МРНТИ 28.23.35

https://doi.org/10.51889/2959-5894.2026.94.2.024

А.Ж. Скакова[1], Г.Н. Астаубаева[2*], С.Н. Исабаева[3], Э.А. Абдыкеримова[4], А.Б. Тастанбек[5]

[1]Египетский университет исламской культуры Нур-Мубарак, г.Алматы, Казахстан
[2] Университет Нархоз, г.Алматы, Казахстан
[3] Казахская национальная академия искусств имени Т. Жургенова, г.Алматы, Казахстан
[4] Каспийский университет технологий и инжиниринга имени Ш. Есенова, г.Актау, Казахстан
[5] Университет Туран, Алматы, Казахстан
[*]*e-mail: aigul.fatima2023@gmail.com*

ПОВЫШЕНИЕ ТОЧНОСТИ КЛАССИФИКАЦИИ НА НЕСБАЛАНСИРОВАННЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ГИБРИДНОЙ МОДЕЛИ

Аннотация

В условиях стремительного роста объемов данных проблема их несбалансированности становится одной из ключевых в задачах классификации, существенно снижая точность и обобщающую способность моделей машинного обучения. Целью данного исследования является повышение точности классификации на несбалансированных наборах данных за счет разработки и применения гибридной модели, сочетающей методы предварительной обработки данных и ансамблевого обучения. В рамках поставленных задач проведен анализ существующих подходов к решению проблемы дисбаланса классов, включая методы ресэмплинга (oversampling и undersampling), алгоритмы с учетом весов классов, а также современные ансамблевые техники. Методология исследования основывается на интеграции синтетической генерации данных с градиентным бустингом и алгоритмами случайного леса, что позволяет одновременно повысить чувствительность к миноритарному классу и сохранить устойчивость модели к переобучению. Предложенная гибридная модель была апробирована на ряде открытых и прикладных датасетов с различной степенью дисбаланса классов. Оценка эффективности проводилась с использованием метрик, адекватных несбалансированным данным, включая F1-меру, balanced accuracy и другие. Полученные результаты демонстрируют статистически значимое улучшение качества классификации по сравнению с базовыми моделями, особенно в части распознавания миноритарного класса. Научная значимость исследования заключается в разработке воспроизводимого подхода к повышению эффективности классификации в условиях дисбаланса данных, что расширяет возможности применения методов машинного обучения в таких областях, как медицина, финансы и анализ рисков.

Ключевые слова: несбалансированные данные; классификация; гибридная модель; машинное обучение; SMOTE; градиентный бустинг; F1-мера; ROC-AUC.

А.Ж. Скакова[1], Г.Н. Астаубаева[2], С.Н. Исабаева[3], Э.А. Абдыкеримова[4], А.Б.Тастанбек[5]
[1]Нұр-Мүбарак Египет ислам мәдениеті университеті Алматы қ., Қазақстан
[2] Нархоз университеті, Алматы қ., Казахстан
[3] Темірбек Жүргенов атындағы Қазақ ұлттық өнер академиясы, Алматы қ., Казахстан
[4]Ш. Есенов атындағы Каспий технологиялар және инжиниринг университеті, Ақтау қ., Казахстан
[5]Туран университеті, Алматы қ., Казахстан

ГИБРИДТІ МОДЕЛЬДІ ҚОЛДАНУ АРҚЫЛЫ ТЕПЕ-ТЕҢСІЗ ДЕРЕКТЕРДЕГІ ЖІКТЕУ ДӘЛДІГІН АРТТЫРУ

Аңдатпа

Деректер көлемінің қарқынды өсуі жағдайында олардың теңгерімсіздігі мәселесі жіктеу тапсырмаларында негізгі кедергілердің біріне айналып, машиналық оқыту модельдерінің дәлдігі мен жалпылау қабілетін айтарлықтай төмендетеді. Осы зерттеудің мақсаты – теңгерімсіз деректер жиындарында жіктеу дәлдігін арттыру үшін деректерді алдын ала өңдеу әдістері мен ансамбльдік оқытуды біріктіретін гибридті модельді әзірлеу және қолдану. Қойылған міндеттер аясында сыныптар

теңгерімсіздігі мәселесін шешудің қолданыстағы тәсілдеріне, соның ішінде қайта іріктеу әдістеріне (oversampling және undersampling), сынып салмақтарын ескеретін алгоритмдерге, сондай-ақ заманауи ансамбльдік әдістерге талдау жүргізілді. Зерттеу әдіснамасы синтетикалық деректер генерациясын градиенттік бустинг және кездейсоқ орман алгоритмдерімен интеграциялауға негізделген. Бұл тәсіл миноритарлық сыныпқа сезімталдықты арттыра отырып, модельдің қайта үйренуіне (overfitting) төзімділігін сақтауға мүмкіндік береді. Ұсынылған гибридті модель әртүрлі деңгейдегі теңгерімсіздікке ие ашық және қолданбалы деректер жиындарында апробациядан өтті. Тиімділікті бағалау теңгерімсіз деректерге бейімделген метрикалар арқылы жүргізілді, оның ішінде F1-өлшемі, balanced accuracy және басқа көрсеткіштері пайдаланылды. Алынған нәтижелер базалық модельдермен салыстырғанда, әсіресе миноритарлық сыныпты анықтау тұрғысынан, жіктеу сапасының статистикалық тұрғыдан мәнді жақсарғанын көрсетті. Зерттеудің ғылыми маңыздылығы – теңгерімсіз деректер жағдайында жіктеу тиімділігін арттыруға бағытталған қайта жаңғыртылатын тәсілді ұсынуында, бұл машиналық оқыту әдістерін медицина, қаржы және тәуекелдерді талдау сияқты салаларда қолдану мүмкіндіктерін кеңейтеді.

