ABSTRAK

Data Tinggi Muka Air (TMA) yang akurat merupakan parameter yang penting dalam pengelolaan sumber daya air dan mitigasi risiko banjir. Adanya data yang hilang akibat kegagalan sensor atau kendala teknis dapat menghambat keakuratan analisis hidrologi. Penelitian ini mengevaluasi empat model imputasi data -*Random Forest* (RF), XGBoost, *Convolutional Long Short-Term Memory* (ConvLSTM), dan *Nonlinear AutoRegressive with Exogenous Inputs Long Short-Term Memor*y (NARX_LSTM)- untuk melengkapi data yang hilang pada *Automatic Water Level Recorder* (AWLR) Kretek di Daerah Aliran Sungai (DAS) Opak, Yogyakarta. Model ini memanfaatkan data tinggi muka air Pulo dan curah hujan yng berasal dari 9 (Sembilan) pos hujan, terdiri dari 5 (lima) pos *Automatic Rainfall Recorder* (ARR) yang diperoleh dari Dinas Pekerjaan Umum, Perumahan, dan Energi Sumber Daya Mineral (DPUPESDM) Daerah Istimewa Yogyakarta, dan 4 (empat) pos *Automatic Rain Gauge* (ARG) yang diperoleh dari Badan Meteorologi Klimatologi dan Geofisika, sebagai variabel prediktor.

Periode data yang hilang mulai dari Juli hingga November 2018, merupakan periode kritis dalam pemantauan hidrologi. Evaluasi model dilakukan dengan membandingkan hasil imputasi terhadap *ground truth,* menggunakan metrik *Mean Squared Error* (MSE) dan *Mean Absolute Error* (MAE).

Hasil penelitian menunjukkan bahwa model NARX-LSTM tanpa *hyperparameter tuning* pada skenario 9 pos curah hujan memberikan performa terbaik, dengan nilai MAE rata-rata sebesar 0.0785 dan RMSE sebesar 0.1001. Model ConvLSTM juga mencatat akurasi tinggi, diikuti oleh XGBoost dan *Random Forest*. Penambahan jumlah data curah hujan secara konsisten meningkatkan performa model, namun proses *hyperparameter tuning* hanya memberikan peningkatan signifikan pada model *tree-based*. Seluruh model mampu menangkap pola data tinggi muka air yang memiliki data hilang sesuai dengan kategori *Missing Not at Random* (MNAR).

Studi ini memberikan wawasan mengenai efektivitas berbagai pendekatan dalam menangani data hidrologi yang hilang dan dapat menjadi referensi dalam pengembangan sistem peringatan dini serta manajemen sumber daya air.

Kata kunci: tinggi muka air, imputasi data, Random Forest, XGBoost, ConvLSTM, NARX-LSTM, Sungai Opak, Curah Hujan.

ABSTRACT

Accurate Water Level data is a crucial parameter in water resources management and flood risk mitigation. Missing data due to sensor failures or technical issues can hinder the accuracy of hydrological analysis. This study evaluates four data imputation models -Random Forest (RF), XGBoost, Convolutional Long Short-Term Memory (ConvLSTM), and Nonlinear Autoregressive with Exogenous Inputs Long Short-Term Memory (NARX-LSTM)- to reconstruct missing data at the Automatic Water Level Recorder (AWLR) in the Opak River Basin, Yogyakarta. These models utilize Pulo water level data and rainfall data from Dinas Pekerjaan Umum, Perumahan, dan Energi Sumber Daya Mineral (DPUPESDM) Special Region of Yogyakarta and Badan Meteorologi Klimatologi dan Geofisika as predictor variable.

The missing data period, spanning July to November 2018, represents a critical phase in hydrological monitoring. The models' performance is evaluated by comparing the imputed data results with the ground truth, using metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE)

The results show that the NARX-LSTM model without hyperparameter tuning using the 9-station rainfall configuration achieved the best performance, with an average MAE of 0.0785 and RMSE of 0.1001. ConvLSTM also demonstrated high accuracy, followed by XGBoost and Random Forest. Increasing the number of rainfall data consistently improved model performance, while hyperparameter tuning only provided significant improvement for tree-based models. All models were able to capture the underlying patterns of water level data, despite the missing data occurring sequentially and categorized as Missing Not at Random (MNAR).

This study provides insights into the effectiveness of various approaches in addressing missing hydrological data and serves as reference for developing early warning systems and water resources management.

Keywords: water level data imputation, missing data, Random Forest, XGBoost, ConvLSTM, NARX-LSTM, Opak River, rainfall.