ABSTRACT

The lack of completeness of rainfall data from rain gauge measurements might influence the quality of the data and weather prediction results due to missing values in the dataset. Several recent studies have used a machine learning model with an Artificial Neural Network (ANN) algorithm as the imputation approach to fill in missing rainfall data with high accuracy. However, the input feature selection approach continues to provide insufficient attention to the variable features that are most relevant to the target variable, and the imputation operation is still performed noniteratively, resulting in suboptimal rainfall estimation model accuracy.

This work aims to identify strategies for improving the accuracy of ANN imputation methods in estimating rainfall missing data when rainfall datasets and non-rainfall meteorological features are used as predictive input variables. This study provides an imputation framework with two approaches for improving the accuracy of the rainfall imputation approach using an ANN algorithm. First, feature selection is utilized to find the most important elements from a combination of nearby meteorological and rainfall stations. use iterative imputation using an ANN model. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). The suggested imputation framework's performance is compared to noniterative imputation methods.

The results of the experiment demonstrate that feature selection based on the spatial-temporal connection of rainfall effectively enhances the performance of ANN-based rainfall imputation, as demonstrated by the evaluation values of $R^2 = 0.861$, MAE = 2.413, and RMSE = 4.882. In addition, the iterative imputation method using the ANN model outperforms noniterative imputation methods for filling in missing rainfall data at all levels of data loss evaluated, ranging from 5% to 40%. Based on the evaluation factor, the suggested framework performs best at a rainfall data loss rate of 5%, with $R^2 = 0.890$, RMSE = 2.183, and MAE = 1.932.

Keywords: precipitation, imputation, ANN, missing value, rain gauge

INTISARI

Kurangnya kelengkapan data curah hujan yang dikumpulkan oleh sensor *rain gauge* dapat mempengaruhi kualitas data dan hasil prakiraan cuaca karena adanya *missing value* pada dataset. Metode imputasi untuk mengisi data curah hujan yang hilang dengan akurasi yang baik pada beberapa penelitian terakhir adalah menggunakan model *machine learning* dengan algoritma *Artificial Neural Network* (ANN). Namun pendekatan pemilihan fitur masukan masih kurang memperhatikan fitur variabel yang paling relevan terhadap variabel target dan operasi imputasi masih dilakukan secara *noniterative* yang mengakibatkan akurasi model estimasi curah hujan menjadi kurang maksimal.

Penelitian ini mencoba mengetahui metodologi yang dapat diterapkan untuk meningkatkan akurasi metode imputasi menggunakan model ANN dalam memprediksi data curah hujan yang hilang saat menggunakan dataset curah hujan dan variabel meteorologi bukan hujan sebagai variabel *input* prediksi. Penelitian ini mengusulkan kerangka imputasi dengan dua strategi. Pertama, *features selection* digunakan untuk menentukan fitur yang paling relevan dari kombinasi data meteorologi dan curah hujan tetangga yang digunakan sebagai *input* untuk estimasi nilai yang hilang. Kedua, *iterative imputation* yaitu metode imputasi *iterative* menggunakan model ANN untuk mengisi data curah hujan yang hilang. Untuk mengetahui kinerja model imputasi, kami membandingkan kinerja model usulan dengan metode imputasi *noniterative* menggunakan evaluasi *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE) dan *Coefficient of Determination* (R^2).

Hasil pengujian menunjukkan bahwa seleksi fitur berdasarkan hubungan spasial dan temporal curah hujan berhasil meningkatkan performa imputasi curah hujan berbasis model ANN dengan nilai evaluasi RMSE = 4,882, MAE = 2,413, dan R^2 = 0,861. Selain itu metode imputasi *iterative* menggunakan model ANN mampu mengungguli performa metode imputasi *noniterative* untuk mengisi data curah hujan yang hilang pada setiap tingkatan kehilangan data yang diuji yaitu 5% hingga 40%. *Framework* yang diusulkan memiliki performa terbaik pada tingkat kehilangan data curah hujan 5% berdasarkan metrik evaluasi yaitu R^2 = 0,890, RMSE = 2,183, dan MAE = 1, 932.

Kata Kunci: curah hujan, imputasi, ANN, *missing value, rain gauge.*