Machine Learning Approaches for Forecasting Ammonia Concentrations and Performance Outcomes in Wastewater Treatment Processes

Abduljaber Abdulqader
a.abdulqader2@ncl.ac.uk

Newcastle University

UK infrastructure + Add to myFT

Veolia chief driven 'nuts' by UK water utilities' failure to use AI to detect leaks

England's water companies are lagging behind other more water-stressed nations, says Estelle Brachlianoff

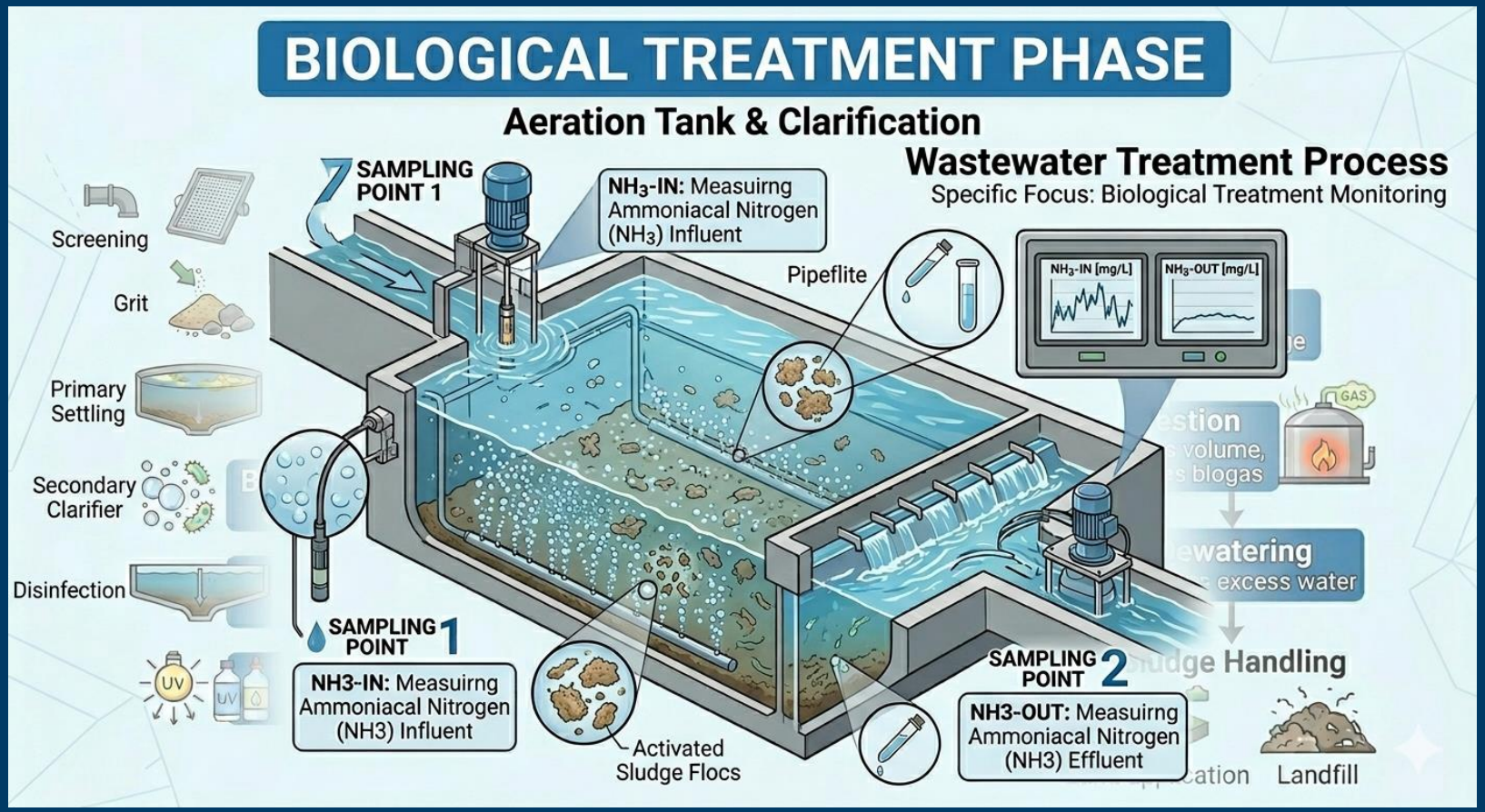
Newcastle University

Content

- Biological Treatment in **Wastewater Treatment Plants (**WWTPs)
- WWTPs Challenges
- Research Objective
- Data Selection Approaches
- Modelling Approaches
- Results and Performance
- Future Work
- Any Questions?

Biological Treatment in WWTPs

- Wastewater Treatment Plants (WWTPs) consist of multiple treatment phases
- This research focuses on the biological treatment phase
- Treatment methods and environmental conditions vary significantly
- Limited availability of data samples poses a challenge
- WWTPs contribute to greenhouse gas (GHG) emissions – Biggest source of emission for the water industry.



Source: Process flow adapted from Metcalf & Eddy; Visualization generated via Gemini AI (2026).

