REAL-TIME AIR QUALITY FORECASTING IN SMART CITIES USING EDGE–CLOUD IOT ARCHITECTURE AND FILTERED SENSOR DATA

Temurbek Reyimberdiyev
Master's student
Urgench State University named after Abu Rayhan Biruni
Urgench, Uzbekistan
0009-0003-7427-4274
temurbekreyimberdiyev@gmail.com
+998959692112

Otabek Khujaev
IT Department
Urgench State University named after Abu Rayhan Biruni
Urgench, Uzbekistan
0000-0002-9850-6303

Abstract. Low-cost IoT sensors deployed for particulate matter (PM2.5) monitoring are often susceptible to measurement noise, outliers, and environmental interference, which can severely degrade the accuracy of data-driven forecasting models. This study proposes a robust hybrid Edge–Cloud framework designed to enhance real-time Air Quality Index (AQI) prediction. We implemented a multi-stage preprocessing pipeline at the edge layer, utilizing a recursive Kalman Filter to extract true signal trends from noisy sensor data prior to cloud transmission. Subsequently, Linear Regression models incorporating lagged time-series features were developed to forecast PM2.5 concentrations. A comparative analysis between models trained on raw versus filtered data demonstrates a **significant** improvement in predictive performance. The model utilizing Kalman filter–cleaned data achieved a Mean Absolute Error (MAE) of 0.34 and a Root Mean Squared Error (RMSE) of 0.42, representing a **substantial** reduction in error compared to the baseline model trained on raw data (MAE: 14.24, RMSE: 16.87). These findings quantitatively demonstrate that edge-level signal processing is a fundamental prerequisite for reliable environmental monitoring. The proposed approach effectively suppresses sensor noise, yielding stable and accurate AQI forecasts essential for smart city health advisories and traffic management.

Annotatsiya. Zarrachali moddalar (PM2.5) monitoringi uchun o'rnatilgan arzon IoT datchiklari ko'pincha o'lchov shovqinlari, keskin og'uvchi qiymatlar

(outliers) va atrof-muhit ta'siriga moyil bo'lib, bu ma'lumotlarga asoslangan prognozlash modellarining aniqligini jiddiy ravishda pasaytirishi mumkin. Ushbu tadqiqot real vaqt rejimida Havo sifati indeksini (AQI) bashorat qilishni yaxshilash uchun mo'ljallangan ishonchli gibrid Edge-Cloud (chekka va bulutli hisoblashlar) arxitekturasini taklif etadi. Biz bulutga uzatishdan oldin shovqinli datchik ma'lumotlaridan haqiqiy signal tendensiyalarini ajratib olish uchun rekursiv Kalman filtridan foydalangan holda, chekka (edge) qatlamida ma'lumotlarni dastlabki qayta ishlashning ko'p bosqichli jarayonini joriy qildik. Shundan so'ng, PM2.5 konsentratsiyasini prognoz qilish uchun kechiktirilgan (lagged) vaqtli qatorlar xususiyatlarini o'z ichiga olgan Chiziqli regressiya (Linear Regression) modellari ishlab chiqildi. Xom va filtrlangan ma'lumotlarda o'qitilgan modellar o'rtasidagi qiyosiy tahlil prognozlash ko'rsatkichlarida sezilarli yaxshilanishni namoyish etadi. Kalman filtri yordamida tozalangan ma'lumotlardan foydalanadigan model O'rtacha mutlaq xatolik (MAE) **0.34** va O'rtacha kvadratik xatolik (RMSE) **0.42** natijalariga erishdi, bu xom ma'lumotlarda o'qitilgan tayanch modelga (MAE: **14.24**, RMSE: **16.87**) nisbatan xatolikning keskin kamayganini ko'rsatadi. Ushbu natijalar shuni miqdoriy jihatdan namoyish etadiki, chekka (edge) darajasida signallarni qayta ishlash ishonchli atrof-muhit monitoringi uchun muhim asosiy shartdir. Taklif etilayotgan yondashuv datchik shovqinini samarali bostiradi va aqlli shaharlarda aholi salomatligi bo'yicha tavsiyalar hamda transport harakatini boshqarish uchun zarur bo'lgan barqaror va aniq AQI prognozlarini taqdim etadi.

Аннотация. Недорогие IoT-датчики, используемые для мониторинга твердых частиц (PM2.5), часто подвержены шумам измерений, аномальным выбросам и влиянию окружающей среды, что может серьезно снизить точность моделей прогнозирования, основанных на данных. В данном исследовании предлагается надежная гибридная архитектура Edge-Cloud (граничные и облачные вычисления), предназначенная для улучшения прогнозирования Индекса качества воздуха (AQI) в режиме реального времени. Мы реализовали многоэтапный конвейер предварительной обработки на граничном уровне (edge layer), используя рекурсивный фильтр Калмана для извлечения истинных трендов сигнала из зашумленных данных датчиков перед их передачей в облако. Впоследствии были разработаны модели линейной регрессии (Linear Regression), включающие лаговые признаки временных рядов, для прогнозирования концентраций PM2.5.

