

**Форат Марк Игоревич**, магистрант, Новосибирский государственный технический университет, г. Новосибирск

**Алетдинова Анна Александровна**, доктор экономических наук, доцент, Новосибирский государственный технический университет, г. Новосибирск

## **МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ УРОВНЕЙ ЗАГРЯЗНЕНИЯ АТМОСФЕРНОГО ВОЗДУХА**

### **Аннотация**

*В статье рассматриваются методы машинного обучения для классификации индекса качества воздуха (AQI). Проведён обзор публикаций 2019–2025 гг., охватывающий классические алгоритмы (SVM, k-NN, решающие деревья), ансамблевые модели (случайный лес, градиентный бустинг), нейросетевые подходы (CNN, LSTM, трансформеры) и гибридные архитектуры со стекингом. Установлено доминирование случайного леса (40 % работ, accuracy 93–99,3 %) и градиентного бустинга (94,0–99,8 %). Рекордную точность 99,99 % достиг стекинг с мета-эвристической оптимизацией PSO-GWO. Среди нейросетевых архитектур CNN-LSTM показал наилучший баланс точности и вычислительной эффективности (94,0–97,5 %). Для преодоления дисбаланса классов стандартом де-факто стал SMOTE (прирост точности 5–25 %). Выявлены ключевые проблемы: кросс-городская деградация моделей (8–30 %), отсутствие единой международной шкалы AQI, высокие вычислительные требования трансформеров. Перспективными направлениями являются трансформеры с weather-aware attention, мультимодальные модели, федеративное обучение и объяснимый искусственный интеллект.*

**Ключевые слова:** классификация качества воздуха, индекс качества воздуха, машинное обучение, случайный лес, глубокое обучение, гибридные модели, дисбаланс классов, SMOTE

# MACHINE LEARNING METHODS FOR AIR POLLUTION LEVEL CLASSIFICATION

## Abstract

*The article examines machine learning methods for Air Quality Index (AQI) classification. A review of publications from 2019 to 2025 covers classical algorithms (SVM, k-NN, decision trees), ensemble models (random forest, gradient boosting), neural network approaches (CNN, LSTM, transformers), and hybrid architectures with stacking. Random forest dominates (40 % of studies, accuracy 93–99.3 %) alongside gradient boosting (94.0–99.8 %). The record accuracy of 99.99 % was achieved by stacking with PSO-GWO meta-heuristic optimization. Among neural architectures, CNN-LSTM showed the best balance of accuracy and computational efficiency (94.0–97.5 %). SMOTE became the de facto standard for class imbalance mitigation (5–25 % accuracy gain). Key challenges identified include cross-city model degradation (8–30 %), the absence of a unified international AQI scale, and high computational requirements of transformers. Promising directions include transformers with weather-aware attention, multimodal models, federated learning, and explainable artificial intelligence.*

Keywords: air quality classification, air quality index, machine learning, random forest, deep learning, hybrid models, class imbalance, SMOTE

## Введение

Загрязнение атмосферного воздуха остаётся приоритетной экологической угрозой XXI века: по оценкам ВОЗ, ежегодно от воздействия аэрозольных загрязнителей преждевременно погибают около 7 млн человек. Интегральным показателем совокупного влияния приоритетных загрязнителей (PM<sub>2,5</sub>, PM<sub>10</sub>, O<sub>3</sub>, NO<sub>2</sub>, SO<sub>2</sub>, CO) служит индекс качества воздуха (Air Quality Index, AQI), перевод которого в дискретные категории представляет собой многоклассовую задачу классификации с порядковой целевой переменной [1].

За 2018–2025 гг. число публикаций в области AQI-классификации возросло в 5–10 раз [2], что обусловлено накоплением массивных датасетов (CPCSV Индии [3], EPA США [4], EEA Европы [5]), совершенствованием алгоритмов и ростом вычислительной доступности глубокого обучения. Эволюция методологии прошла четыре этапа: классические алгоритмы (SVM, k-NN, решающие деревья) → ансамблевые методы (случайный лес, XGBoost, CatBoost) → глубокое обучение (CNN, LSTM, трансформеры) → гибридные архитектуры со стекингом и мета-эвристической оптимизацией. Наряду с прогрессом остаются системные проблемы: дисбаланс классов (Good и Moderate — до 80 % наблюдений, Hazardous — менее 1 %), кросс-городская деградация моделей (8–30 % потери точности) [6] и отсутствие единой международной шкалы AQI.

Настоящий обзор ставит цели: систематизировать методы машинного обучения для AQI-классификации за 2019–2025 гг.; сравнить подходы по ключевым метрикам, датасетам и архитектурам; выявить тренды, пробелы и перспективные направления.