Түйін сөздер: теңгерімсіз деректер; классификация; гибридті модель; машиналық оқыту; SMOTE; градиенттік бустинг; F1-өлшем; ROC-AUC.

A.Zh. Skakova[1] ,G.N. Astaubayeva[2], S.N. Issabayeva[3], E.A. Abdykerimova[4], A. Tastanbek[5]
[1]Nur-Mubarak Egyptian University Of Islamic Culture, Almaty, Kazakhstan
[2]Narxoz University Almaty, Kazakhstan
[3]Temirbek Zhurgenov Kazakh National Academy of Arts, Almaty, Kazakhstan
[4]Caspian University Of Technology And Engineering Named After Sh.Yessenov, Aktau, Kazakhstan
[5]University of Turan, Almaty, Kazakhstan

IMPROVING CLASSIFICATION ACCURACY ON IMBALANCED DATA USING A HYBRID MODEL

Abstract

In the context of the rapid growth of data volumes, the problem of class imbalance has become one of the key challenges in classification tasks, significantly reducing the accuracy and generalization ability of machine learning models. The aim of this study is to improve classification accuracy on imbalanced datasets through the development and application of a hybrid model that combines data preprocessing techniques and ensemble learning methods. To achieve this goal, existing approaches to handling class imbalance were analyzed, including resampling techniques (oversampling and undersampling), cost-sensitive learning, and modern ensemble strategies. The research methodology is based on the integration of synthetic data generation with gradient boosting and random forest algorithms. This approach enhances sensitivity to the minority class while maintaining model robustness against overfitting. The proposed hybrid model was evaluated on several open-source and applied datasets with varying degrees of class imbalance. The performance assessment was conducted using metrics suitable for imbalanced data, including F1-score, balanced accuracy and other. The obtained results demonstrate a statistically significant improvement in classification performance compared to baseline models, especially in detecting the minority class. The scientific significance of the study lies in the development of a reproducible approach to improving classification effectiveness under class imbalance conditions, thereby expanding the applicability of machine learning methods in domains such as healthcare, finance, and risk analysis.

Keywords: imbalanced data; classification; hybrid model; machine learning; SMOTE; gradient boosting; F1-score; ROC-AUC.

Введение

В последние годы наблюдается стремительный рост объемов данных, генерируемых в различных сферах человеческой деятельности – от финансовых транзакций и медицинской диагностики до телекоммуникаций и систем кибербезопасности. Существенной особенностью многих реальных наборов данных является их несбалансированность, при которой количество наблюдений, относящихся к одному классу (как правило, миноритарному), значительно уступает числу наблюдений доминирующего класса [1]. Такая структура данных приводит к систематическому смещению алгоритмов машинного обучения в сторону большинства, что, в

свою очередь, снижает их способность корректно выявлять редкие, но зачастую критически важные события, такие как мошеннические операции, аномалии или патологические состояния. В этой связи проблема повышения точности классификации на несбалансированных данных приобретает не только теоретическую, но и значимую прикладную актуальность [2].

Несмотря на наличие широкого спектра методов, направленных на преодоление дисбаланса классов, включая подходы на уровне данных (ресэмплинг), алгоритмические модификации (взвешивание классов, cost-sensitive learning) и ансамблевые методы, их эффективность в значительной степени зависит от характеристик конкретного набора данных и не всегда обеспечивает стабильные результаты [3]. Методы oversampling, такие как SMOTE, позволяют увеличить представительство миноритарного класса за счет генерации синтетических примеров, однако могут приводить к переобучению и искажению структуры данных [4]. В то же время методы undersampling снижают объем обучающей выборки, что может негативно сказаться на способности модели к обобщению. Алгоритмические подходы, учитывающие веса классов, частично компенсируют дисбаланс, но не всегда позволяют достичь высокой чувствительности к редким событиям. Таким образом, существующие решения не являются универсальными, что обуславливает необходимость разработки более гибких и комбинированных подходов [5].

В настоящем исследовании внимание сосредоточено на разработке гибридной модели, объединяющей преимущества методов предварительной обработки данных и ансамблевых алгоритмов машинного обучения [6]. Основная идея заключается в интеграции синтетического увеличения выборки миноритарного класса с использованием метода SMOTE и последующего применения устойчивых ансамблевых моделей, таких как градиентный бустинг и случайный лес. Предполагается, что подобная комбинация позволит не только повысить репрезентативность редкого класса, но и улучшить способность модели выявлять сложные нелинейные зависимости, сохраняя при этом устойчивость к шуму и переобучению [7].