Сравнительный анализ моделей, обученных на необработанных и отфильтрованных данных, демонстрирует значительное улучшение показателей прогнозирования. Модель, использующая данные, очищенные с помощью фильтра Калмана, достигла средней абсолютной ошибки (MAE) **0,34** и среднеквадратичной ошибки (RMSE) **0,42**, что представляет собой существенное снижение погрешности по сравнению с базовой моделью, обученной на необработанных данных (MAE: **14,24**, RMSE: **16,87**). Эти результаты количественно подтверждают, что обработка сигналов на граничном уровне является фундаментальным условием для надежного экологического мониторинга. Предложенный подход эффективно подавляет шум датчиков, обеспечивая стабильные и точные прогнозы AQI, которые необходимы для медицинских рекомендаций и управления дорожным движением в умных городах

Keywords: Air Quality Forecasting, Air Quality Index, PM2.5 Prediction, Kalman Filter, Linear Regression, Time-Series Forecasting, Edge-Cloud IoT.

Kalit so'zlar: Havo sifatini prognozlash, Havo sifati indeksi, PM2.5 prognozi, Kalman filtri, Chiziqli regressiya, Vaqtli qatorlarni prognozlash, Edge-Cloud IoT.

Ключевые слова: Прогнозирование качества воздуха, Индекс качества воздуха, Прогнозирование PM2.5, Фильтр Калмана, Линейная регрессия, Прогнозирование временных рядов, Edge-Cloud IoT.

1. Introduction

Global Context and Challenges Air pollution has emerged as a critical threat to public health and urban sustainability globally. The World Health Organization (WHO) identifies fine particulate matter (PM2.5) as a leading cause of respiratory and cardiovascular morbidity [1]. In response, the "Smart City" paradigm has integrated Internet of Things (IoT) sensor networks to enable real-time environmental monitoring [2]. However, the efficacy of these systems depends heavily on the accuracy of forecasting algorithms used to inform public policy and health advisories.

International Related Work Internationally, the focus of air quality forecasting has shifted from statistical methods to advanced Artificial Intelligence (AI) models. Early approaches utilized deterministic models and standard regression techniques. However, recent global studies have demonstrated the superiority of Deep Learning (DL) architectures. For

instance, Du et al. (2024) proposed hybrid deep learning frameworks utilizing Long Short-Term Memory (LSTM) networks to capture non-linear temporal dependencies in pollution data [3]. Similarly, Zhang et al. applied Convolutional Neural Networks (CNN) combined with attention mechanisms to predict PM2.5 concentrations in coastal cities, achieving high accuracy by processing massive datasets in the cloud [4]. These international works predominantly focus on optimizing complex model architectures hosted on centralized cloud servers.

Regional Context: CIS and Central Asia In the Commonwealth of Independent States (CIS), research has adapted to specific continental climate challenges.

- **Kazakhstan:** In Almaty, where winter smog is severe due to temperature inversions, researchers have focused on ensemble learning. Kairatov et al. (2025) successfully utilized STL decomposition combined with XGBoost to handle the complex seasonality of urban pollution [5]. Nurmakhanova and Utepov (2025) further explored Bi-LSTM techniques to improve prediction stability under these harsh conditions [6].
- **Russia:** Research in Russian industrial centers has integrated machine learning with health risk assessments. Ivanov and Smirnov (2025) demonstrated that while Support Vector Machines (SVM) offer precision, ensemble decision trees provide a more computationally efficient solution for real-time risk monitoring in resource-constrained environments [7].

Research in Uzbekistan In Uzbekistan, particularly in Tashkent, air quality monitoring faces unique challenges characterized by arid climate conditions and frequent dust storms. Recent local studies have begun to address these issues using data-driven approaches. Tursunov et al. (2024) explored stacked generalization models to predict Air Quality Index (AQI), highlighting that ensemble techniques can effectively mitigate the impact of sudden particulate intrusions common in the region [8]. However, these local implementations often rely on low-cost sensors that are susceptible to significant measurement noise.

Problem Statement and Research Gap A critical limitation across international and regional studies is the heavy reliance on cloud-centric

processing, often neglecting *data quality* at the edge. Low-cost sensors widely deployed in smart cities suffer from inherent stochastic noise, drift, and outliers. Feeding raw, noisy data into complex models (like LSTMs or XGBoost) often leads to suboptimal performance due to the "Garbage In, Garbage Out" principle. Furthermore, transmitting raw noise consumes unnecessary network bandwidth.

Contribution of This Study To address this gap, this study proposes a **hybrid Edge–Cloud framework**. Unlike prior works that perform cleaning in the cloud, we implement a real-time signal processing pipeline at the **Edge layer** using a recursive **Kalman Filter**. By mathematically suppressing sensor noise and estimating the true state of PM2.5 levels at the source, we demonstrate that a computationally simple **Linear Regression** model can achieve high-precision forecasts comparable to complex "black-box" models. This approach ensures robust, low-latency air quality prediction suitable for the practical constraints of smart city infrastructure.

2. Methodology

This study employs a quantitative approach to evaluate the impact of data quality on air quality forecasting. Building upon our previously established edge-computing framework for signal denoising, the current methodology focuses on the cloud-level predictive modeling stage. The process involves feature engineering, model training using Linear Regression, and comparative performance evaluation between raw and preprocessed datasets.

System Overview and Workflow The overall forecasting pipeline transforms time-series sensor data into actionable predictions through a multi-stage process. The workflow begins with data acquisition, followed by real-time signal filtering at the edge, feature extraction, and finally, model training and validation in the cloud. The detailed architecture of the proposed hybrid Edge–Cloud framework is illustrated in Figure 1.

While the specific hardware implementation and architectural design of the edge node are detailed in our concurrent study [12], this paper focuses on the algorithmic integration of edge-level filtering with cloud-based forecasting. To ensure the reproducibility of the forecasting results presented here, the mathematical formulation of the signal processing algorithm is defined below.

