

ABSTRAK

Leukemia adalah salah satu kanker darah paling mematikan yang sangat membutuhkan deteksi dini untuk pengobatan yang efektif. Namun, metode diagnosis konvensional sering kali subjektif, memakan waktu, dan mahal, sehingga menimbulkan tantangan terutama di daerah dengan keterbatasan sumber daya. Studi ini menyajikan analisis komparatif komprehensif dari dua algoritma pembelajaran mesin yang banyak digunakan - Random Forest (RF) dan Logistic Regression (LR) - untuk prediksi leukemia menggunakan kumpulan data akses terbuka dari 10.661 gambar sel darah mikroskopis yang telah diproses sebelumnya dari Kaggle. Kumpulan data tersebut secara hati-hati dipartisi menjadi set pelatihan (80%) dan pengujian (20%), dengan praproses yang ketat termasuk normalisasi gambar dan ekstraksi fitur. Evaluasi kami menggabungkan beberapa metrik kinerja: akurasi, sensitivitas, spesifisitas, dan AUC. Hasilnya menunjukkan bahwa kinerja Random Forest lebih unggul dengan akurasi klasifikasi sebesar 85,23%, spesifisitas sebesar 0,9351, sensitivitas sebesar 0,6774, dan AUC sebesar 0,8881, jauh mengungguli LR yang mencapai akurasi sebesar 78,11%, spesifisitas sebesar 0,8363, sensitivitas sebesar 0,6742, dan AUC sebesar 0,8120. Temuan ini menunjukkan bahwa metode ensemble seperti RF sangat cocok untuk mendeteksi salah satu kanker darah yang paling mematikan, leukemia, karena kemampuannya untuk menangani interaksi fitur yang kompleks dalam data pencitraan medis. Meskipun kedua algoritma tersebut berpotensi sebagai pendukung keputusan klinis, penelitian di masa mendatang dapat menguji teknik pembelajaran mendalam dan kumpulan data yang lebih besar untuk meningkatkan akurasi dan keandalan model.

Kata kunci: Random Forest, Regresi Logistik, Prediksi, leukemia, python

ABSTRACT

Leukemia is one of the deadliest blood cancers that urgently requires early detection for effective treatment. However, conventional diagnosis methods are often subjective, time-consuming, and expensive, posing challenges especially in resource-constrained areas. This study presents a comprehensive comparative analysis of two widely-used machine learning algorithms - Random Forest (RF) and Logistic Regression (LR) - for leukemia prediction using an open-access dataset of 10,661 preprocessed microscopic blood cell images from Kaggle. The dataset was carefully partitioned into training (80%) and testing (20%) sets, with rigorous preprocessing including image normalization and feature extraction. Our evaluation incorporated multiple performance metrics: accuracy, sensitivity, specificity, and AUC. The results show that Random Forest's performance is superior with a classification accuracy of 85.23%, specificity of 0.9351, sensitivity of 0.6774, and AUC of 0.8881, significantly outperforming LR which achieved an accuracy of 78.11%, specificity of 0.8363, sensitivity of 0.6742, and AUC of 0.8120. These findings suggest that ensemble methods like RF are particularly well-suited for detecting one of the most deadly blood cancers, leukemia, due to their ability to handle complex feature interactions in medical imaging data. While both algorithms have potential as clinical decision support, future research can test deep learning techniques and larger datasets to improve the accuracy and reliability of the model.

Keywords: *Random Forest, Logistic Regression, Prediction, leukemia, python*