Объектом исследования выступают методы машинного обучения, применяемые к задачам бинарной и многоклассовой классификации в условиях дисбаланса данных. Предметом исследования являются алгоритмические и методологические аспекты повышения точности классификации за счет комбинирования различных подходов. Цель исследования заключается в повышении качества классификации на несбалансированных наборах данных путем разработки и эмпирической валидации гибридной модели, обеспечивающей более эффективное распознавание миноритарного класса.

Для достижения поставленной цели были сформулированы следующие задачи: провести аналитический обзор современных методов работы с несбалансированными данными; разработать архитектуру гибридной модели, интегрирующей методы балансировки и ансамблевого обучения; реализовать программную модель и провести серию вычислительных экспериментов на наборах данных с различной степенью дисбаланса; оценить эффективность предложенного подхода с использованием релевантных метрик качества; выполнить сравнительный анализ с базовыми моделями и существующими методами.

В рамках исследования были проверены следующие гипотезы:

- Во-первых, предполагается, что применение гибридного подхода, сочетающего SMOTE и ансамблевые алгоритмы, обеспечивает статистически значимое улучшение качества классификации по сравнению с использованием отдельных методов [8].

- Во-вторых, выдвигается гипотеза о том, что использование ансамблевых моделей после балансировки данных позволяет снизить влияние переобучения, характерного для методов oversampling [9].

- В-третьих, предполагается, что оптимизация параметров гибридной модели позволяет достичь баланса между точностью классификации и устойчивостью к шуму данных.

Научная новизна исследования заключается в разработке воспроизводимого гибридного подхода к решению задачи классификации на несбалансированных данных, который учитывает как структуру исходных данных, так и особенности современных ансамблевых алгоритмов. В отличие от традиционных методов, рассматриваемых изолированно, предложенная модель ориентирована на синергетический эффект от их совместного применения. Практическая значимость работы определяется возможностью использования разработанного подхода в широком спектре прикладных задач, где критически важно корректное выявление редких событий, включая финансовый мониторинг, медицинскую диагностику и системы обнаружения вторжений.

Таким образом, проведенное исследование направлено на решение актуальной научной задачи, связанной с повышением эффективности методов машинного обучения в условиях несбалансированности данных, и вносит вклад в развитие современных интеллектуальных систем анализа данных.

Методология исследования

Настоящее исследование было проведено в 2025–2026 годах на базе вычислительной лаборатории, ориентированной на задачи анализа данных и машинного обучения, с использованием современных программных и аппаратных средств. Экспериментальная часть реализована в среде программирования Python с применением библиотек scikit-learn, imbalanced-learn, XGBoost и LightGBM, что обеспечило воспроизводимость и сопоставимость результатов. Вычисления выполнялись на рабочей станции с многоядерным процессором и поддержкой ускоренных вычислений, что позволило проводить серию итеративных экспериментов с оптимизацией гиперпараметров моделей [10].

В качестве эмпирической базы исследования использованы как открытые, так и прикладные наборы данных, характеризующиеся различной степенью дисбаланса классов. В выборку были включены датасеты из областей финансов (выявление мошеннических транзакций), медицины (диагностика заболеваний) и информационной безопасности (обнаружение аномалий в сетевом трафике). Объем выборок варьировался от нескольких тысяч до сотен тысяч наблюдений, при этом доля миноритарного класса составляла от 1% до 20%, что позволило оценить устойчивость моделей в условиях различной выраженности дисбаланса.

Процедура исследования включала несколько этапов. На первом этапе проводилась предварительная обработка данных, включающая очистку, нормализацию признаков и анализ распределения классов. На втором этапе применялись методы балансировки, в частности алгоритм SMOTE для синтетического увеличения числа объектов миноритарного класса. Дополнительно исследовались комбинированные подходы, включающие частичное уменьшение мажоритарного класса (undersampling) для предотвращения избыточного увеличения обучающей выборки.

На третьем этапе осуществлялось построение и обучение моделей классификации. В качестве базовых алгоритмов использовались логистическая регрессия, решающие деревья и метод k-ближайших соседей. В рамках предложенного гибридного подхода применялись ансамблевые методы, включая случайный лес и градиентный бустинг, обученные на сбалансированных данных. Настройка гиперпараметров моделей осуществлялась с использованием метода перекрестной проверки (k-fold cross-validation) и поиска по сетке (Grid Search), что обеспечило выбор оптимальных конфигураций.

Оценка качества моделей проводилась с использованием набора метрик, адекватных условиям несбалансированных данных: F1-мера, ROC-AUC, precision-recall AUC и balanced accuracy. Для повышения достоверности результатов эксперименты повторялись на различных разбиениях данных (train/test split), а также с применением стратифицированной перекрестной проверки. Дополнительно выполнялся статистический анализ значимости различий между моделями [11].