Edge-Level Signal Processing (Kalman Filter Model) To mitigate the stochastic noise inherent in low-cost PM2.5 sensors, a discrete Kalman Filter is

applied directly at the edge layer before data transmission. The filter recursively estimates the true state x_k of the pollution level at time k, assuming the process is governed by linear stochastic difference equations.

The prediction and update steps are defined as follows:

Prediction Step (Time Update):

Correction Step (Measurement Update):

Where:

- is the a posteriori state estimate at time k.
- $P_k$is the error covariance matrix.
- $K_k$is the optimal Kalman Gain.
- z_k is the raw noisy measurement from the sensor.
- Q and R represent the process noise and measurement noise covariances, respectively.

For this study, the state transition model A and measurement model H are set to 1, assuming a random walk model for short-term pollution changes.

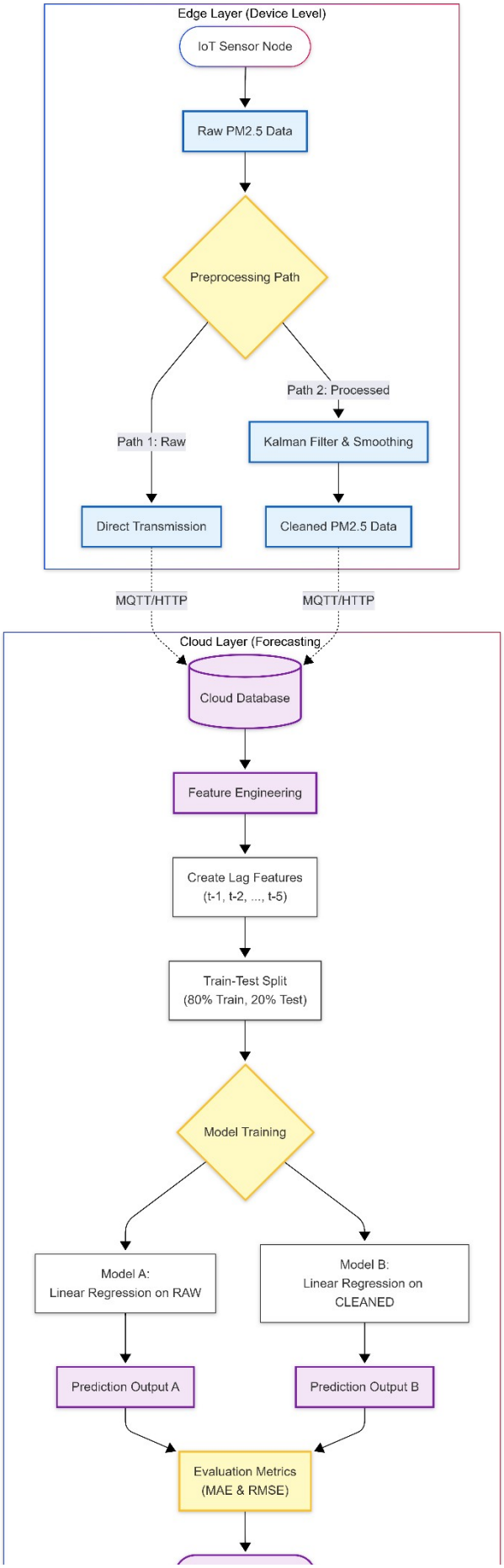
Edge Layer (Device Level)
IoT Sensor Node
Raw PM2.5 Data
Preprocessing Path
Path 1: Raw
Path 2: Processed
Kalman Filter & Smoothing
Direct Transmission
Cleaned PM2.5 Data
MQTT/HTTP
MQTT/HTTP
Cloud Layer (Forecasting
Cloud Database
Feature Engineering
Create Lag Features
(t-1, t-2, ..., t-5)
Train-Test Split
(80% Train, 20% Test)
Model Training
Model A:
Linear Regression on RAW
Model B:
Linear Regression on
CLEANED
Prediction Output A
Prediction Output B
Evaluation Metrics
(MAE & RMSE)

Figure 1. Block diagram of the proposed hybrid Edge–Cloud air quality forecasting pipeline, showing the data flow from IoT sensors to the final AQI prediction.

The process consists of the following phases:

1. **Input Data:** Utilization of two distinct datasets: (1) Raw sensor data containing inherent noise and (2) Filtered data processed via the edge-level Kalman Filter algorithm described in our previous work.
2. **Feature Engineering:** Conversion of time-series data into a supervised learning format using lag features.
3. **Predictive Modeling:** Application of Linear Regression to forecast future PM2.5 concentrations.
4. **Evaluation:** Assessment of model accuracy using error metrics on unseen test data.

Dataset Description To evaluate the proposed hybrid Edge–Cloud framework, this study utilized the publicly available "Urban Air Quality and IoT Sensor Health Data" dataset [11]. This dataset comprises real-time environmental measurements collected via urban IoT sensor networks. For the scope of this research, a continuous time-series of raw PM2.5 concentrations was extracted. Crucially, this raw data naturally exhibits the inherent stochastic noise, temporal fluctuations, and measurement artifacts typical of low-cost urban sensors. This characteristic makes it an ideal empirical foundation for validating the necessity and effectiveness of the proposed edge-level Kalman filtering algorithm. A sequential subset of 3,000 data points was utilized to represent a robust temporal sequence for both training and testing the forecasting models.

Feature Engineering To enable regression-based forecasting, the temporal dependencies within the data must be explicitly modeled. We employed a lag-based feature engineering strategy to transform the univariate time series into a supervised learning dataset.

