Knowledge-based Feature Selection Enhances Wastewater Treatment Modelling

Senyuan Gu and Haoran Duan
UNSW Water Research Centre

Introduction & Aim

- Traditional data-driven models in wastewater treatment often struggle with small, high-dimensional datasets, leading to unstable and less interpretable results.
- To address this, we developed a knowledge-guided feature selection framework that integrates mechanistic understanding with statistical correlation analysis.
- Our goal: To enhance prediction accuracy, model stability, and interpretability by embedding domain expertise into the data-driven workflow.

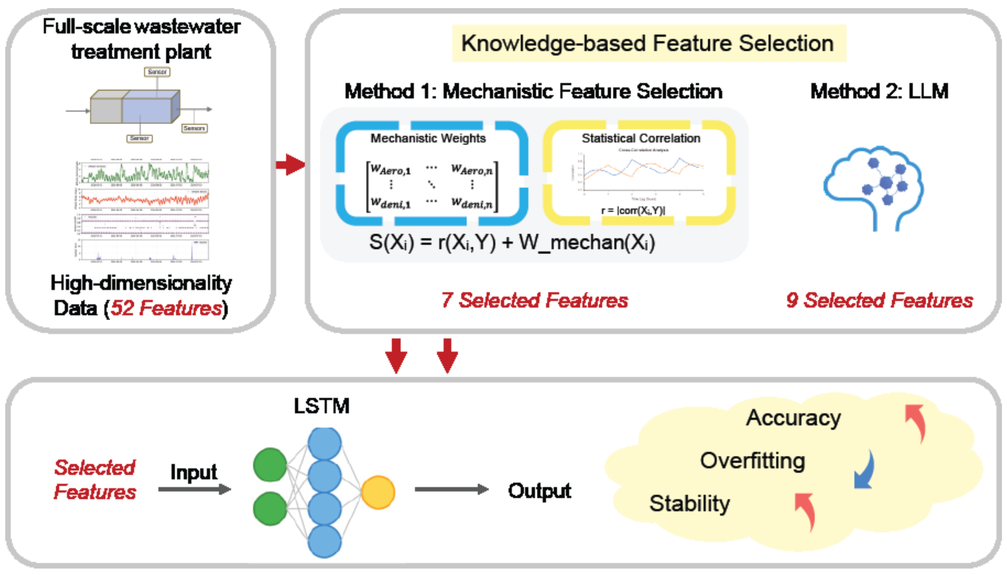


Figure 1: Overview of the proposed knowledge-driven framework.

Methods

- Data Source:51 days of full-scale WWTP (Queensland, Australia) SCADA data at 15-min resolution.
- Knowledge-Driven Feature Selection:Calculate statistical correlation (SC-score) and mechanistic importance (MI-score).Remove redundant variables (similarity < 0.85).Select optimal subset of features combining knowledge and data evidence.
- LLM-Assisted Selection:Large language models (GPT-4, Gemini 2.0, Claude 3.5) were used to interpret process descriptions and suggest relevant features, further refining domain-informed selection.