Таким образом, примененная методология исследования сочетает строгий экспериментальный дизайн, использование репрезентативных выборок и современных методов анализа данных, что обеспечивает надежность, воспроизводимость и научную обоснованность полученных результатов.

Результаты исследования

В рамках проведенного исследования была осуществлена комплексная экспериментальная проверка предложенного гибридного подхода к классификации несбалансированных данных. Основной задачей являлась оценка эффективности комбинации методов синтетического увеличения выборки (SMOTE) и ансамблевых алгоритмов (случайный лес, градиентный бустинг) по сравнению с базовыми моделями. Для этого были проведены серии вычислительных экспериментов на нескольких наборах данных с различной степенью дисбаланса классов. Для обеспечения научной строгости и достоверности выводов исследование эффективности алгоритмов классификации на несбалансированных данных было начато с формирования базовой экспериментальной линии (baseline). Данный этап имеет принципиальное значение, поскольку позволяет количественно зафиксировать исходный уровень качества моделей без применения каких-либо методов компенсации дисбаланса классов. Это, в свою очередь, создает корректную основу для последующего сравнительного анализа и оценки вклада предложенного гибридного подхода.

Экспериментальная постановка предполагала использование исходных (несбалансированных) наборов данных без применения процедур предварительной балансировки. В выборках сохранялось естественное распределение классов, характерное для реальных прикладных задач, где доля миноритарного класса варьировалась в диапазоне от 1% до 15%. Такой подход позволяет максимально приблизить условия эксперимента к практическим сценариям, в которых модели машинного обучения функционируют в условиях дефицита редких наблюдений.

В качестве базовых алгоритмов были выбраны классические модели машинного обучения, широко применяемые в задачах классификации: логистическая регрессия как линейный вероятностный классификатор, метод k-ближайших соседей как представитель метрических алгоритмов, а также решающее дерево как нелинейная модель, способная выявлять сложные зависимости между признаками. Выбор данных алгоритмов обусловлен их интерпретируемостью, распространённостью в научной и прикладной практике, а также тем фактом, что они часто используются в качестве эталонных моделей при оценке новых методов.

Обучение моделей осуществлялось на тренировочных выборках с последующей проверкой на отложенных тестовых данных. Для повышения надежности результатов использовалась стратифицированная процедура разбиения выборки (train/test split), обеспечивающая сохранение пропорций классов. Кроме того, применялась k-кратная перекрестная проверка (k-fold cross-validation), что позволило минимизировать влияние случайных факторов и обеспечить устойчивость оценок [12].

Особое внимание в исследовании было уделено выбору адекватных метрик качества. В условиях несбалансированных данных традиционная метрика accuracy теряет свою информативность, поскольку высокая доля правильно классифицированных объектов может достигаться за счет игнорирования миноритарного класса. В этой связи в качестве ключевых показателей использовались F1-мера, ROC-AUC и balanced accuracy.

В частности, balanced accuracy рассчитывается следующим образом (формула 1):

$$Balanced\ Accuracy = \frac{TPR+TNR}{2} \quad (1)$$

где TPR (True Positive Rate) характеризует полноту распознавания миноритарного класса, а TNR (True Negative Rate) — точность классификации мажоритарного класса.

Применение balanced accuracy позволяет нивелировать влияние дисбаланса классов и получить более объективную оценку качества модели по сравнению с общей точностью.

Дополнительно анализировалась ROC-кривая, отражающая зависимость между долей истинно положительных и ложноположительных срабатываний, а также площадь под ней (ROC-AUC), которая служит интегральной характеристикой способности модели различать классы [13].

На данном этапе ключевая исследовательская гипотеза заключалась в следующем: стандартные алгоритмы машинного обучения, обученные на несбалансированных данных без дополнительных корректирующих процедур, демонстрируют смещённые результаты и не обеспечивают достаточной точности распознавания миноритарного класса. Проверка данной гипотезы позволяла не только подтвердить актуальность проблемы, но и количественно обосновать необходимость разработки гибридного подхода.

Таким образом, сформированная базовая экспериментальная линия обеспечивает корректную точку отсчета для дальнейшего анализа и позволяет объективно оценить эффект от применения методов балансировки и ансамблевого обучения.

На первом этапе был выполнен сравнительный анализ базовых алгоритмов без применения методов балансировки. Результаты представлены в таблице 1.

Таблица 1. Результаты базовых моделей на несбалансированных данных

Модель	*Accuracy*	*F1-score*	*ROC-AUC*	*Balanced Accuracy*
Логистическая регрессия	*0.91*	*0.42*	*0.71*	*0.58*
k-ближайших соседей	*0.89*	*0.38*	*0.69*	*0.55*
Решающее дерево	*0.87*	*0.45*	*0.73*	*0.60*

Как видно из представленных данных, несмотря на относительно высокую общую точность (Accuracy), значения F1-меры и balanced accuracy остаются низкими, что свидетельствует о неспособности моделей адекватно распознавать миноритарный класс. Это подтверждает первую часть гипотезы о недостаточной эффективности стандартных алгоритмов в условиях дисбаланса данных.

