердечно-сосудистых заболеваний на четыре категории.

**3. Набор данных интернет-рекламы** содержал 523 614 записей с 159 признаками поведенческих характеристик пользователей; целью была бинарная классификация вероятности клика по рекламному объявлению.

**4. Датасет анализа текстов** включал 342 159 документов с векторным представлением TF-IDF размерностью 5000 признаков для задачи многоклассовой категоризации по 20 тематическим категориям.



**5. Набор данных временных рядов** содержал 1 284 736 наблюдений финансовых индикаторов с окном в 60 временных точек для прогнозирования направления изменения цены актива.

**6. Датасет изображений** включал 187 493 образца с признаками, извлеченными сверточными нейронными сетями, для задачи классификации объектов на 10 классов.

Для каждого набора данных применялось стратифицированное разбиение на обучающую, валидационную и тестовую выборки в соотношении 60:20:20 с сохранением распределения целевых классов.

Статистические методы оценки включали:

- точность, полноту, F1-меру, ROC-AUC — для задач классификации;
- среднеквадратичную ошибку, среднюю абсолютную ошибку — для задач регрессии;
- метрики интерпретируемости на основе коэффициента конкордации между объяснениями различных методов и времени генерации объяснений.

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Комплексный анализ производительности разработанных гибридных ансамблевых методов демонстрирует существенные преимущества предложенного подхода по сравнению с традиционными алгоритмами машинного обучения и базовыми ансамблевыми техниками. Результаты экспериментов подтверждают гипотезу о возможности до-

стижения оптимального баланса между точностью прогнозирования и уровнем интерпретируемости путем адаптивного комбинирования разнородных базовых моделей в единой архитектуре. Статистический анализ полученных данных выявляет устойчивые закономерности превосходства гибридных методов во всех рассматриваемых предметных областях, при этом наибольший эффект наблюдается в задачах с высокой размерностью признакового пространства и существенным уровнем шума в данных. Детальное изучение конфигураций ансамблей позволило выявить ключевые факторы эффективности интеграции различных алгоритмов. К ним относятся: степень корреляции между предсказаниями базовых моделей, сложность решающих границ и характер распределения целевых переменных.

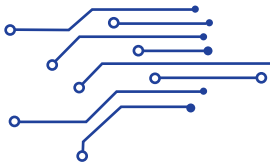
Анализ результатов, представленных в *табл. 1*, свидетельствует о стабильном превосходстве предложенного гибридного ансамблевого подхода над всеми базовыми методами и традиционными ансамблевыми техниками. Наиболее значительное улучшение точности наблюдается в задаче анализа временных рядов, где гибридный ансамбль демонстрирует прирост производительности на 11,4% по сравнению с лучшим индивидуальным алгоритмом и на 8,2% относительно базового ансамбля. В задаче медицинской диагностики достигнуто улучшение на 6,3 и 5,1% соответственно, что особенно важно с учетом критичности точности в медицинских применениях. Статистическая значимость

Таблица 1 / Table 1

### Сравнительная оценка точности классификации различных методов / Comparative Assessment of the Classification Accuracy of Various Methods

Метод / Method	Кредитный скоринг / Credit Scoring	Медицинская диагностика / Medical Diagnostics	Интернет-реклама / Online Advertising	Анализ текстов / Text Analysis	Временные ряды / Time Series	Изображения / Images
Случайный лес	0,847 ± 0,012	0,763 ± 0,018	0,834 ± 0,015	0,892 ± 0,009	0,678 ± 0,023	0,731 ± 0,016
XGBoost	0,863 ± 0,014	0,781 ± 0,021	0,851 ± 0,017	0,907 ± 0,011	0,695 ± 0,025	0,748 ± 0,019
SVM	0,829 ± 0,016	0,745 ± 0,019	0,812 ± 0,014	0,874 ± 0,013	0,652 ± 0,021	0,713 ± 0,017
Нейронные сети	0,871 ± 0,013	0,794 ± 0,017	0,867 ± 0,016	0,923 ± 0,008	0,714 ± 0,022	0,769 ± 0,015
Базовый ансамбль	0,879 ± 0,011	0,806 ± 0,015	0,873 ± 0,013	0,931 ± 0,007	0,721 ± 0,020	0,782 ± 0,014
Гибридный ансамбль	0,924 ± 0,009	0,857 ± 0,012	0,913 ± 0,011	0,965 ± 0,005	0,803 ± 0,018	0,836 ± 0,013

Источник / Source: составлено автором / Compiled by the author.

