

ABSTRAK

Tuberkulosis merupakan salah satu penyakit menular yang masih menjadi permasalahan kesehatan serius di Indonesia sehingga diperlukan metode diagnosis yang cepat dan akurat untuk membantu proses pengambilan keputusan. Penelitian ini bertujuan untuk membangun model klasifikasi diagnosis penyakit tuberkulosis berdasarkan data klinis menggunakan pendekatan *machine learning* dengan algoritma *Random Forest* serta mengukur performa model berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*. Data yang digunakan dalam penelitian ini adalah *Tuberculosis X-Ray Dataset (Synthetic)* yang diunggah oleh Arif Miah pada platform Kaggle. Proses penelitian meliputi tahap pra-proses data, pembagian data dengan beberapa rasio, pelatihan model, serta evaluasi performa model menggunakan metrik evaluasi klasifikasi. Hasil penelitian menunjukkan bahwa model *Random Forest* menghasilkan nilai *accuracy* berturut-turut sebesar 70,1%, 70,6%, dan 71,2% pada rasio data *split* 70:30, 80:20, dan 90:10, yang menunjukkan bahwa model mampu mengklasifikasikan status pasien tuberkulosis dengan tingkat ketepatan sekitar 70% dari keseluruhan data uji, baik untuk kelas tuberkulosis maupun normal. Selain itu, nilai *precision*, *recall*, dan *F1-score* yang diperoleh relatif serupa dan tidak berbeda secara signifikan. Keseragaman performa antar rasio data *split* menunjukkan bahwa performa model lebih dipengaruhi oleh karakteristik fitur pada *dataset* dibandingkan oleh variasi pembagian data, serta menunjukkan stabilitas algoritma *Random Forest* dalam melakukan klasifikasi diagnosis tuberkulosis.

Kata kunci: tuberkulosis; klasifikasi; *random forest*; *machine learning*.

ABSTRACT

Tuberculosis is one of the infectious diseases that remains a serious health problem in Indonesia requiring fast and accurate diagnostic method to aid the decision-making process. This study aims to build a tuberculosis diagnosis classification model based on clinical data using a machine learning approach with the Random Forest algorithm and to measure the model's performance based on accuracy, precision, recall, and F1-score values. The data used in this study is the Tuberculosis X-Ray Dataset (Synthetic) uploaded by Arif Miah on the Kaggle platform. The research process includes data preprocessing, data splitting with several ratios, model training, and model performance evaluation using classification evaluation metrics. The results show that the Random Forest model produced accuracy values of 70.1%, 70.6%, and 71.2% for data split ratios of 70:30, 80:20, and 90:10, indicating that the model is capable of classifying the status of tuberculosis patients with an accuracy rate of around 70% of the total test data, both for the tuberculosis and normal classes. In addition, the precision, recall, and F1-score values obtained were relatively similar and not significantly different. The consistency of performance across data split ratios indicates that the model's performance is more influenced by the feature characteristics in the dataset than by variations in data division, and demonstrates the stability of the Random Forest algorithm in classifying tuberculosis diagnoses.

Keywords: *tuberculosis; classification; random forest; machine learning.*