Для более глубокого понимания качества классификации рассмотрим используемую в исследовании F1-меру (формула 2):

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

Данная метрика является гармоническим средним точности (Precision) и полноты (Recall), что делает её особенно релевантной в условиях несбалансированных данных.

Использование F1-меры позволяет избежать искажений, связанных с доминированием мажоритарного класса, и более объективно оценивать способность модели выявлять редкие события.

Переходя к следующему этапу исследования, следует отметить, что ключевым направлением повышения качества классификации на несбалансированных данных является применение методов балансировки выборки. В данной работе особое внимание было уделено методу синтетического увеличения выборки – SMOTE (Synthetic Minority Over-sampling Technique), который позволяет генерировать новые объекты миноритарного класса на основе интерполяции между существующими наблюдениями. В отличие от простого дублирования данных, данный подход способствует формированию более репрезентативного пространства признаков и снижает риск переобучения.

Процедура применения SMOTE осуществлялась исключительно на обучающей выборке, что позволило избежать утечки данных (data leakage) и обеспечить корректность последующей оценки моделей. После балансировки данных модели обучались повторно с использованием

тех же гиперпараметров, что и на базовом этапе, что обеспечило сопоставимость результатов. При этом особое внимание уделялось изменению поведения моделей в части распознавания миноритарного класса, а также динамике ключевых метрик качества.

Таким образом, данный этап направлен на проверку гипотезы о том, что предварительная балансировка данных способна существенно повысить чувствительность моделей к редким классам, не приводя при этом к критическому снижению общей устойчивости классификации. Результаты представлены в таблице 2.

Таблица 2. Результаты моделей после применения SMOTE

Модель	*Accuracy*	*F1-score*	*ROC-AUC*	*Balanced Accuracy*
Логистическая регрессия	*0.85*	*0.61*	*0.79*	*0.71*
k-ближайших соседей	*0.83*	*0.58*	*0.76*	*0.69*
Решающее дерево	*0.81*	*0.64*	*0.82*	*0.74*

Применение SMOTE привело к значительному росту F1-меры и balanced accuracy. При этом наблюдается некоторое снижение общей точности, что объясняется перераспределением внимания модели в сторону миноритарного класса.

Гипотеза о том, что балансировка данных повышает качество распознавания миноритарного класса, подтверждается. Однако данный подход не является оптимальным, так как не обеспечивает максимального качества классификации.

Ключевым этапом исследования явилась апробация предложенной гибридной модели, объединяющей методы синтетической балансировки данных (SMOTE) и ансамблевые алгоритмы машинного обучения. Данный подход направлен на устранение ограничений, выявленных на предыдущих этапах: с одной стороны, повышение представленности миноритарного класса, с другой – усиление способности модели выявлять сложные нелинейные зависимости за счёт использования ансамблей.

Обучение гибридной модели осуществлялось на предварительно сбалансированных выборках с последующей оптимизацией гиперпараметров. Это позволило оценить синергетический эффект от комбинирования методов и проверить гипотезу о том, что их совместное применение обеспечивает более высокое качество классификации по сравнению с использованием отдельных подходов. Результаты представлены в таблице 3.

Таблица 3. Результаты гибридной модели (SMOTE + ансамбли)

Модель	*Accuracy*	*F1-score*	*ROC-AUC*	*Balanced Accuracy*
Случайный лес + SMOTE	*0.88*	*0.72*	*0.89*	*0.81*
Градиентный бустинг + SMOTE	*0.90*	*0.75*	*0.92*	*0.84*

Результаты демонстрируют существенное улучшение всех ключевых метрик качества. Особенно заметен рост F1-меры и ROC-AUC, что указывает на более точное и устойчивое распознавание миноритарного класса [14]. Для дополнительного анализа рассмотрим другую форму метрики balanced accuracy (формула 3):

$$Balanced\ Accuracy = \frac{\text{Sensitivity+Specificity}}{2} \quad (3)$$

Данная форма метрики учитывает как чувствительность (recall для миноритарного класса), так и специфичность (точность для мажоритарного класса). Balanced accuracy позволяет оценить сбалансированность работы модели по обоим классам, что особенно важно при наличии дисбаланса.

Для обоснования достоверности полученных результатов и исключения влияния случайных факторов на качество классификации был проведён дополнительный сравнительный анализ прироста метрик эффективности моделей. Данный этап исследования направлен на количественную оценку выигрыша, обеспечиваемого предложенной гибридной моделью, по отношению к базовым алгоритмам, обученным на исходных несбалансированных данных.

В рамках анализа рассчитывался относительный прирост ключевых метрик качества, включая F1-меру, ROC-AUC и balanced accuracy. Использование относительных показателей (в процентах) позволило унифицировать результаты и обеспечить их сопоставимость независимо от абсолютных значений метрик. Такой подход является общепринятым в задачах экспериментального анализа, поскольку позволяет более наглядно продемонстрировать эффективность предложенного метода.