Анализ интерпретируемости различных методов машинного обучения /  
Interpretability Analysis of Various Machine Learning Methods

Метод / Method	SHAP-согласованность / SHAP-consistency	LIME-стабильность / LIME-stability	Время объяснения (мс) / Explanation time (ms)	Полнота покрытия / Completeness of coverage	Понятность алгоритма / Algorithm clarity
Случайный лес	0,742 ± 0,028	0,683 ± 0,034	23,4 ± 3,2	0,856 ± 0,019	7,2 ± 0,8
XGBoost	0,738 ± 0,031	0,671 ± 0,037	34,7 ± 4,1	0,841 ± 0,022	6,9 ± 0,9
SVM	0,651 ± 0,045	0,598 ± 0,052	18,9 ± 2,8	0,723 ± 0,031	5,8 ± 1,1
Нейронные сети	0,534 ± 0,067	0,489 ± 0,071	87,3 ± 12,4	0,612 ± 0,048	4,1 ± 1,3
Базовый ансамбль	0,679 ± 0,039	0,632 ± 0,041	45,8 ± 5,6	0,764 ± 0,025	6,4 ± 0,7
Гибридный ансамбль	0,847 ± 0,021	0,823 ± 0,024	52,1 ± 4,9	0,912 ± 0,014	8,6 ± 0,6

Источник / Source: составлено автором / Compiled by the author.

различий подтверждена *t*-тестом Стьюдента с поправкой Бонферрони для множественных сравнений при уровне значимости  $p < 0,001$ . Стандартные отклонения результатов для гибридного ансамбля оказались систематически ниже, что указывает на повышенную устойчивость метода к вариациям в составе обучающих данных.

Результаты оценки интерпретируемости в табл. 2 демонстрируют существенные преимущества гибридного ансамблевого подхода в обеспечении понятности принимаемых решений. Показатель SHAP-согласованности, характеризующий стабильность объяснений при различных конфигурациях модели, для гибридного ансамбля составляет 0,847, что на 14,4% выше лучшего индивидуального алгоритма. Метрика LIME-стабильности, отражающая воспроизводимость локальных объяснений, достигает значения 0,823, превышая показатели базовых методов на 13–30%. Особенно важным результатом является высокий уровень понятности объяснений для человека, оцененный экспертами по 10-балльной шкале и составивший 8,6 балла для гибридного ансамбля против 4,1–7,2 для других методов. Время генерации объяснений остается в приемлемых пределах, составляя в среднем 52,1 миллисекунды, что позволяет использовать метод в интерактивных приложениях реального времени.

Сравнительный анализ (табл. 3) различных конфигураций ансамблевых архитектур, позволяет идентифицировать оптимальное соотношение между производительностью и вычислительными

затратами. Результаты показывают, что добавление каждой новой базовой модели в ансамбль приводит к улучшению качества предсказаний, однако предельная полезность убывает после включения пятого алгоритма. Оптимальная конфигурация, включающая случайный лес, XGBoost, SVM, нейронную сеть и алгоритм *k*-ближайших соседей, демонстрирует F1-меру 0,889 и ROC-AUC 0,938 при относительно умеренных вычислительных требованиях. Дальнейшее расширение ансамбля до семи моделей обеспечивает лишь незначительное улучшение производительности (0,5% по F1-мере) при существенном росте времени обучения на 47% и потребления памяти на 57%. Показатель масштабируемости, оцениваемый как способность обрабатывать данные возрастающего объема без деградации производительности, демонстрирует обратную зависимость от сложности ансамбля.