## **Результаты исследования**

### *Классические методы машинного обучения*

Классические алгоритмы — метод опорных векторов (SVM), k ближайших соседей (k-NN) и решающие деревья (DT) — остаются востребованными благодаря интерпретируемости и низким вычислительным затратам, служа базовым уровнем для сравнения с ансамблевыми моделями. SVM демонстрирует наибольшую вариативность результатов: ассигасу колеблется от 76 % до 97,98 %, что обусловлено сильной зависимостью от типа ядра (RBF, полиномиальное, линейное) и специфики региона [7][8]. Алгоритм k-NN достигает 78–98,5 % ассигасу, однако требует обязательного масштабирования признаков и чувствителен к дисбалансу классов [9]. Решающие деревья показывают диапазон

74,8–100 % accuracy, при этом максимальные значения сопряжены с высоким риском переобучения [10].

### *Ансамблевые методы*

Ансамблевые методы систематически превосходят классические по точности и устойчивости. Случайный лес (Random Forest) является наиболее часто применяемым алгоритмом: он используется в более чем 40 % публикаций и обеспечивает 93–99,3 % accuracy благодаря механизму бэггинга и устойчивости к зашумлённым данным [1][7][11]. Среди градиентного бустинга XGBoost достигает 95,7–97,14 % accuracy [4][12], CatBoost показывает 95,6–99,8 % и отличается эффективной обработкой категориальных признаков [12][13], LightGBM достигает 94–99,75 %, являясь наиболее быстрым алгоритмом семейства [12].

Методы комбинирования разнородных моделей демонстрируют наивысшие показатели. Голосование (k-NN + RF + SVM) достигает 99,68–99,84 % accuracy [14]. Стекинг (RF + XGBoost + Extra Trees) с SMOTE-балансировкой показывает 96,45 % [4]. Наибольший результат — 99,99 % accuracy и F1-мера 0,9999 — достигнут стекингом (RF + Extra Trees + MLP) с оптимизацией гибридным алгоритмом PSO-GWO [1].

**Таблица 1. Сравнение классических и ансамблевых методов**

| Метод             | Accuracy, % | Ключевая особенность                        |
|-------------------|-------------|---------------------------------------------|
| SVM               | 76–97,98    | Зависимость от ядра и региона               |
| Random Forest     | 93–99,3     | Лидер по частоте (40 %+); устойчив к шуму   |
| XGBoost           | 95,7–97,14  | Оптимальное соотношение точности и скорости |
| CatBoost          | 95,6–99,8   | Эффективен с категориальными признаками     |
| Stacking / Voting | 96,45–99,99 | Наивысшая точность при мета-оптимизации     |

Анализ таблицы показывает, что ансамблевые методы обеспечивают accuracy выше 95 % независимо от региона, тогда как классические алгоритмы характеризуются значительным разбросом результатов (до 22 п.п. для SVM). Наибольшую точность демонстрируют стекинг-схемы с оптимизацией PSO-GWO, что свидетельствует о синергетическом эффекте комбинирования разнородных моделей.

### *Нейросетевые методы*

Нейросетевые подходы охватывают архитектуры от многослойных перцептронов (MLP) до глубоких моделей на базе CNN, LSTM и трансформеров. MLP остаётся базовым инструментом: трёхслойная архитектура с 128 нейронами достигает ~99 % accuracy на сбалансированных данных (SMOTE), тогда как на несбалансированных показатель снижается до 93,5 % [9]. Основное ограничение MLP — отсутствие механизмов учёта пространственно-временных зависимостей.

Свёрточные нейронные сети применяются в конфигурациях 1D-CNN для временных рядов и 2D-CNN для изображений атмосферы: 2D-CNN на базе ResNet достигает 97,36 % accuracy [15]. Гибридная архитектура CNN-LSTM, где свёрточные слои извлекают локальные признаки, а LSTM моделирует временную динамику, является доминирующей схемой: DCNN-LSTM на данных Дели позволил достичь 97,48 % accuracy и AUC = 0,97 [15].

Трансформерные архитектуры обеспечивают учёт глобальных зависимостей за счёт механизма внимания. AQIFormer — первый трансформер для классификации AQI — достиг 89,96 % accuracy, а механизм адаптации под метеоусловия снизил кросс-городскую деградацию до 8,29 % [6]. Каскадная архитектура Transformer-CNN-BiLSTM показала 95,24 % accuracy [6]. Несмотря на преимущества, трансформеры требуют значительных вычислительных ресурсов.

### *Гибридные модели*

Гибридные модели, комбинирующие нейросетевые и ансамблевые компоненты, демонстрируют наивысшую точность. Двухстадийная модель CBLA (CNN-BiLSTM-Attention + XGBoost) достигла  $R^2 = 0,928$  при  $RMSE = 10,29$  мкг/м<sup>3</sup> [1]. Стекинг (RF + Extra Trees + MLP) с PSO-GWO показывает рекордные 99,99 % [1]. Модель CNN+Autoencoder+BiLSTM с оптимизатором PUGWO достигла  $R^2 = 0,961$  [16].

