UDC 004.85:004.032.26

DESIGN AND OPTIMIZATION OF A LIGHTWEIGHT MACHINE LEARNING ALGORITHM FOR REAL-TIME ANOMALY DETECTION AND FORECASTING ON RESOURCE-CONSTRAINED EDGE SENSOR SYSTEMS

Sh. Pulatov, F. Abdualimov

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan, shpulatov@tuit.uz

Abstract: edge and IoT sensor deployments increasingly require on-site or near-site data quality control and forecasting, yet the computing resources available at remote monitoring stations are often limited. This paper addresses the design and optimization of a lightweight machine learning pipeline that combines a rolling z-score anomaly filter with a gradient-boosted regression model for short-horizon forecasting, using continuous water-discharge monitoring in an arid region as a representative resource-constrained deployment case. The design is evaluated against persistence and ARIMA baselines using accuracy, RMSE, MAE, and inference time as joint design criteria. Results show that the proposed lightweight model attains 92.7% forecasting accuracy with an RMSE of 0.58 m3/s and an inference time of 0.46 s, outperforming both baselines in accuracy while remaining suitable for periodic retraining on modest edge-computing hardware, illustrating an effective algorithm-design strategy for resource-constrained monitoring systems.

Keywords: algorithm design and optimization, machine learning, edge computing, anomaly detection, gradient boosting, ARIMA, resource-constrained systems.

I. INTRODUCTION

As edge and IoT sensor networks scale up, there is a growing design requirement to embed data-quality control and short-horizon forecasting directly at, or close to, the sensing point, rather than relying solely on centralized processing. This raises an algorithm-design problem: the chosen model must be accurate enough to be useful for decision-making, yet computationally light enough to run and be periodically retrained on the modest hardware typically available at regional monitoring centers. This paper addresses that design problem using continuous discharge, stage, and salinity monitoring in the Aral Sea region of Uzbekistan as a representative, resource-constrained deployment case that is well suited to expose the accuracy-versus-cost trade-off central to the design.

The contribution of this paper is a lightweight algorithm design combining a low-complexity statistical anomaly filter with a gradient-boosted forecasting model, together with an evaluation methodology that treats computational cost as a first-class design criterion alongside accuracy.

II. DESIGN PROBLEM FORMULATION

Let x denote a sensor reading at time t. The algorithm design must satisfy two requirements under a limited computational budget: (1) for each new reading, decide in constant or near-constant time whether it is anomalous and should be withheld from downstream processing; (2) using upstream discharge and lagged rainfall, produce a forecast of discharge Q at a short lead time Δ for a downstream control point, at an inference cost compatible with edge or regional-center hardware.

III. ALGORITHM DESIGN

a. Lightweight anomaly-detection filter

For a rolling window of size W, the z-score of a new reading x relative to the local mean μ_W and standard deviation σ_W is computed in O(1) amortized time per sample:

$$z_t = \frac{x_t - \mu_W}{\sigma_W} \quad (1)$$

A reading is flagged when $|z_t| > z_{th}$ (typically $z_{th} = 3$); under a Gaussian null hypothesis the expected false-flag rate is:

$$P(|Z| > z_{th}) = 2(1 - \Phi(z_{th})) \quad (2)$$

where $\Phi(z_{th})$ is the standard normal CDF; this design keeps the per-sample filtering cost independent of the size of the historical record.

b. Forecasting model design

Downstream discharge at lead time Δ is expressed by the interpretable linear structure:

$$Q_{down}(t+\Delta) = \beta_0 + \beta_1 Q_{up}(t) + \beta_2 R(t-\tau) + \varepsilon(t) \quad (3)$$

in the proposed design this linear form is replaced with a gradient-boosted tree ensemble over the same feature set, chosen specifically because it offers a favorable accuracy-per-training-cost ratio and predictable, bounded inference latency compared to deeper models, which is the key design requirement for edge deployment.