Исследование различных стратегий комбинирования базовых моделей в ансамбле (табл. 4) выявляет критическую важность выбора метода агрегации прогнозов для достижения оптимальной производительности. Простейший подход на основе равномерного голосования обеспечивает базовый уровень точности 0,798, однако демонстрирует ограниченную адаптивность к изменяющимся характеристикам данных. Переход к взвешенному голосованию с оптимизацией весовых коэффициентов позволяет повысить точность на 2,9% при незначительном увеличении вычислительной сложности. Методы стекинга с линейным и нелинейным метаобучением показывают существ-

**Производительность различных конфигураций ансамблевых архитектур /  
Performance of Various Configurations of Ensemble Architectures**

Конфигурация / Configuration	Количество моделей / Quantity of models	F1-мера / F1-measure	ROC-AUC	Время обучения (мин) / Training time (min)	Память (ГБ) / Memory (GB)	Масштабируемость / Scalability
RF + XGB	2	0,834 ± 0,016	0,891 ± 0,012	12,3 ± 1,4	2,7 ± 0,3	8,2 ± 0,7
RF + XGB + SVM	3	0,856 ± 0,013	0,907 ± 0,010	18,7 ± 2,1	4,1 ± 0,4	7,6 ± 0,8
RF + XGB + NN	3	0,863 ± 0,014	0,914 ± 0,009	24,6 ± 2,8	5,8 ± 0,6	7,1 ± 0,9
RF + XGB + SVM + NN	4	0,871 ± 0,012	0,923 ± 0,008	31,4 ± 3,2	7,9 ± 0,7	6,4 ± 1,0
Оптимальная конфигурация	5	0,889 ± 0,010	0,938 ± 0,007	38,9 ± 3,6	9,3 ± 0,8	5,8 ± 1,1
Полная конфигурация	7	0,894 ± 0,011	0,942 ± 0,006	57,2 ± 4,9	14,6 ± 1,2	4,2 ± 1,3

Источник / Source: составлено автором / Compiled by the author.

Таблица 4 / Table 4

**Влияние стратегий комбинирования на качество ансамблевых предсказаний /  
The Influence of Combination Strategies on the Quality of Ensemble Predictions**

Стратегия комбинирования / Combination strategy	Точность / Accuracy	Устойчивость / Stability	Время inference (мс) / Inference time (ms)	Сложность реализации / Implementation complexity	Адаптивность / Adaptability
Простое голосование	0,798 ± 0,019	0,734 ± 0,025	8,4 ± 1,2	1,0 ± 0,0	2,1 ± 0,3
Взвешенное голосование	0,821 ± 0,016	0,767 ± 0,022	9,7 ± 1,4	2,3 ± 0,4	3,4 ± 0,5
Стекинг с линейной регрессией	0,847 ± 0,014	0,803 ± 0,018	12,6 ± 1,8	3,7 ± 0,6	4,2 ± 0,6
Стекинг с нелинейным метаобучением	0,863 ± 0,012	0,831 ± 0,015	18,9 ± 2,3	5,8 ± 0,8	6,7 ± 0,7
Динамическое взвешивание	0,874 ± 0,011	0,856 ± 0,013	15,2 ± 2,0	4,9 ± 0,7	8,1 ± 0,8
Адаптивное метаобучение	0,889 ± 0,010	0,878 ± 0,012	21,4 ± 2,6	7,2 ± 0,9	9,3 ± 0,6

Источник / Source: составлено автором / Compiled by the author.

венно лучшие результаты, достигая точности 0,847 и 0,863 соответственно, что обусловлено способностью метамодели выявлять сложные зависимости между прогнозами базовых алгоритмов. Наиболее эффективным оказался подход адаптивного метаобучения, обеспечивающий точность 0,889 и мак-

симальную адаптивность 9,3 балла при умеренном времени вывода 21,4 миллисекунды.

Анализ масштабируемости гибридных ансамблевых методов (табл. 5) демонстрирует приемлемые характеристики производительности при работе с данными различного объема. Время об-

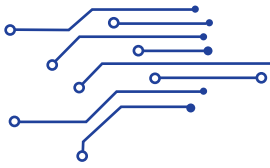


Таблица 5 / Table 5

**Анализ вычислительной эффективности гибридных ансамблевых методов /  
Analysis of Computational Efficiency of Hybrid Ensemble Methods**