Newcastle University

WWTPs Challenges

- The microbial ecosystem exhibits instability
- Variations in the quality of the influent water are frequently observed
- The physicochemical composition of the wastewater may impede the efficiency of the treatment process

Why Ammonia NH3?

Key Pollutant: NH3 (ammonia) is a major pollutant from wastewater that plants must monitor and remove.

Protects Aquatic Life: High NH3 is toxic to fish; it depletes oxygen via nitrification and causes eutrophication (algal blooms).

Meets Regulations: Strict limits (often 5-15 mg/L NH3) prevent fines and ensure safe discharge.

Optimises Processes: adjusting aeration, cuts energy 15-45%, stabilises conversion to nitrate, and lowers costs.

Research Objective

- The research aims to improve process predictability in wastewater treatment through advanced prediction techniques, enabling early intervention

- Specifically, machine learning models trained on historical data are used to predict ammonia spikes and support faster operational responses.

Newcastle University

Data

A 5-year time series dataset includes 257 data points and 71 attributes.

Chemical data

COD	NH3	MLSS	MLVSS	Nitrate	Nitrite	Sulphate	Phosphate	Fluoride	Chloride	pH	Temperature	DO
32	28	2.1	1.7	2.45	0.38	87.92	10.23	0.6	79.5	7.04	14.5	2.6

Taxonomy data

Kingdom	Phylum	Class	Order	Family	Genus

Water Flow

INF (Influent): This refers to the incoming wastewater that enters the treatment plant.

EFF (Effluent): This is the treated wastewater that is discharged from the treatment plant into the environment.

Newcastle University

Data Selection Approaches

Data Selection based on certain NH3 threshold

Adding Previous value

Permutation Importance
Random Forest Feature Importance
SHAP (SHapley Additive exPlanations)
SMOTE

