ABSTRAK

ALVIRA ANGRAINI. Analisis Prediksi Polutan PM2.5 dengan Reduksi Dimensi PCA dan *Deep Learning* LSTM Pada Data Deret Waktu Multivariat. Dibimbing oleh HENDRA KURNIAWAN dan MUHAMAD FADLI.

PM2.5 merupakan indikator kualitas udara paling berbahaya bagi kesehatan manusia. Penelitian ini menganalisis pengaruh seleksi fitur menggunakan *Principal Component Analysis* (PCA) terhadap akurasi model *Long Short-Term Memory* (LSTM) dalam memprediksi konsentrasi PM2.5, dengan membandingkan RMSE, MAE, dan R^2 antara LSTM Murni dan PCA-LSTM pada data deret waktu multivariat di Serpong, Tangerang Selatan. Data terdiri atas 15 variabel input, termasuk *history* PM2.5 dan *wind direction* yang ditransformasi ke sin dan cos, dicatat tiap 30 menit sepanjang 2023. Pra-pemrosesan meliputi normalisasi *Min-Max*, transformasi *wind direction*, dan korelasi Pearson. Kedua model menggunakan parameter sama hasil *grid search* (*units* 32, *dropout* 0,1, *learning rate* 0,001, *batch size* 32), dengan pembagian data 70:15:15 dan *lookback* 12 langkah waktu. PCA diterapkan berdasarkan kriteria eigenvalue Kaiser (>1) dan *loading factor* (|loading| > 0,5), menghasilkan 4 fitur terpilih (O3, Wind_Dir_Sin, Wind_Dir_Cos, Air_Pressure) yang digabung dengan data historis PM2.5, sehingga total fitur PCA-LSTM berjumlah 5 dari 15 variabel. Hasil menunjukkan PCA-LSTM lebih unggul (RMSE 7,4813; MAE 5,0647; R^2 0,8688) dibanding LSTM Murni (RMSE 7,7751; MAE 5,4971; R^2 0,8583). Seleksi fitur PCA menyaring variabel kurang berkontribusi dan redundansi antar fitur, sehingga LSTM belajar pola temporal lebih efisien dan generalisasi lebih baik. Penelitian ini menghasilkan prototipe sistem *web* pemantauan kualitas udara.

Kata kunci: *Deep Learning*, LSTM, PM2.5, PCA, prediksi kualitas udara.

ABSTRACT

ALVIRA ANGRAINI. Prediction Analysis of PM2.5 Pollutant Using PCA Dimensionality Reduction and Deep Learning LSTM on Multivariate Time series Data. Supervised by HENDRA KURNIAWAN and MUHAMAD FADLI.

PM2.5 is the air quality indicator most hazardous to human health. This study analyzes the impact of feature selection using Principal Component Analysis (PCA) on the accuracy of a Long Short-Term Memory (LSTM) model for predicting PM2.5 concentrations by comparing RMSE, MAE, and R^2 values between a standard LSTM and a PCA-LSTM model using multivariate time-series data from Serpong, South Tangerang. The dataset comprises 15 input variables—including historical PM2.5 data and wind direction (transformed into sine and cosine components)—recorded at 30-minute intervals throughout 2023. Pre-processing steps included Min-Max normalization, wind direction transformation, and Pearson correlation analysis. Both models utilized identical parameters derived from a grid search (32 units, 0.1 dropout, 0.001 learning rate, and a batch size of 32), with a 70:15:15 data split and a lookback window of 12 time steps. PCA was applied based on Kaiser's eigenvalue criterion (>1) and a loading factor threshold (|loading| > 0.5), resulting in four selected features (O3, Wind_Dir_Sin, Wind_Dir_Cos, and Air_Pressure); these were combined with historical PM2.5 data, bringing the total feature count for the PCA-LSTM model to five out of the original fifteen variables. The results demonstrate that the PCA-LSTM model outperformed the standard LSTM model (RMSE: 7.4813, MAE: 5.0647, R^2: 0.8688 vs. RMSE: 7.7751, MAE: 5.4971, R^2: 0.8583). PCA feature selection filtered out variables with low contribution and reduced inter-feature redundancy, enabling the LSTM to learn temporal patterns more efficiently and achieve better generalization. This research resulted in the development of a prototype web-based air quality monitoring system.

Keywords: Deep Learning, LSTM, PM2.5, PCA, Air Quality Prediction.