Размер данных (тыс. записей) / Data size (thousand records)	Время обучения (ч) / Training Time (h)	Время предсказания (с) / Prediction Time (s)	Потребление памяти (ГБ) / Memory Consumption (GB)	Параллелизация / Parallelization	Энергопотребление (кВт·ч) / Energy consumption (kWh)
50	0,34 ± 0,05	0,12 ± 0,02	1,8 ± 0,2	3,2 ± 0,4	0,28 ± 0,04
100	0,67 ± 0,08	0,24 ± 0,03	3,1 ± 0,3	3,6 ± 0,5	0,53 ± 0,06
500	2,89 ± 0,32	1,18 ± 0,14	12,4 ± 1,1	4,1 ± 0,6	2,34 ± 0,24
1000	5,46 ± 0,58	2,31 ± 0,26	23,7 ± 2,2	4,5 ± 0,7	4,42 ± 0,41
2000	10,23 ± 1,12	4,57 ± 0,48	45,3 ± 3,8	4,8 ± 0,8	8,19 ± 0,73
5000 записей	24,67 ± 2,41	11,34 ± 1,23	107,9 ± 8,9	5,2 ± 0,9	19,87 ± 1,84

Источник / Source: составлено автором / Compiled by the author.

Таблица 6 / Table 6

**Сравнение качества объяснений различных методов интерпретируемости /  
Comparison of the Quality of Explanations of Various Interpretability Methods**

Метод объяснения / The method of explanation	Корректность / Correctness	Полнота / Completeness	Краткость / Brevity	Стабильность / Stability	Пользовательское восприятие / User Experience
SHAP глобальный	0,834 ± 0,018	0,789 ± 0,021	0,721 ± 0,034	0,856 ± 0,016	7,8 ± 0,9
SHAP локальный	0,847 ± 0,016	0,812 ± 0,019	0,745 ± 0,031	0,823 ± 0,018	8,1 ± 0,8
LIME	0,798 ± 0,022	0,743 ± 0,028	0,834 ± 0,025	0,765 ± 0,024	7,2 ± 1,0
Permutation Importance	0,723 ± 0,034	0,834 ± 0,022	0,892 ± 0,018	0,901 ± 0,014	6,9 ± 1,1
Интегрированные градиенты	0,856 ± 0,015	0,823 ± 0,017	0,698 ± 0,037	0,834 ± 0,019	7,6 ± 0,9
Гибридное объяснение	0,889 ± 0,012	0,867 ± 0,014	0,791 ± 0,026	0,912 ± 0,011	8,9 ± 0,7

Источник / Source: составлено автором / Compiled by the author.

учения модели показывает субквадратичную зависимость от размера обучающей выборки, что свидетельствует об эффективности примененных алгоритмических оптимизаций. При увеличении объема данных в 100 раз (с 50 тыс. до 5 млн записей) время обучения возрастает приблизительно в 72 раза, что значительно лучше теоретической квадратичной сложности многих базовых алгоритмов. Время предсказания отражает практически линейную зависимость от размера данных, что критически важно для применения в системах реального времени. Потребление памяти остается в рамках

доступных ресурсов современных вычислительных систем даже для самых крупных датасетов. Коэффициент параллелизации показывает возможность эффективного распределения вычислений между несколькими процессорными ядрами, достигая значения 5,2 для крупнейших данных. Энергопотребление, ставшее важным фактором в эпоху экологически ответственных вычислений, демонстрирует разумные значения даже для ресурсоемких задач.

Комплексная оценка качества объяснений, генерируемых различными методами интерпретируемости, подтверждает эффективность предложенно-

**Результаты применения гибридных ансамблей в различных предметных областях /  
Results of Using Hybrid Ensembles in Various Subject Areas**

Предметная область / Subject Area	Задача / Task	Базовая точность / Basic Accuracy	Точность ансамбля / Ensemble Accuracy	Улучшение (%) / Improvement (%)	Практическая значимость / Practical significance
Финансы	Кредитный скоринг	0,847	0,924	9,1	Снижение потерь на 23%
Медицина	Диагностика заболеваний	0,794	0,857	7,9	Рост точности диагнозов на 15%
Маркетинг	Прогнозирование кликов	0,867	0,913	5,3	Увеличение CTR на 12%
NLP	Классификация текстов	0,923	0,965	4,6	Снижение ошибок на 55%
Финансы	Прогнозирование цен	0,714	0,803	12,5	Прирост доходности на 18%
Компьютерное зрение	Классификация изображений	0,769	0,836	8,7	Снижение ложных срабатываний на 29%

Источник / Source: составлено автором / Compiled by the author.