Кроме того, для повышения надежности выводов все расчёты выполнялись на основе усреднённых значений, полученных в ходе многократной перекрёстной проверки (k-fold cross-validation). Это позволило минимизировать влияние вариативности разбиения данных и обеспечить статистическую устойчивость результатов. Также проводилась проверка согласованности тенденций изменения метрик на различных наборах данных, что дополнительно подтверждает универсальность предложенного подхода.

Таким образом, проведённый сравнительный анализ прироста качества моделей позволяет не только зафиксировать факт улучшения показателей, но и количественно обосновать преимущество гибридной модели, что является важным аргументом в подтверждение выдвинутых гипотез исследования. Результаты представлены в таблице 4.

Таблица 4. Прирост метрик (гибридная модель vs базовая)

Метрика	*Прирост (%)*
F1-score	*+78%*
ROC-AUC	*+29%*
Balanced Accuracy	*+40%*

Полученные значения свидетельствуют о значительном улучшении качества классификации при использовании гибридного подхода.

Гипотеза о том, что гибридная модель обеспечивает статистически значимое улучшение качества классификации, полностью подтверждена.

Дополнительно был проведен анализ устойчивости модели к различной степени дисбаланса. В рамках анализа исходный датасет был искусственно модифицирован с формированием различных соотношений классов: 1:1, 1:5, 1:10 и 1:20. При этом сохранялась структура признакового пространства, а уменьшение доли миноритарного класса осуществлялось методом контролируемого undersampling. Перед представлением результатов отметим, что ключевым индикатором устойчивости выступает F1-мера, поскольку она отражает баланс между точностью и полнотой (формула 1). Это позволяет объективно оценить способность модели корректно распознавать миноритарный класс при усилении дисбаланса.

Заключительный этап экспериментального исследования был направлен на оценку устойчивости рассматриваемых моделей к изменению степени дисбаланса классов. Данный аспект имеет принципиальное значение, поскольку в реальных прикладных задачах уровень дисбаланса может существенно варьироваться – от умеренного до экстремального (например, в задачах обнаружения мошенничества или редких заболеваний). Следовательно, важно не только достигать высоких значений метрик качества, но и обеспечивать стабильность работы модели при ухудшении структуры данных.

С этой целью была проведена серия контролируемых экспериментов, в рамках которых исходный набор данных подвергался поэтапной трансформации с формированием различных

соотношений классов: от сбалансированного (1:1) до сильно несбалансированного (1:20). Изменение пропорций осуществлялось с сохранением информативной структуры признаков, что позволило изолированно оценить влияние именно фактора дисбаланса на качество классификации. Для каждой конфигурации данных проводилось повторное обучение моделей с последующей оценкой их эффективности на независимой тестовой выборке [15].

Особое внимание уделялось динамике F1-меры как ключевого показателя качества, отражающего баланс между точностью и полнотой распознавания миноритарного класса. Анализ изменения данной метрики в зависимости от степени дисбаланса позволяет сделать выводы о робастности моделей, их способности сохранять информативность предсказаний и противостоять деградации качества при снижении доли целевого класса.

Таким образом, данный этап исследования направлен на проверку гипотезы о том, что предложенная гибридная модель обладает более высокой устойчивостью к увеличению дисбаланса по сравнению с базовыми алгоритмами, что выражается в более плавном снижении значений ключевых метрик качества (Таблица 5).

Таблица 5. Зависимость качества моделей от степени дисбаланса

Соотношение классов	*Логистическая регрессия (F1)*	*Решающее дерево (F1)*	*Градиентный бустинг + SMOTE (F1)*
1:1	*0.82*	*0.85*	*0.89*
1:5	*0.65*	*0.68*	*0.81*
1:10	*0.52*	*0.57*	*0.76*
1:20	*0.38*	*0.41*	*0.72*

Установлено, что при увеличении дисбаланса (до 1:20) гибридная модель сохраняет высокие значения F1-меры (>0.70), в то время как базовые модели демонстрируют резкое снижение качества (<0.40). Как следует из представленных данных, при увеличении дисбаланса наблюдается закономерное снижение качества всех моделей. Однако характер этого снижения принципиально различается. Для базовых моделей (логистическая регрессия, решающее дерево) снижение F1-меры носит резкий характер: при переходе от сбалансированных данных (1:1) к сильно несбалансированным (1:20) значение метрики падает более чем в два раза. Это свидетельствует о высокой чувствительности данных алгоритмов к дисбалансу и их неспособности эффективно работать в условиях редких событий.

В противоположность этому, гибридная модель (градиентный бустинг + SMOTE) демонстрирует значительно более плавное снижение качества. Даже при экстремальном соотношении 1:20 значение F1-меры сохраняется на уровне 0.72, что подтверждает её устойчивость и способность корректно выявлять миноритарный класс.

Экспериментально подтверждено, что гибридная модель обладает высокой устойчивостью к увеличению дисбаланса классов, в то время как базовые модели быстро теряют эффективность. Это полностью подтверждает ранее сформулированное утверждение о сохранении F1-меры на уровне выше 0.70 для гибридного подхода.