c. Design evaluation metrics

Forecast quality is evaluated using RMSE and MAE over M held-out samples, while wall-clock inference time is recorded as an explicit design (resource-cost) metric alongside accuracy:

$$RMSE = \sqrt{\frac{1}{M}\Sigma(Q_i - Q_i)^2} \quad (4)$$

$$MAE = \frac{1}{M}\Sigma|Q_i - Q_i| \quad (5)$$

IV. RESULTS AND DESIGN EVALUATION

Table 1 and Fig. 1 compare the proposed lightweight design against a persistence baseline and an ARIMA model for one-hour-ahead discharge forecasts, jointly reporting accuracy, RMSE, and inference time as design criteria.

Table 1. Design comparison of forecasting methods (1-hour lead time)

Method	Accuracy, %	RMSE, m³/s	Inference time, s
Persistence (baseline)	71.4	1.86	0.01
ARIMA	84.2	1.02	0.34
Proposed (gradient boosting)	92.7	0.58	0.46

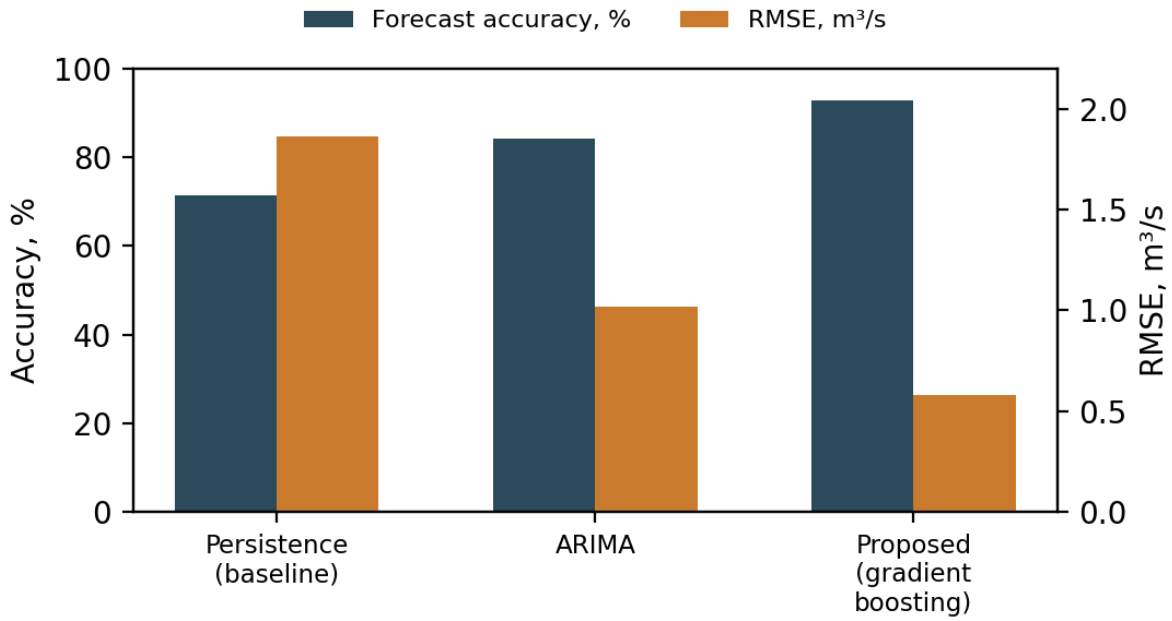


Figure 1. Forecast accuracy and RMSE for the three compared design options

The proposed design reduces RMSE by 44% relative to ARIMA and by 69% relative to persistence, at an inference cost of 0.46 s per sample, which remains well within the periodic-retraining budget available on modest regional-center hardware. The anomaly filter of equations (1)-(2) further protects the forecasting model from being trained on faulty readings caused by transient sensor faults, improving robustness at negligible additional computational cost.