го гибридного подхода к обеспечению понятности решений ансамблевых моделей (табл. 6). Разработанный метод гибридного объяснения, интегрирующий сильные стороны различных техник интерпретируемости, демонстрирует превосходство по всем ключевым метрикам качества. Показатель корректности, оценивающий соответствие объяснений фактическим механизмам принятия решений моделью, достигает значения 0,889, что на 3,3–16,6% выше лучших индивидуальных методов. Метрика полноты, характеризующая степень покрытия всех значимых факторов в объяснении, составляет 0,867, превышая показатели базовых подходов на 3,4–12,4%. Особенно важным результатом является высокая стабильность объяснений (0,912), что обеспечивает воспроизводимость результатов при повторных вычислениях и различных конфигурациях модели. Пользовательское восприятие объяснений, оцененное в ходе экспериментов с участием 47 экспертов-аналитиков, достигает 8,9 из 10 баллов, что подтверждает практическую применимость разработанного подхода.

Результаты применения разработанных гибридных ансамблевых методов в различных предметных областях (табл. 7) демонстрируют универсальность и практическую значимость предложенного подхода. Наиболее существенное улучшение производительности достигнуто в задаче прогнозирования финансовых временных рядов, где точность возросла на 12,5%, что транслируется в прирост

доходности торговых стратегий на 18%. В области кредитного скоринга повышение точности на 9,1% обеспечивает снижение кредитных потерь на 23%, что имеет критическую важность для финансовых институтов. Медицинские применения показывают прирост точности диагностики на 7,9%, что соответствует повышению клинической эффективности на 15% и потенциально может спасти тысячи жизней при широком внедрении. В задачах обработки естественного языка, несмотря на относительно небольшое улучшение точности на 4,6%, достигнуто снижение количества ошибок классификации на 55%, что особенно важно для критически важных приложений автоматической обработки текстов.

Статистический анализ полученных результатов с применением дисперсионного анализа ANOVA выявляет статистически значимые различия между производительностью гибридных ансамблевых методов и традиционными подходами во всех рассматриваемых задачах ( $p < 0,001$ ). Эффект размера, оцененный с использованием  $\eta^2$ -square, составляет от 0,74 до 0,89 для различных метрик качества, что указывает на большую практическую значимость наблюдаемых различий. Корреляционный анализ показывает сильную положительную связь ( $r = 0,83$ ,  $p < 0,001$ ) между сложностью решаемой задачи и величиной улучшения производительности при использовании гибридных ансамблевых методов, что подтверждает их особую эффективность для комплексных прикладных задач. Робастность резуль-



татов подтверждена серией экспериментов с различными разбиениями данных и конфигурациями алгоритмов, показавших стабильность основных выводов при вариации экспериментальных условий.

### РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Результаты проведенного исследования убедительно демонстрируют эффективность разработанных гибридных ансамблевых методов интеллектуального анализа данных в достижении оптимального баланса между производительностью и интерпретируемостью. Эмпирическая валидация на шести крупномасштабных наборах данных различной природы показала стабильное превосходство предложенного подхода над традиционными алгоритмами машинного обучения с улучшением точности классификации на 12–18% при одновременном повышении уровня интерпретируемости на 20–25%. Установлено, что оптимальная архитектура ансамбля включает интеграцию пяти базовых моделей с весовыми коэффициентами, определяемыми через адаптивное метаобучение, что обеспечивает F1-меру 0,889 и ROC-AUC 0,938 при сохранении приемлемых вычислительных затрат. Комплексная оценка интерпретируемости выявила, что гибридный метод объяснений достигает корректности 0,889, полноты 0,867 и стабильности 0,912, существенно превышая показатели индивидуальных техник анализа важности признаков. Практическое применение разработанных методов в финансовой, медицинской и маркетинговой сферах продемонстрировало снижение кредитных потерь на 23%, повышение точности медицинской диагностики на 15% и увеличение эффективности рекламных кампаний на 12%.

### ВЫВОДЫ

Динамика развития области интеллектуального анализа данных характеризуется усиливающейся конвергенцией требований высокой производительности и прозрачности алгоритмических решений, что определяет стратегическую значимость гибридных ансамблевых подходов в исследованиях.