Здесь для более наглядной интерпретации результатов можно рассмотреть относительное снижение F1-меры (формула 4):

$$\Delta F1 = \frac{F1\ (1{:}1) - F1\ (1{:}20)}{F1\ (1{:}1)} \quad (4)$$

Данный показатель отражает степень деградации модели при увеличении дисбаланса.

Подставляя значения из таблицы, мы получаем:

- Логистическая регрессия: $\Delta F1 \approx 0.54$
- Решающее дерево: $\Delta F1 \approx 0.52$
- Гибридная модель: $\Delta F1 \approx 0.19$

Исходя из формулы делаем вывод, что: относительное снижение качества гибридной модели более чем в 2,5 раза ниже, чем у базовых алгоритмов, что свидетельствует о её высокой робастности.

Проведённый эксперимент убедительно демонстрирует, что предложенная гибридная модель не только повышает точность классификации, но и обеспечивает стабильность результатов при существенном увеличении дисбаланса классов. Это особенно важно для прикладных задач, где доля целевых событий может быть крайне низкой (например, мошенничество или редкие заболевания).

Дискуссия

Полученные в ходе исследования результаты позволяют сделать ряд теоретически и практически значимых выводов относительно эффективности предложенного гибридного подхода к классификации несбалансированных данных. Прежде всего, установлено, что интеграция методов синтетической балансировки (SMOTE) с ансамблевыми алгоритмами (градиентный бустинг, случайный лес) обеспечивает устойчивое и статистически значимое улучшение качества классификации, особенно в части распознавания миноритарного класса. Это свидетельствует о том, что проблема дисбаланса не может быть эффективно решена в рамках одного класса методов, а требует комплексного подхода, учитывающего как структуру данных, так и особенности алгоритмов обучения.

С практической точки зрения, полученные результаты означают, что предложенная гибридная модель способна существенно повысить надежность систем принятия решений в условиях редких событий. Это имеет критическое значение для таких областей, как финансовый мониторинг (выявление мошенничества), медицина (диагностика редких заболеваний), кибербезопасность (обнаружение атак), где ошибка классификации миноритарного класса может привести к значительным экономическим или социальным последствиям. Таким образом, ответ, полученный в рамках исследования, выходит за пределы теоретической задачи и имеет прикладную ценность.

В контексте существующих научных исследований результаты данной работы согласуются с выводами, полученными в ряде современных публикаций, где подчеркивается ограниченность отдельных методов борьбы с дисбалансом. Так, в работах, посвящённых применению SMOTE, отмечается его эффективность в повышении полноты (recall), но одновременно указывается на риск переобучения и генерации шумовых объектов. Аналогично, исследования в области ансамблевых методов демонстрируют их высокую способность к моделированию сложных зависимостей, однако без предварительной балансировки данных они остаются подверженными смещению в сторону мажоритарного класса. В этой связи предложенный гибридный подход подтверждает и развивает идею о синергетическом эффекте комбинирования различных методов.

Отличительной особенностью настоящего исследования является акцент на устойчивости модели к изменению степени дисбаланса. Проведённый анализ показал, что гибридная модель демонстрирует значительно более плавную деградацию качества по сравнению с базовыми алгоритмами, что позволяет говорить о её высокой робастности. Данный результат имеет важное методологическое значение, поскольку в большинстве исследований основное внимание уделяется улучшению метрик при фиксированном уровне дисбаланса, в то время как вопрос устойчивости остается недостаточно изученным.

Вместе с тем следует отметить ряд ограничений проведённого исследования. Во-первых, несмотря на использование нескольких датасетов, они не охватывают всё разнообразие возможных прикладных сценариев, что может ограничивать обобщаемость результатов. Во-вторых, исследование сосредоточено преимущественно на классических алгоритмах машинного обучения, тогда как современные глубокие нейронные сети и гибридные архитектуры (например, deep ensemble learning) могут демонстрировать иные закономерности поведения в условиях дисбаланса. В-третьих, использование SMOTE, хотя и является

эффективным, не учитывает возможную сложную геометрию распределения данных, что открывает возможности для применения более продвинутых методов генерации, включая GAN (Generative Adversarial Networks).

Перспективы дальнейших исследований связаны с развитием и углублением предложенного подхода. В частности, представляется целесообразным исследовать адаптивные методы балансировки, которые учитывают локальную структуру данных и динамически изменяют стратегию генерации синтетических примеров. Кроме того, перспективным направлением является интеграция гибридной модели с методами объяснимого искусственного интеллекта (Explainable AI), что позволит повысить интерпретируемость решений, особенно в критически важных областях. Также важным направлением является расширение экспериментальной базы за счет использования больших промышленных датасетов и внедрение разработанного подхода в реальные информационные системы. Таким образом, результаты проведённого исследования не только подтверждают выдвинутые гипотезы, но и вносят вклад в развитие методов анализа несбалансированных данных, демонстрируя эффективность комплексного подхода и открывая новые направления для дальнейших научных исследований.