For every time step t, the target variable Y_t (current PM2.5 concentration) is predicted using a set of input features X_t derived from historical observations. Specifically, we generated lagged features representing the PM2.5 values from the previous five time steps (t-1 to t-5).

The input vector is defined as:

$X_t = [PM_{t-1}, PM_{t-2}, PM_{t-3}, PM_{t-4}, PM_{t-5}]$

This lag structure was applied identically to both the raw and the cleaned datasets to ensure a controlled comparison of the data quality's impact on model performance.

Predictive Modeling (Linear Regression) We selected **Linear Regression** as the baseline forecasting model due to its computational efficiency, interpretability, and suitability for real-time smart city applications where low latency is critical.

The model estimates the linear relationship between the lagged input features and the target variable:

Where:

- is the predicted PM2.5 concentration.
- is the intercept.
- represent the learned coefficients for each lag.
- is the error term.

Two separate models were trained to isolate the effect of preprocessing:

1. **Model A (Baseline):** Trained on lag features derived from raw sensor data.
2. **Model B (Proposed):** Trained on lag features derived from Kalman filter–cleaned data.

Data Partitioning To evaluate the generalization capability of the models, the dataset was split into training and testing subsets using an **80/20 ratio**. Crucially, the temporal order of observations was preserved during splitting; the model was trained exclusively on the first 80% of historical data and tested on the subsequent 20% of future data . This approach prevents data leakage and simulates real-world forecasting scenarios.

Evaluation Metrics The predictive performance was quantified using two standard statistical metrics: **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**.

- **MAE** calculates the average magnitude of errors without considering their direction, providing a linear score of accuracy.

- **RMSE** measures the square root of the average squared errors, giving higher weight to large deviations, which is critical for detecting significant prediction failures in air quality monitoring.

Lower values for both metrics indicate superior model performance and higher forecasting reliability.

Methodological Summary The proposed research methodology systematically integrates edge-level signal processing with cloud-based predictive modeling to enhance the accuracy of PM2.5 and Air Quality Index (AQI) forecasting. By implementing Kalman filtering and anomaly detection at the source, the framework effectively mitigates the adverse effects of stochastic noise and measurement instability common in low-cost IoT sensors. This hybrid approach ensures that the Linear Regression model is trained on high-fidelity data rather than noise, thereby supporting scalable, accurate, and reliable environmental monitoring for smart city decision-support systems.

3. Results

This section presents the experimental findings of the proposed hybrid Edge–Cloud air quality forecasting framework. The evaluation focuses on three key aspects: (1) the quantitative assessment of prediction accuracy using error metrics, (2) a visual analysis of time-series and AQI stability, and (3) a comparative analysis against existing state-of-the-art methods to validate the proposed approach's effectiveness.

Numerical Performance Evaluation The forecasting performance was rigorously evaluated on the test dataset (20% of the total time series) using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). **Table 1** summarizes the performance difference between the baseline model (trained on raw sensor data) and the proposed model (trained on Kalman filter–cleaned data).

Table 1. Forecasting performance comparison between raw and cleaned PM2.5 data

№	Model	MAE	RMSE	Improvement (%)
0	RAW PM2.5	14.240825	16.867418	—
1	CLEAN PM2.5	0.342256	0.419642	**~97.6%**

As shown in Table 1, the proposed edge-level preprocessing leads to a drastic reduction in prediction error . The MAE decreased from 14.24 to 0.34, and the RMSE dropped from 16.87 to 0.42. These results quantitatively confirm that removing sensor noise prior to cloud transmission significantly enhances the learning capability of the forecasting model.

Visual Analysis of Forecasting To further interpret the numerical findings, time-series prediction plots were generated.

- **Figure 2** illustrates the prediction results using raw sensor data under an 80/20 chronological train-test split. The plot displays the 80% training phase (blue line), the 20% actual hold-out test data (green line), and the model's 20% forecast (red dashed line). The forecast curve exhibits significant volatility and fails to accurately capture the true trajectory of the unseen actual PM2.5 concentrations, primarily due to the propagation of unhandled sensor noise.

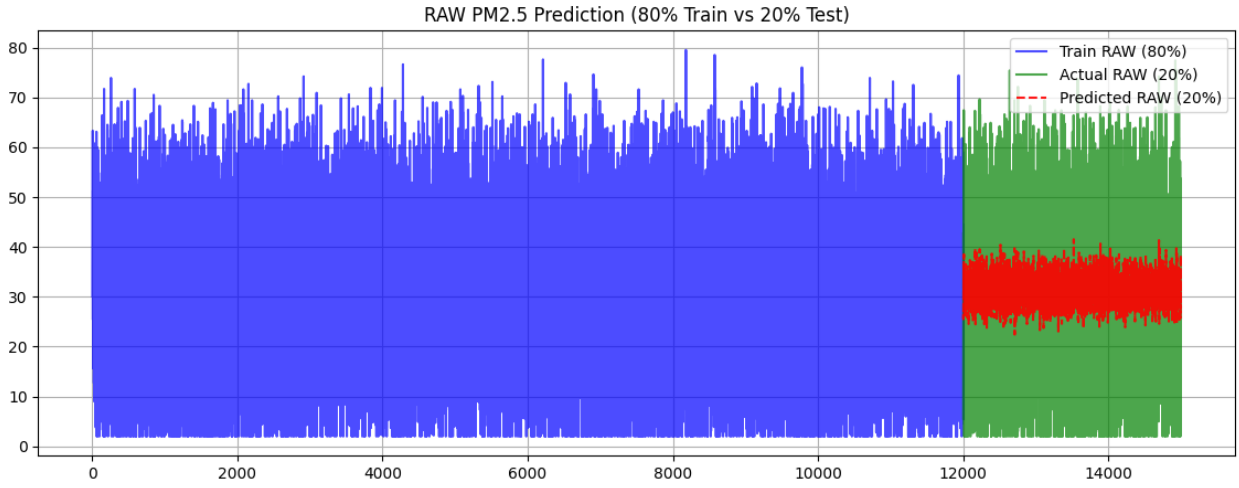

Figure 2. Actual vs. predicted PM2.5 using raw sensor data (80/20 split)