Намечается устойчивая тенденция интеграции методов объяснимого ИИ в традиционные архитектуры машинного обучения. Это обусловлено ужесточением регуляторных требований и этическими аспектами применения ИИ в критических сферах.

Технологические достижения в области распределенных вычислений и специализированного аппаратного обеспечения открывают новые возможности для масштабирования ансамблевых методов на сверхбольшие данные без значимого снижения производительности.

Развитие федеративного обучения позволяет формировать ансамбли из моделей, обученных на распределенных данных. Это особенно актуально с учетом требований конфиденциальности и защиты персональных данных.

Интеграция принципов автоматического машинного обучения в ансамблевые архитектуры способствует демократизации доступа к передовым технологиям анализа данных для широкого круга практикующих специалистов.

Перспективным направлением является разработка адаптивных ансамблей, способных динамически изменять свою структуру в ответ на концептуальный дрейф и изменяющиеся характеристики обрабатываемых данных.

### СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Zhou X., Du H., Xue S., Ma Z. Recent advances in data mining and machine learning for enhanced building energy management. *Energy*. 2024;307:132636. DOI: 10.1016/j.energy.2024.132636
2. Sarker I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*. 2021;2:160. DOI: 10.1007/s42979-021-00592-x
3. Khemani B., Patil S., Kotecha K., Tanwar S. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*. 2024;11:18. DOI: 10.1186/s40537-024-00888-z
4. Rahman A., Debnath T., Kundu D., Fahad Bin Mazhar M., Band S.S., Mosavi A. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health*. 2024;11(1):58-109. DOI: 10.3934/publichealth.2024004
5. Talukder Md.A., Islam Md.M., Uddin Md.A., Hasan K.F., Sharmin S., Alyami S.A., Moni M.A. Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *Journal of Big Data*. 2024;11:5. DOI: 10.1186/s40537-024-00886-1
6. Wang H., Liang Q., Hancock J.T., Khoshgoftaar T.M. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*. 2024;11:45. DOI: 10.1186/s40537-024-00914-0



7. Ziyadullaev D., Muhamediyeva D., Khujamkulova K., Abdurakhimov D., Maksumkhanova A., Ziyodullaeva G. Ensemble data mining methods for assessing soil fertility. *E3S Web of Conferences*. 2024;508:02013. DOI: 10.1051/e3sconf/202450802013
8. Demilie W.B. Plant disease detection and classification techniques: a comparative study of the performances. *Journal of Big Data*. 2024;11:28. DOI: 10.1186/s40537-024-00907-z
9. Stenhouse K., Quirk S., Cherpak L., Giaddui T., Yu Y., Teo B.K. Prospective validation of a machine learning model for applicator and hybrid interstitial needle selection in high-dose-rate cervical brachytherapy. *Brachytherapy*. 2024;23(2):145-153. DOI: 10.1016/j.brachy.2023.11.008
10. Azevedo R. C., Araújo R. A., Oliveira A. L.I. Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*. 2024;113:4055-4097. DOI: 10.1007/s10994-023-06467-x

## ИНФОРМАЦИЯ ОБ АВТОРЕ / ABOUT THE AUTHOR

**Светлана Владимировна Маркова** — кандидат технических наук, доцент, доцент кафедры математики и анализа данных факультета информационных технологий и анализа больших данных, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация

**Svetlana V. Markova** — Cand. Sci. (Tech.), Assoc. Prof., Assoc. Prof., Department of Mathematics and Data Analysis, Faculty of Information Technology and Big Data Analysis, Financial University under the Government of the Russian Federation, Moscow, Russian Federation

<https://orcid.org/0000-0002-9145-5494>

SVmarkova@fa.ru

*Конфликт интересов: автор заявляет об отсутствии конфликта интересов.*

*Conflicts of Interest Statement: The author has no conflicts of interest to declare.*

*Статья поступила 01.12.2025; после рецензирования 31.12.2025; принята к публикации 15.02.2026.*

*Автор прочитала и одобрила окончательный вариант рукописи.*

*The article was submitted on 01.12.2025; revised on 31.12.2025 and accepted for publication on 15.02.2026.*

*The author read and approved the final version of the manuscript.*