INTISARI

PENANGANAN DATA TIDAK SEIMBANG MENGGUNAKAN *CASCADE MODELING* PADA KASUS KLASIFIKASI TUBERKULOSIS BERBASIS CITRA *CHEST X-RAY*

Oleh

Nurraudya Tuz Zahra
22/499728/PPA/06343

Di dalam diagnosis medis, seringkali terjadi ketidakseimbangan kelas karena distribusi kelas pada dataset klinis yang berukuran besar cenderung tidak seimbang. Ketidakseimbangan kelas terjadi ketika kelas dalam dataset tidak terdistribusi secara merata, dengan beberapa kelas memiliki jumlah sampel yang jauh lebih sedikit (kelas minoritas) dibandingkan kelas lainnya (kelas mayoritas). Hal ini terutama terjadi pada dataset penyakit tuberkulosis TBX 11K, dimana jumlah sampel tuberkulosis jauh lebih sedikit daripada sampel *non*-tuberkulosis. Kondisi seperti ini dapat mempengaruhi kinerja model klasifikasi dan mengakibatkan menurunnya kinerja pada kelas minoritas. Untuk menangani masalah tersebut, penelitian ini menggunakan pendekatan *Cascade Modeling* dengan metode klasifikasi *Random Forest* (RF).

Cascade Modeling merupakan metode yang terdiri dari serangkaian model atau langkah-langkah pemrosesan yang dijalankan secara berurutan. Metode *Random Forest* (RF) dimulai untuk mengklasifikasikan kelas *Non*-Tuberkulosis dan Tuberkulosis. Untuk kelas *Non*-Tuberkulosis, metode RF baru diimplementasikan untuk mengklasifikasikan kelas *Healthy* dan *Sick*&*Non*-TB. Proses tersebut diulangi sampai kelas tunggal diperoleh. Penelitian ini melakukan tahap pra-pemrosesan dengan mengubah ukuran dimensi citra dari 512×512 menjadi 224×224 agar dapat digunakan sebagai *input* pada *base* model ekstraksi fitur. Metode ekstraksi fitur yang digunakan adalah arsitektur VGG16, dan fitur-fitur tersebut kemudian dijadikan *input* untuk proses klasifikasi dengan metode *Random Forest* (RF).

Hasil penelitian menunjukkan bahwa pendekatan *Cascade Modeling* berhasil meningkatkan performa pada kelas minoritas, khususnya kelas *Active* Tuberkulosis (ATB). Model tanpa pendekatan *Cascade Model* menunjukkan efisiensi waktu yang baik dalam proses klasifikasi data jika dibandingkan dengan model dengan pendekatan *Cascade*.

Kata Kunci: *Imbalance Data*, Data *Chest X-Ray*, *Cascade Modeling*, *Convolutional Neural Network* (CNN) *feature extraction*, dan *Random Forest* (RF).

ABSTRACT

HANDLING IMBALANCED DATA USING CASCADE MODELING FOR CHEST X-RAY IMAGE BASED TUBERCULOSIS CLASSIFICATION

By

Nurraudya Tuz Zahra
22/499728/PPA/06343

In medical diagnosis, class imbalance often occurs because the distribution of classes in large clinical datasets is unbalanced. Class imbalance occurs when the classes in a dataset are not evenly distributed, with some classes having a much smaller number of samples (minority class) than other classes (majority class). This is especially the case in the TBX 11K tuberculosis disease dataset, where the number of tuberculosis samples is much less than non-tuberculosis samples. Conditions like this can affect the performance of the classification model and result in decreased performance in the minority class. To handle this problem, this research uses a Cascade Modeling approach with the Random Forest (RF) classification method.

Cascade Modeling is a method that consists of a series of models or processing steps that are executed sequentially. The Random Forest (RF) method is started for classifying the class Non-Tuberculosis and Tuberculosis. For Non-Tuberculosis, the new RF method is implemented to classify the class healthy and sick. The process is repeated until the single class is obtained. This research conducted a pre-processing stage by resizing the image dimension from 512×512 to 224×224 so that it can be used as input for the base feature extraction model. The feature extraction method used is the VGG16 architecture, and the features are then used as input for the classification process using the Random Forest (RF) method.

The research results show that the Cascade Modeling approach has succeeded in improving performance in minority classes, especially the Active Tuberculosis (ATB) class. The model without the Cascade Model approach showed good time efficiency in the data classification process when compared to the model with the Cascade approach.

Keyword: *Imbalance Data*, *Chest X-Ray Data*, *Cascade Modeling*, *Convolutional Neural Network* (CNN) *feature extraction*, dan *Random Forest* (RF).