**Таблица 2. Сравнение нейросетевых и гибридных методов**

| Метод             | Accuracy / $R^2$              | Ключевая особенность                    |
|-------------------|-------------------------------|-----------------------------------------|
| MLP               | 93,5–99 %                     | Простой, требует feature engineering    |
| CNN (1D/2D)       | 94–97,36 %                    | Иерархическое извлечение признаков      |
| CNN-LSTM          | 94–97,48 %                    | Доминирующая DL-архитектура             |
| Transformer       | 89,96–95,24 %                 | Глобальное внимание; высокие требования |
| Гибриды (стекинг) | 99,99 % / $R^2 = 0,928–0,961$ | Наивысшая точность                      |

Анализ таблицы демонстрирует, что CNN-LSTM-гибриды обеспечивают наилучший баланс между точностью и вычислительной эффективностью, тогда как трансформеры уступают из-за высоких требований к данным. Наивысшие показатели достигаются гибридными моделями с мета-эвристической оптимизацией.

#### *Работа с дисбалансом классов*

Дисбаланс классов является фундаментальной проблемой AQI-классификации: категории Good и Moderate составляют до 80 % наблюдений, тогда как Hazardous — менее 1 % [10]. SMOTE стал стандартом де-факто для oversampling: прирост accuracy составляет 5–25 %, F1-мера возрастает с 0,805 до 0,9999 [1][17]. SMOTE-Tomek и SMOTE-ENN дополняют базовый метод удалением перекрывающихся пар и шумовых образцов. Borderline-SMOTE фокусируется на граничных экземплярах, повышая TP rate на 45,2 % [18].

ADASYN с адаптивной генерацией показывает сопоставимые результаты (F1 = 0,939 против 0,934 у SMOTE) [19].

Среди функций потерь Focal Loss ( $\gamma = 2,0$ ) фокусируется на «трудных» примерах, решая проблему исчезающих градиентов для миноритарных классов [20]. Cost-sensitive learning через взвешивание классов (inverse frequency, scale\_pos\_weight) позволяет обходиться без изменения данных, что является преимуществом для больших датасетов [10].

**Таблица 3. Методы работы с дисбалансом классов**

| Метод                   | Прирост ассигасу | Ключевое преимущество                 |
|-------------------------|------------------|---------------------------------------|
| SMOTE                   | 5–25 %           | Стандарт де-факто; простая реализация |
| SMOTE-Tomek/ENN         | 10–20 %          | Удаление шумов и перекрытий           |
| Borderline-SMOTE        | TP rate +45,2 %  | Фокус на граничных экземплярах        |
| Focal Loss              | F1 до 0,9999     | Работа с hard examples в DL           |
| Cost-sensitive learning | 3–8 %            | Без изменения данных                  |

#### *Сравнительный анализ*

Сводное сравнение методов по диапазонам ассигасу демонстрирует чёткую иерархию. Классические алгоритмы показывают 68–100 % с большим разбросом. Ансамблевые методы обеспечивают стабильно высокие результаты (93–99,8 %). Нейросетевые подходы демонстрируют 89–99 %, при этом CNN-LSTM лидирует среди мономодельных архитектур. Гибридные модели достигают 99,99 %, однако требуют значительных вычислительных ресурсов и экспертной настройки.

По регионам доминирует Индия (датасет CPRCB [3], 15+ публикаций), за которой следуют США (EPA [4]), Китай, Таиланд [14] и Корея [7]. Отсутствие единого бенчмарка затрудняет прямое сравнение: шкала EPA США содержит 6 классов (0–500), CPRCB Индии — 6 классов с другими порогами, SAQI Европы

— 5 классов (0–100+). Ключевые нерешённые проблемы включают кросс-городскую деградацию (8–30 %), интерпретируемость моделей (SHAP, LIME, Grad-CAM) и вычислительную сложность трансформеров.

**Таблица 4. Сводное сравнение методов**

| Метод        | Accuracy, % | Ключевое преимущество              | Ограничение                   |
|--------------|-------------|------------------------------------|-------------------------------|
| Классические | 68–100      | Интерпретируемость, скорость       | Нестабильность, переобучение  |
| Ансамблевые  | 93–99,8     | Устойчивость, высокая точность     | Сложность настройки           |
| MLP          | 93,5–99     | Простота реализации                | Нет учёта временной структуры |
| CNN-LSTM     | 94–97,5     | Пространственно-временные паттерны | Требует больших данных        |
| Гибриды      | до 99,99    | Наивысшая точность                 | Вычислительная сложность      |

**Заключение**

Проведённый обзор публикаций 2019–2025 гг. демонстрирует устойчивый прогресс в применении методов машинного обучения для классификации уровней загрязнения атмосферного воздуха. Случайный лес доминирует в литературе (40 % работ, accuracy 93–99,3 %), градиентный бустинг (XGBoost, CatBoost, LightGBM) обеспечивает 94,0–99,8 %. Среди нейросетевых архитектур CNN-LSTM показал наилучший баланс точности и вычислительной эффективности (94,0–97,5 %). Гибридные модели со стекингом и мета-эвристической оптимизацией достигают рекордной точности 99,99 % [1]. Для преодоления дисбаланса классов стандартом де-факто стал SMOTE (прирост 5–25 %).