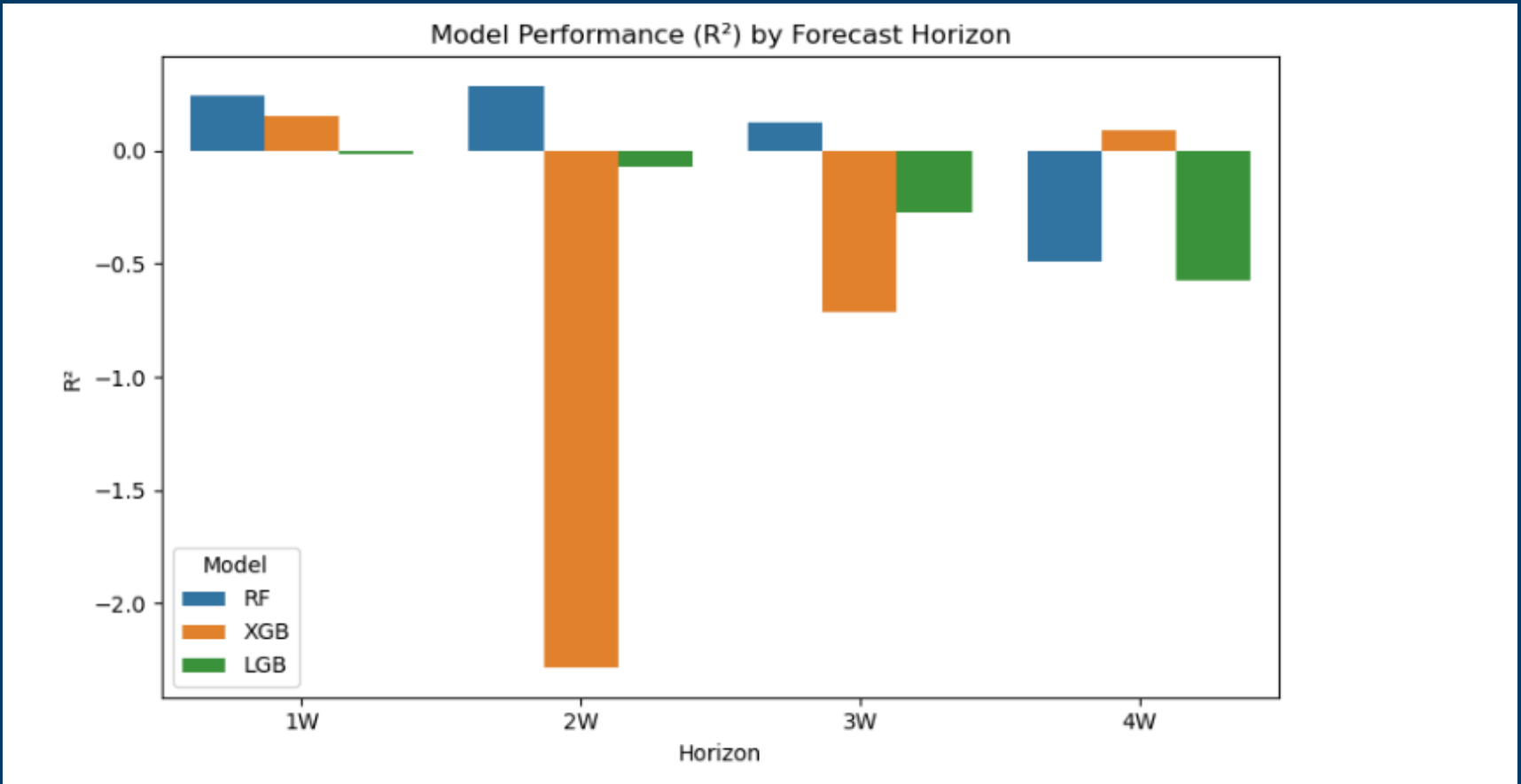
Newcastle University

Modelling Approaches

- Model performances varied by data selection methods, largely depending on the datasets and prediction horizons..
- Analyse feature importance to identify key predictors
- Shortlisted three models:
- Random Forest (RF), XGBoost (XGB), and LightGBM (LGB).
  - Random Forest uses bagging to builds independent trees on bootstrapped data subsets with random feature selection.
  - XGB uses sequential gradient boosting with level-wise tree growth
  - LGB boosts faster via leaf-wise growth, gradient-based sampling

Newcastle University

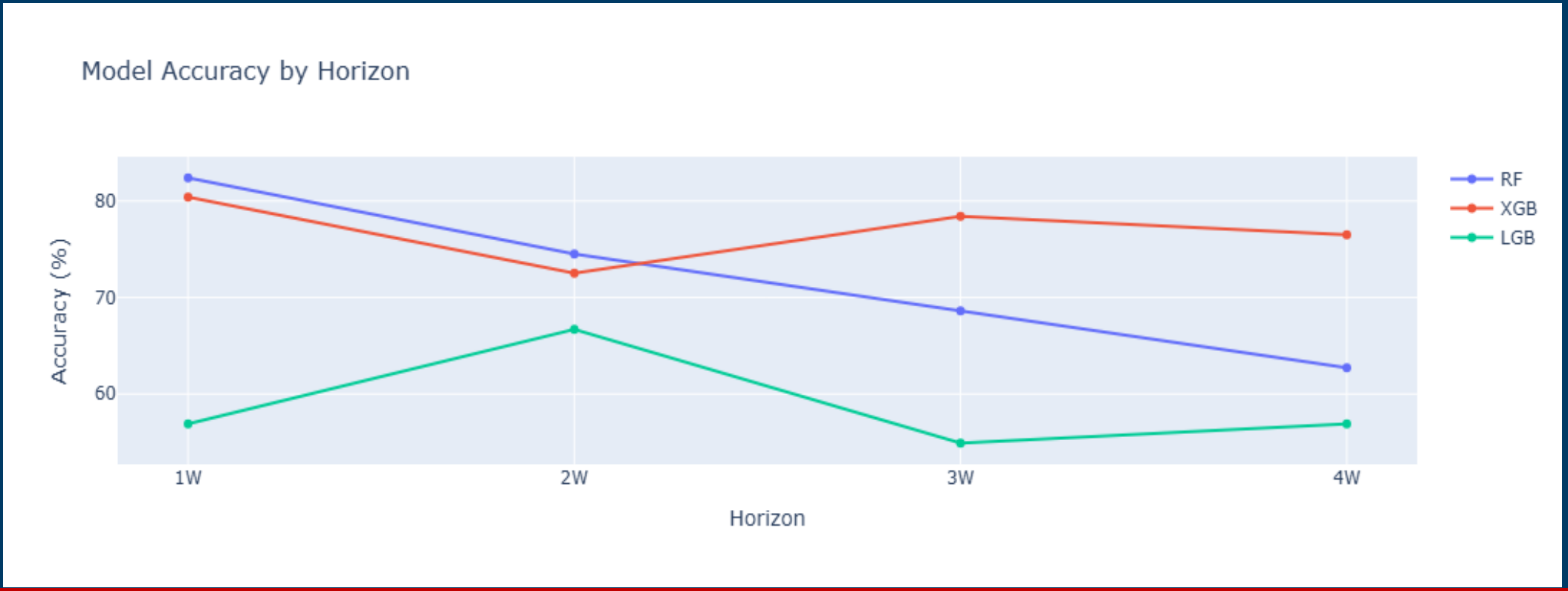
Results and Performance



Using the best feature importance results to train three shortlisted machine learning models.

Converting the prediction to True or False based on a threshold; i.e., how likely is it that NH_3 will exceed the threshold?

Horizon	RF Accuracy	XGB Accuracy	LGB Accuracy
1W	82.4%	80.4%	56.9%
2W	74.5%	72.5%	66.7%
3W	68.6%	78.4%	54.9%
4W	62.7%	76.5%	56.9%



Future Work

- Securing more data would help, but that would mean waiting for this new data to be sampled and produced.
- Include the taxonomy data to show which microbe is contributing to the consumption of NH_3.
- Understand how many data points are required to confirm that the model is predicting at its best ability with a small dataset.

Newcastle University

Thanks for listening. Any questions or comments?

a.abdulqader2@ncl.ac.uk