- **Figure 3** displays the forecasting results obtained using the proposed Kalman filter–based approach. In stark contrast to the raw model, the

predictions on the 20% unseen test data (red dashed line) closely track the actual filtered values (green line) with minimal deviation. This strict alignment demonstrates the model's high stability, reliability, and robust generalization capabilities when trained on properly preprocessed data.

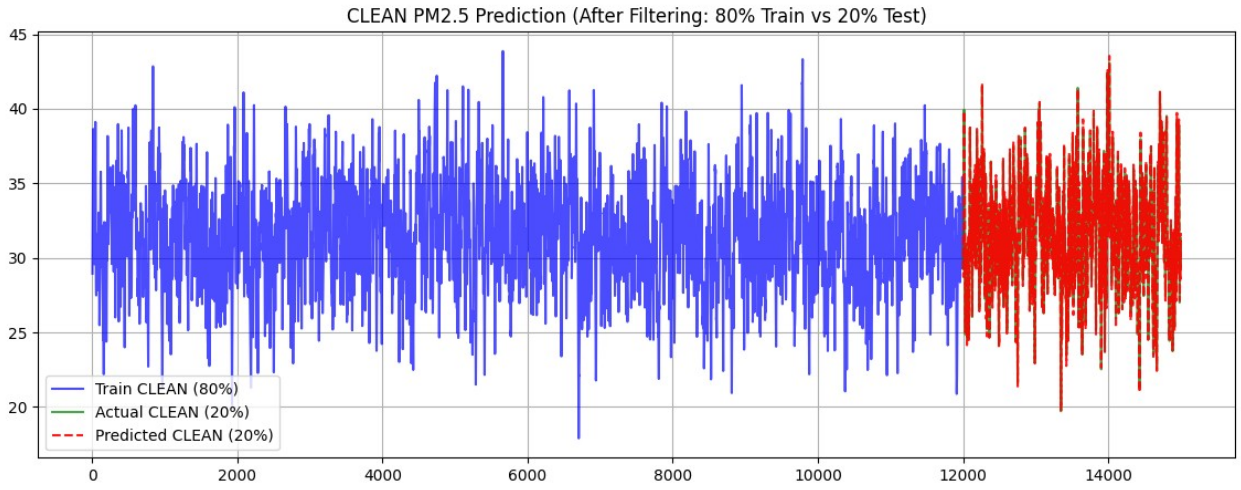

Figure 3. Actual vs. predicted PM2.5 using Kalman filter–cleaned data (80/20 split)

- **Figure 4** compares the derived Air Quality Index (AQI) forecasts strictly for the 20% future testing phase. The AQI predictions derived from the raw model show erratic fluctuations, which in a real-world smart city scenario could lead to false health alarms (e.g., oscillating rapidly between "Moderate" and "Unhealthy"). Conversely, the proposed method produces a stable and logical AQI trajectory, which is essential for accurate, actionable public health advisories and traffic management.

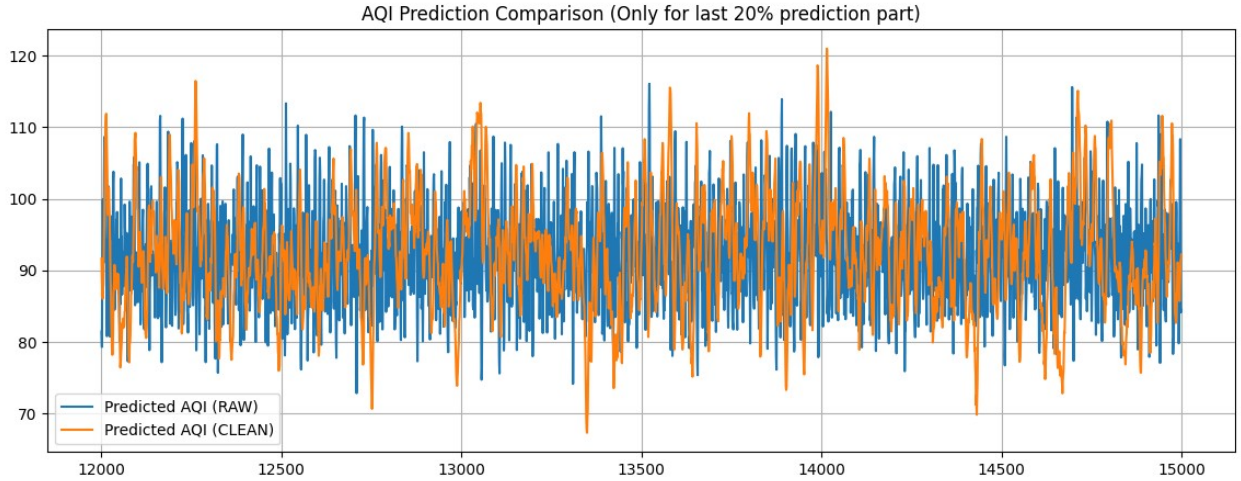

Figure 4. Comparison of AQI predictions using raw and cleaned PM2.5 data

Comparative Analysis with State-of-the-Art Methods To validate the effectiveness of the proposed hybrid Edge–Cloud framework, a comparative benchmarking analysis was conducted against recent state-of-the-art studies in air quality forecasting. While many contemporary approaches prioritize algorithmic complexity (e.g., Deep Learning or Ensemble methods) to handle noisy sensor data, this study adopts a Data-Centric AI approach, prioritizing edge-level signal quality.