Knowledge-Driven Feature Selection

LLM-Assisted Feature Selection

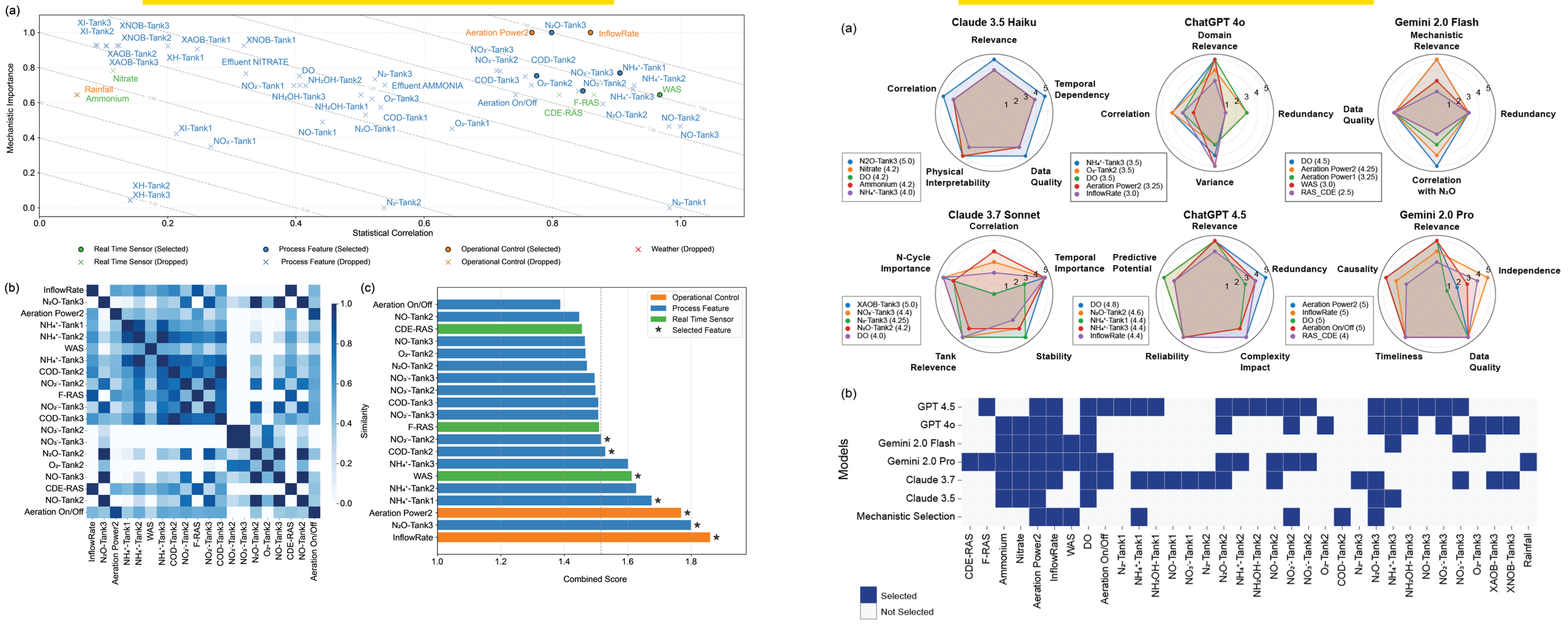
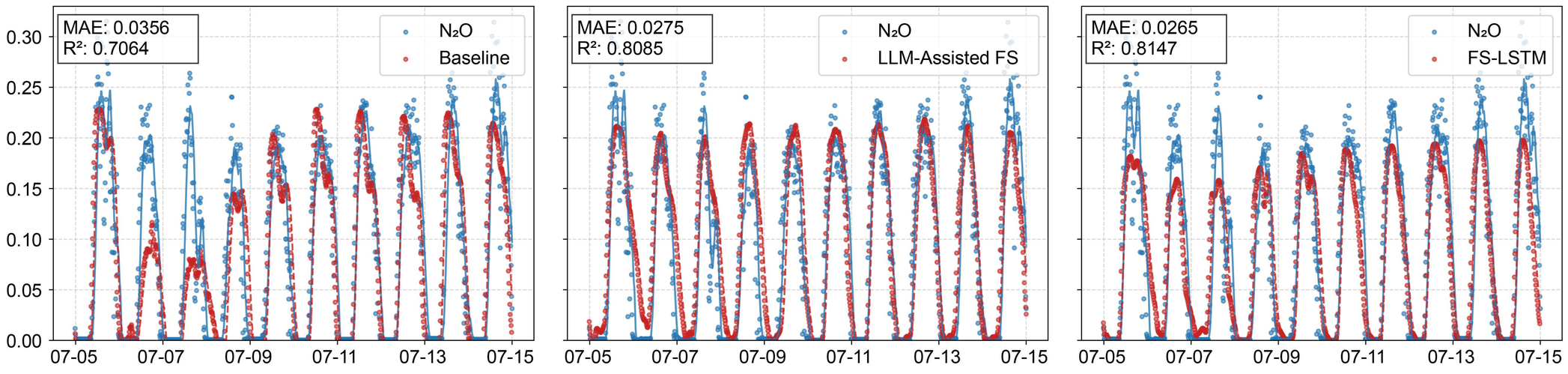


Figure 2&3: Workflow of knowledge- and LLM-assisted feature selection.

Results & Discussion

- FS-LSTM achieved R^2 = 0.812, MAE = 0.027 — significantly outperforming attention models.
- LLM-assisted feature selection (Gemini 2.0 Flash) achieved R^2 = 0.809.
- Demonstrates robust performance under small-sample, high-dimensional data conditions.



Model Performance:

- *Baseline LSTM:* R^2 = 0.706, MAE = 0.036.
- *Knowledge-Driven FS-LSTM:* R^2 = 0.812, MAE = 0.027.
- *LLM-Assisted FS:* R^2 = 0.809, MAE = 0.028.

Conclusions:

- Knowledge-guided feature selection improves accuracy.
- LLMs can enhance domain knowledge synthesis for feature discovery.
- Framework generalizable to other environmental systems.