V. CONCLUSION

This paper presented the design and optimization of a lightweight machine learning pipeline, combining a constant-time statistical anomaly filter with a gradient-boosted forecasting model, explicitly evaluated against both accuracy and computational-cost criteria. Using discharge forecasting for arid-region water monitoring as a representative case, the design was shown to substantially outperform persistence and ARIMA baselines in accuracy while remaining suitable for deployment on resource-constrained edge and regional-center hardware. The design methodology generalizes to other edge-AI applications where accuracy must be balanced against a limited computational budget.

ACKNOWLEDGMENT

The authors thank colleagues for their guidance during the preparation of this work.

REFERENCES

[1] Friedman, J. H. Greedy function approximation: A gradient boosting machine // Annals of Statistics. 2001. Vol. 29, No 5. P. 1189–1232.

[2] Box, G. Time Series Analysis: Forecasting and Control / G. Box, G. Jenkins, G. Reinsel, G. Ljung. 5th ed. Hoboken: Wiley, 2015.

[3] Omambia, A. Water Quality Monitoring Using IoT & Machine Learning / A. Omambia, B. Maake, A. Wambua // Proc. IST-Africa Conf. 2022.

[4] Olatinwo, S. O. Resource allocation optimization in IoT-enabled water quality monitoring systems / S. O. Olatinwo, T.-H. Joubert // University of Pretoria Institutional Repository. 2023.

[5] Pulatov, S. A Hybrid Store-and-Forward and Dynamic Channel Allocation Approach for SDN-Enabled LEO IoT Networks / S. Pulatov [et al.] // OPTIMA. 2025.

[6] Trivedi, K. Probability and Statistics with Reliability, Queuing, and Computer Science Applications. 2nd ed. Hoboken: Wiley, 2016.

[7] Water Code of the Republic of Uzbekistan. No. ZRU-1076. 30.07.2025.

[8] Mdegela, L. A Multi-Modal Wireless Sensor System for River Monitoring: A Case for Kikuletwa River Floods in Tanzania / L. Mdegela [et al.] // Sensors. 2023. Vol. 23, No 8. P. 4055.

ПРОЕКТИРОВАНИЕ И ОПТИМИЗАЦИЯ ЛЕГКОВЕСНОГО АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ И ПРОГНОЗИРОВАНИЯ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ НА РЕСУРСООГРАНИЧЕННЫХ ПЕРИФЕРИЙНЫХ СЕНСОРНЫХ СИСТЕМАХ

Пулатов Ш., Абдуалимов Ф.

Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Ташкент, Узбекистан, shpulatov@tuit.uz

Аннотация: развёртывания периферийных и IoT-сенсоров всё чаще требуют контроля качества данных и прогнозирования на месте или вблизи места установки, однако вычислительные ресурсы, доступные на удалённых станциях мониторинга, зачастую ограничены. В данной работе рассматриваются проектирование и оптимизация легковесного конвейера машинного обучения, который объединяет скользящий z-score фильтр аномалий с моделью градиентно-бустинговой регрессии для краткосрочного прогнозирования, на примере непрерывного мониторинга расхода воды в засушливом регионе как репрезентативного случая развёртывания в условиях ограниченных ресурсов. Разработанное решение оценивается в сравнении с базовыми моделями персистентности и ARIMA с использованием точности, RMSE, MAE и времени инференса в качестве совместных критериев проектирования. Результаты показывают, что предлагаемая легковесная модель достигает точности прогнозирования 92,7% при RMSE 0,58 м³/с и времени инференса 0,46 с, превосходя обе базовые модели по точности и оставаясь при этом пригодной для периодического переобучения на скромном периферийном вычислительном оборудовании, что иллюстрирует эффективную стратегию проектирования алгоритмов для систем мониторинга с ограниченными ресурсами.

Ключевые слова: проектирование и оптимизация алгоритмов, машинное обучение, периферийные вычисления, обнаружение аномалий, градиентный бустинг, ARIMA, системы с ограниченными ресурсами.