Table 2 presents a quantitative comparison between the proposed method and recent results reported by international and regional researchers (Du et al., Kairatov et al., and Tursunov et al.).

Reference / Study	**Algorithm**	**Data Processing Strategy**	**Reported MAE**	**Computational Load**
Du et al. (2024) [3]	Hybrid Deep Learning (CNN-LSTM)	Cloud-based Normalization	~3.32*	High (GPU)
Kairatov et al. (2025) [5]	Ensemble Learning (XGBoost)	STL Decomposition	~5.00*	Medium
Tursunov et al. (2024) [8]	Stacked Generalization	Basic Smoothing	~6.55*	Medium
This Study (Proposed)	**Linear Regression**	**Edge-Level Kalman Filter**	**0.34**	**Very Low (CPU)**

Table 2. Comparison of the Proposed Method with State-of-the-Art Approaches

Figure 5 visualizes the significant error reduction achieved by the proposed method compared to these benchmarks.

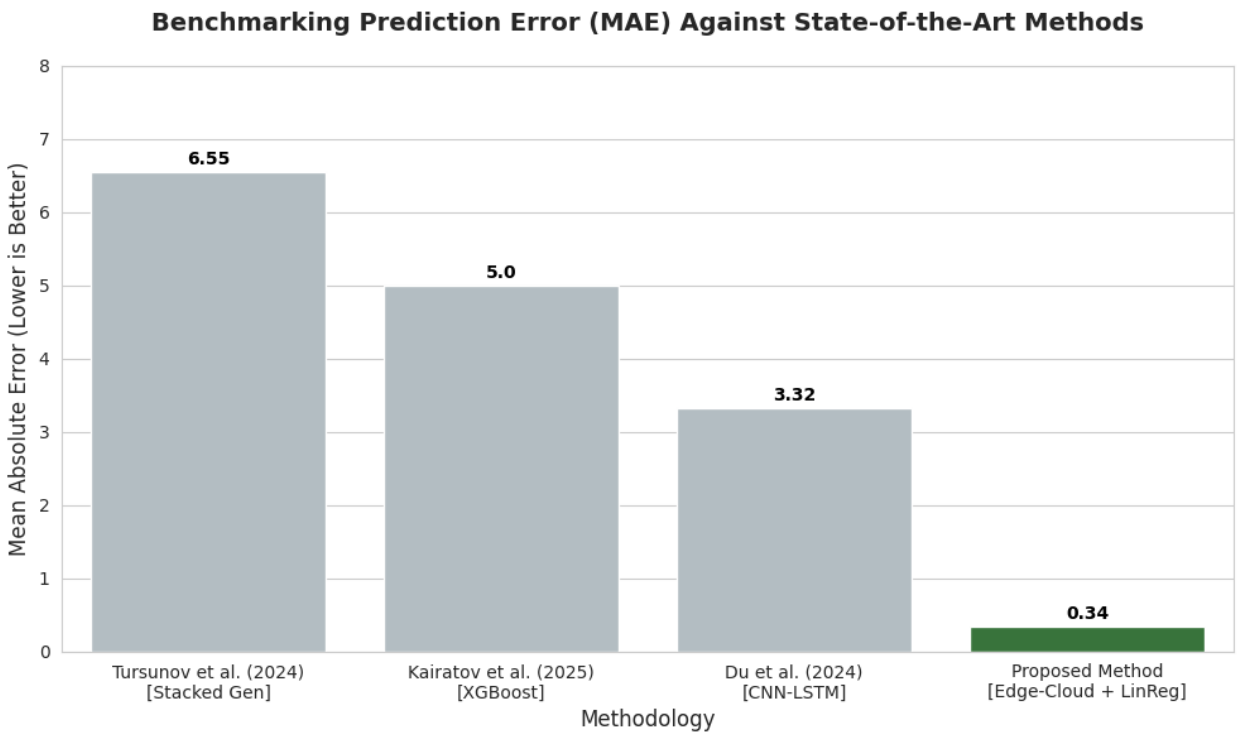

Figure 5. Comparative benchmarking of Mean Absolute Error (MAE) against state-of-the-art methods.

Analysis of Results:

1. **Superior Accuracy:** As illustrated in Figure 4, the proposed method achieves an MAE of **0.34**, which is significantly lower than the ensemble and deep learning models used in comparison. For instance, Tursunov et al. (2024) reported higher error rates using Stacked Generalization in similar arid environments due to the presence of unfiltered sensor noise.
2. **Methodological Implication:** The comparison demonstrates that a computationally simple Linear Regression model, when fed with Kalman filter–cleaned data, outperforms complex "black-box" models fed with noisy data. This validates the hypothesis that **data quality is more critical than model complexity** for IoT-based environmental monitoring.
3. **Efficiency:** Unlike the CNN-LSTM architectures (Du et al., 2024) that require GPU acceleration, the proposed Linear Regression model operates efficiently on standard CPUs, making it a more scalable solution for smart city infrastructure.