Перспективными направлениями являются: трансформеры с механизмами weather-aware attention для снижения кросс-городской деградации; мультимодальные модели, комбинирующие сенсорные данные, спутниковые



изображения и метеорологические параметры; федеративное обучение для обеспечения конфиденциальности при межрегиональном обучении; объяснимый искусственный интеллект (SHAP, LIME) для повышения доверия к моделям в экологических приложениях; edge-deployment оптимизированных моделей для IoT-сенсоров мониторинга качества воздуха.

## Литература

1. Fahim A. et al. Enhancing air quality index classification based on ensemble machine learning techniques // Engineering, Technology and Applied Science Research. – 2025. – Vol. 15. – No. 6. – P. 29325–29333.
2. Agbehadji I.E. et al. Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction // Atmosphere. – 2024. – Vol. 15. – No. 11. – P. 1352.
3. Rao R. Air Quality Data in India (2015–2020) // Kaggle. – URL: <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india> (дата обращения: 13.07.2026).
4. Rao R.S. et al. Multimodal imputation-based stacked ensemble for prediction and classification of air quality index in Indian cities // Computers and Electrical Engineering. – 2024. – Vol. 114. – P. 109098.
5. Air Quality Data // European Environment Agency. – URL: <https://www.eea.europa.eu/en/datahub> (дата обращения: 13.07.2026).
6. AQIFormer: a transformer-based multi-view architecture for cross-city air quality classification // Proceedings of ICVGIP 2025. – ACM, 2025.
7. Choi Y. et al. Utilizing machine learning-based classification models for tracking air pollution sources: a case study in Korea // Aerosol and Air Quality Research. – 2024. – Vol. 24. – P. 230222.
8. Saxena A., Shekhawat S. Ambient air quality classification by grey wolf optimizer based support vector machine // Journal of Environmental and Public Health. – 2017. – Vol. 2017. – P. 3131083.
9. Elabd E. et al. Air quality index (AQI) classification based on hybrid particle swarm and grey wolf optimization with ensemble machine learning model // Scientific Reports. – 2025.
10. Jayadi B.V. et al. Air quality index classification for imbalanced data using machine learning approach // Sistemasi. – 2024. – Vol. 13. – No. 3. – P. 951–958.
11. El Furqany N. Optimizing air quality index classification using multiple machine learning models and oversampling techniques // International Journal of Artificial Intelligence and Machine Learning. – 2025. – Vol. 3. – No. 2.
12. Patil M.C., Katke S.M. Air quality classification and forecasting: insights from multi-source environmental data // JETIR. – 2024. – Vol. 11. – No. 12.
13. Idroes G.M. et al. Urban air quality classification using machine learning approach to enhance environmental monitoring // Leuser Journal of Environmental Studies. – 2023. – Vol. 1. – No. 2. – P. 62–68.
14. Lertnilkarn S. Forecasting air quality index in Thailand using ensemble method // Proceedings of IEEE Conference on Computer Science and Engineering. – 2022.
15. Barthwal A., Goel S. Advancing air quality prediction models in urban India: a deep learning approach integrating DCNN and LSTM architectures // Modeling Earth Systems and Environment. – 2024. – Vol. 10. – No. 2. – P. 2935–2955.
16. Mehmood F. et al. Vision-AQ: explainable multi-modal deep learning for air pollution classification in smart cities // Mathematics. – 2025. – Vol. 13. – No. 18. – P. 3017.

17. Haq M.A. SMOTEDNN: a novel model for air pollution forecasting and AQI classification // Computers, Materials & Continua. – 2022. – Vol. 71. – No. 1.
18. Han H., Wang W.Y., Mao B.H. A new over-sampling method in imbalanced data sets learning: Borderline-SMOTE // Lecture Notes in Computer Science. – Springer, 2005. – Vol. 3644.
19. Zhao X. et al. A deep recurrent neural network for air quality classification // Journal of Information Hiding and Multimedia Signal Processing. – 2018. – Vol. 9. – No. 2. – P. 346–354.
20. Lin T.-Y. et al. Focal loss for dense object detection // Proceedings of IEEE International Conference on Computer Vision (ICCV). – 2017. – P. 2980–2988.