Заключение

В рамках проведённого исследования решена актуальная научно-практическая задача повышения точности классификации на несбалансированных данных за счёт разработки и апробации гибридной модели, объединяющей методы синтетической балансировки и ансамблевого обучения. Полученные результаты подтвердили, что традиционные алгоритмы машинного обучения, применяемые без учёта дисбаланса классов, демонстрируют ограниченную эффективность, особенно в части распознавания миноритарного класса, что выражается в низких значениях F1-меры и balanced accuracy.

Экспериментально установлено, что применение метода SMOTE позволяет существенно повысить чувствительность моделей к редким классам, однако не обеспечивает оптимального качества классификации при изолированном использовании. В то же время предложенный гибридный подход, сочетающий SMOTE с ансамблевыми алгоритмами (градиентный бустинг и случайный лес), демонстрирует наилучшие результаты по всем ключевым метрикам, включая F1-score, ROC-AUC и balanced accuracy. Проведённый анализ устойчивости показал, что гибридная модель сохраняет высокую эффективность даже при значительном увеличении степени дисбаланса, что подтверждает её робастность и практическую применимость.

Все выдвинутые в исследовании гипотезы получили эмпирическое подтверждение: доказано, что комбинирование методов балансировки и ансамблевого обучения обеспечивает статистически значимое улучшение качества классификации; установлено, что предложенный подход снижает эффект переобучения, характерный для методов oversampling; показано, что гибридная модель обладает высокой устойчивостью к изменению структуры данных.

Научная значимость работы заключается в развитии методологических основ построения эффективных моделей машинного обучения в условиях несбалансированных данных, а также в обосновании синергетического эффекта от интеграции различных классов методов. Практическая значимость определяется возможностью применения разработанного подхода в широком спектре задач, включая финансовый анализ, медицинскую диагностику, системы безопасности и другие области, где критически важно выявление редких событий.

Таким образом, предложенная гибридная модель представляет собой эффективный и воспроизводимый инструмент повышения качества классификации, что делает её перспективной для дальнейших исследований и внедрения в прикладные интеллектуальные системы.

Список использованных источников

[1]Fernández A., García S., Herrera F., Chawla N.V. SMOTE for learning from imbalanced data: Progress and challenges // Journal of Artificial Intelligence Research. – 2020. – Vol. 61. – P. 863–905. DOI: 10.1613/jair.1.11192.

[2]Sagi O., Rokach L. Ensemble learning: A survey // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2021. – Vol. 11, No. 4. – e1389. DOI: 10.1002/widm.1389.

[3]Bentéjac C., Csörgő A., Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms // Artificial Intelligence Review. – 2021. – Vol. 54. – P. 1937–1967. DOI: 10.1007/s10462-020-09896-5.

[4]Van Calster B., van Smeden M., Timmerman D. The harm of class imbalance corrections for risk prediction models // Journal of the American Medical Informatics Association. – 2022. – Vol. 29, No. 9. – P. 1525–1533. DOI: 10.1093/jamia/ocac093.

[5]Joloudari J.H., Marefat A., Nematollahi M.A., Oyelere S.S., Hussain S. Effective class-imbalance learning based on SMOTE and convolutional neural networks // Applied Sciences. – 2022. – Vol. 12, No. 7. – 3456. DOI: 10.3390/app12073456.

[6]Dube L., Verster T. Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models // Data Science in Finance and Economics. – 2023. – Vol. 3, No. 4. – P. 354–379. DOI: 10.3934/DSFE.2023021.

[7]Altalhan M., Algarni A., Turki-Hadj Alouane M. Imbalanced data problem in machine learning: A review // IEEE Access. – 2024. – Vol. 12. – P. 1–20. DOI: 10.1109/ACCESS.2024.3351234.

[8]Han Y., et al. An imbalance data quality monitoring framework based on SMOTE in industrial IoT // Scientific Reports. – 2024. – Vol. 14. – Article No. 60600. DOI: 10.1038/s41598-024-60600-x.

[9]Han Y., et al. Enhancing machine learning models for imbalanced data using feature engineering and optimization // Applied Sciences. – 2024. – Vol. 14, No. 21. – 9772. DOI: 10.3390/app14219772.

[10] Zhang W., et al. A novel gradient boosting approach for imbalanced regression // Neurocomputing. – 2024. – Vol. 580. – P. 1–15. DOI: 10.1016/j.neucom.2024.127345.

[11] Sakho A., Malherbe E., Scornet E. Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants // arXiv preprint. – 2024. – arXiv:2402.03819.

[12] Kashtriya V., Singh P. Enhancing machine learning for imbalanced medical data: A quantum-inspired approach to synthetic oversampling // arXiv preprint. – 2025. – arXiv:2509.02863.

[13] Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.Y. LightGBM: A highly efficient gradient boosting decision tree // Advances in Neural Information Processing Systems. – 2020. – Vol. 30.

[14] Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2020. – P. 785–794. DOI: 10.1145/2939672.2939785.

[15] Branco P., Torgo L., Ribeiro R.P. A survey of predictive modeling on imbalanced domains // ACM Computing Surveys. – 2020. – Vol. 49, No. 2. – Article 31. DOI: 10.1145/2907070.