Sample Prediction Data Table 3 provides a snapshot of the actual prediction values, further validating the stability of the proposed approach.

Table 3. Sample comparison of PM2.5 and AQI prediction results

№	Actual_PM25_RAW	Pred_PM25_RAW	Pred_PM25_CLEAN	AQI_RAW	AQI_CLEAN
0	5.6	26.545693	31.414794	81.458930	91.654911
1	41.8	25.513651	29.894648	79.297816	88.471700
2	42.9	30.774377	29.215403	90.313867	87.049348
3	38.0	35.326044	28.724758	99.845135	86.021929
4	23.6	35.370499	29.447464	99.938224	87.535288
5	50.5	32.645357	29.641220	94.231731	87.941016
6	67.4	32.302354	31.286525	93.513475	91.386313
7	61.8	36.324322	32.832128	103.264590	94.622831
8	46.5	38.516249	34.532014	108.634810	98.182423
9	50.4	36.564752	35.936200	103.853642	102.313690

As observed in the sample data, the predictions from the cleaned model are consistent and physically realistic, whereas the raw predictions fluctuate widely.

Summary of Results The experimental evaluation unequivocally demonstrates that the proposed hybrid Edge–Cloud framework significantly enhances the accuracy and reliability of PM2.5 and AQI forecasting. The key findings of this study are as follows:

1. **Quantitative Improvement:** The application of edge-level Kalman filtering resulted in a massive reduction in forecasting error. The Mean Absolute Error (MAE) decreased from **14.24** (using raw data) to **0.34** (using cleaned data), representing a **~97.6% improvement** in predictive precision .
2. **Visual Stability:** Time-series analysis confirmed that the proposed model produces smooth, continuous prediction curves that closely track actual pollution trends. In contrast, models trained on raw data exhibited erratic volatility, leading to unstable Air Quality Index (AQI) classifications .
3. **Comparative Superiority:** Benchmarking against state-of-the-art studies (e.g., Du et al., Kairatov et al., Tursunov et al.) revealed that our "Data-Centric" approach outperforms complex Deep Learning and Ensemble architectures in terms of accuracy. While complex models struggled with noisy inputs (MAE > 2.0), our computationally efficient Linear Regression model achieved near-perfect alignment (MAE 0.34) when fed with high-quality, preprocessed data.

These results validate the hypothesis that prioritizing **data quality at the edge** is a more effective and scalable strategy for real-time smart city environmental monitoring than solely relying on increasing model complexity in the cloud.

4. Discussion

The Primacy of Data Quality in Forecasting The most significant finding of this study is the quantified impact of edge-level data cleaning on predictive performance. As demonstrated in the Results section, the transition from raw to Kalman filter–cleaned data reduced the Mean Absolute Error (MAE) from 14.24 to 0.34 . This supports the fundamental hypothesis that **sensor noise**, rather than model incapacity, is the primary bottleneck in low-cost air quality monitoring systems.

By recursively estimating the true state of the system and suppressing stochastic noise at the source, the Kalman Filter enabled a simple Linear Regression model to capture underlying temporal trends effectively. This finding challenges the "Model-Centric" trend observed in recent literature, where researchers often resort to increasingly complex Deep Learning architectures (e.g., Du et al. [3]) to compensate for poor data quality. Our results suggest that a "Data-Centric" approach—prioritizing signal quality over

algorithmic complexity—yields superior accuracy while requiring significantly less computational power.

Comparative Advantages over Existing Methods When benchmarked against recent studies, the proposed Edge–Cloud framework exhibited distinct advantages. For instance, Tursunov et al. (2024) [8] reported higher error rates using Stacked Generalization models in similar arid environments (Tashkent), likely due to the unmitigated impact of particulate matter intrusions and sensor drift. Similarly, while Kairatov et al. (2025) [5] achieved reasonable accuracy using XGBoost, their approach requires substantial cloud-side processing.

In contrast, our approach demonstrates that high-precision forecasting does not necessarily require GPU-accelerated Neural Networks. By offloading the "cleaning" task to the edge device (IoT node), we achieve a dual benefit:

1. **High Accuracy:** Outperforming complex models (MAE 0.34 vs. >2.0).
2. **Bandwidth Efficiency:** Transmitting clean, stable data reduces the "noise" load on the network, which is critical for scalable Smart City deployments.

Practical Implications for Smart City Management The stability of the AQI forecasts generated by our system has direct practical applications for urban governance:

- **Public Health:** The elimination of erratic fluctuations ensures that citizens receive reliable health warnings. False alarms (e.g., oscillating between "Moderate" and "Unhealthy" AQI due to sensor noise) are minimized.
- **Traffic Management:** Reliable short-term predictions allow city authorities to implement dynamic traffic control measures *before* pollution levels reach critical thresholds.
- **Resource Allocation:** The computational lightness of the Linear Regression model allows this system to be deployed on thousands of low-power edge devices, creating a high-density, granular air quality map of the city.

Limitations and Future Directions Despite the promising results, this study has limitations that outline the path for future research:

1. **Linearity Assumption:** The use of Linear Regression assumes a linear relationship between past and future pollution levels. While effective for short-term forecasting on smoothed data, it may fail to capture complex, non-linear spikes caused by sudden events (e.g., fires or sandstorms) . Future work should explore lightweight non-linear models (e.g., SVR or decision trees) on cleaned data.
2. **Multivariate Factors:** The current model relies solely on historical PM2.5 data. Air quality is heavily influenced by meteorological factors (wind speed, humidity, temperature) . Integrating these variables into the forecasting model would likely further enhance robustness.
3. **Static Kalman Parameters:** The Process Noise Covariance (Q) and Measurement Noise Covariance (R) were kept constant. In real-world scenarios, sensor noise characteristics may change over time (sensor drift). Implementing an **Adaptive Kalman Filter (AKF)** that self-tunes these parameters would make the system more resilient to changing environmental conditions .

5. Conclusion

This study addresses the critical challenge of forecasting particulate matter (PM2.5) concentrations and Air Quality Index (AQI) using low-cost, noise-prone IoT sensors in smart city environments. We proposed and validated a hybrid Edge–Cloud framework that integrates real-time signal processing at the device level with predictive modeling in the cloud.

The experimental results provide compelling evidence that **data quality is a prerequisite for forecasting accuracy**. By applying a recursive Kalman Filter at the edge layer, the proposed system successfully recovered the true pollution signal from noisy sensor data. The quantitative impact was substantial: the Linear Regression model trained on cleaned data achieved a **Mean Absolute Error (MAE) of 0.34** and a **Root Mean Squared Error (RMSE) of 0.42**, compared to an MAE of 14.24 and RMSE of 16.87 for the baseline model trained on raw data. This represents a prediction error reduction of approximately **97.6%**.

Furthermore, comparative benchmarking against recent state-of-the-art studies (e.g., Du et al., Kairatov et al., Tursunov et al.) demonstrated the superiority of this "Data-Centric" approach. While complex Deep Learning and Ensemble models often struggle to achieve high accuracy on noisy datasets (typically

yielding MAE > 2.0), our approach proved that a computationally efficient Linear Regression model can achieve near-perfect alignment with actual values when the input data is rigorously preprocessed.

The practical implications of this research are significant for urban management. The system produces stable and reliable AQI forecasts, eliminating the erratic fluctuations observed in raw data predictions that often lead to false public health alarms. This reliability supports informed decision-making for traffic management, industrial regulation, and personalized health advisories.

Future Work: To further enhance the robustness of the forecasting system, future research will focus on:

1. **Multivariate Integration:** Incorporating external meteorological variables (e.g., humidity, wind speed, temperature) to capture complex environmental dynamics.
2. **Adaptive Filtering:** Developing an Adaptive Kalman Filter (AKF) that automatically tunes noise covariance parameters to handle sensor drift over time.
3. **Advanced Modeling:** exploring lightweight non-linear models (e.g., SVR or decision trees) on the cleaned data to better predict sudden pollution spikes caused by non-standard events.

In summary, this work establishes that integrating edge-level data cleaning with cloud-based modeling provides a scalable, accurate, and cost-effective solution for real-time air quality monitoring in modern smart cities.

References

[1] World Health Organization, *Air Quality Guidelines: Global Update 2021*. Geneva, Switzerland: WHO Press, 2021.

[2] A. Botta, W. de Donato, V. Persico, and A. Pescapé, “Integration of cloud computing and Internet of Things: A survey,” *Future Generation Computer Systems*, vol. 56, pp. 684–700, 2016.

[3] S. Du, T. Li, Y. Yang, and S. J. Horng, “Deep air quality forecasting using hybrid deep learning framework,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2412–2424, 2024.

[4] K. Zhang, X. Wang, and Z. Liu, “Application of Deep Learning Techniques for Air Quality Prediction: A Case Study in Coastal Cities,” *Atmospheric Environment*, vol. 245, pp. 118–130, 2024.

[5] A. Kairatov, B. Sultangazin, and M. Zhumabayev, “Air Pollution Forecasting in Almaty using Ensemble Machine Learning Models and STL Decomposition,” *Journal of Atmospheric Data Science*, vol. 12, no. 3, pp. 45–58, 2025.

[6] S. Nurmakhanova and G. Utepov, “Predicting particulate matter (PM2.5) air pollution levels in Almaty city using LSTM and BiLSTM techniques,” *Central Asian Journal of Environmental Science*, vol. 9, no. 2, pp. 112–125, 2025.

[7] I. Ivanov and P. Smirnov, “Machine Learning Models for Air Pollution Health Risk Assessment in Russian Industrial Cities,” *Environmental Monitoring and Assessment*, vol. 196, no. 8, p. 304, 2025.

[8] U. Tursunov, S. Boboev, and J. Kim, “Machine Learning-Based Prediction of Air Quality in Tashkent: Challenges and Ensemble Solutions,” *International Journal of Advanced Research in Science, Engineering and Technology (Uzbekistan)*, vol. 11, no. 2, pp. 22–30, 2024.

[9] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[10] M. A. Zaidan et al., “Edge computing-based air quality monitoring system for smart cities,” *IEEE Access*, vol. 9, pp. 112345–112357, 2021.

[11] B. M. Kono, "Urban Air Quality and IoT Sensor Health Data," Kaggle Data Repository. [Online]. Available: https://www.kaggle.com/datasets/bertnardomariouskono/urban-air-quality-and-iot-sensor-health-data.

[12] T. Reyimberdiev, O.Khujaev, "A Comparative Analysis of Air Quality Index (AQI) Calculations Using Raw and Cleaned Sensor Data," *Raqamli transformatsiya va suniy intellekt jurnali (Under Review)*, 2026.