RINGKASAN

PERBANDINGAN TEKNIK *RESAMPLING* PADA *DATASET* HAM10000 TERHADAP PERFORMA MODEL KLASIFIKASI KANKER KULIT

Ali Rohman

Kanker kulit adalah jenis kanker yang tumbuh di jaringan kulit yang ditandai dengan adanya perubahan kondisi kulit yang abnormal (lesi kulit). Selain dapat menyebabkan kematian, kanker kulit tentu akan merusak penampilan seseorang. Dan yang menambah permasalahan adalah banyak orang biasa tidak dapat membedakan antara tahi lalat dan kanker kulit *melanoma*, padahal *5-year survival rate* untuk *melanoma* jika dapat terdeteksi dini bisa mencapai 99%. Karena itulah penulis ingin membuat sistem klasifikasi kanker kulit menggunakan *dataset* HAM10000 yang mengklasifikasikan kanker kulit berdasarkan kondisi lesi kulit berpigmen menjadi 7 kelas kanker kulit, yaitu akiec, bcc, bkl, nv, mel, vasc, dan df. Yang menjadi permasalahan adalah *dataset* yang digunakan memiliki distribusi data antar kelas yang tidak seimbang, sehingga perlu dilakukan penyeimbangan kualitas data.

Model arsitektur klasifikasi dibuat sendiri menggunakan metode CNN. Untuk melakukan penyeimbangan dataset, dilakukan dengan teknik *resampling* dengan bantuan *library Python imbalanced-learn.* Pada penelitian ini dilakukan uji coba beberapa teknik *resampling* untuk mendapatkan hasil yang paling optimal. Teknik *resampling* yang digunakan antara lain, *random undersampler, cluster centroids, random oversampler*, SMOTE, ADASYN, gabungan SMOTE dengan *Tomek Link*, dan gabungan SMOTE dengan ENN.

Berdasarkan hasil pelatihan dan pengujian model arsitektur yang telah dirancang dan melewati *trial and error* hingga mencapai kualitas maksimal yang diharapkan. Di antara tiga teknik umum *resampling,* yaitu *undersampling, oversampling*, dan gabungan (*hybrid*), secara umum teknik *resampling* gabungan merupakan teknik yang memberikan hasil paling optimal. Dari beberapa teknik *resampling dataset* HAM10000, teknik *resampling* gabungan antara *oversampling* SMOTE dan *undersampling* ENN memberikan hasil paling optimal dengan nilai *validation accuracy* sebesar 0.986843*, validation precision* sebesar 0.958295*,* dan *validation recall* sebesar 0.949213.

Kata kunci : *resampling dataset*, CNN, *deep learning*, kanker kulit.

SUMMARY

COMPARISON OF RESAMPLING TECHNIQUES IN HAM10000 DATASET ON THE PERFORMANCE OF SKIN CANCER CLASSIFICATION MODEL

Ali Rohman

Skin cancer is a type of cancer that grows in the skin tissue which is characterized by abnormal changes in skin conditions (skin lesions). Besides being able to cause death, skin cancer will certainly damage a person's appearance. And what adds to the problem is that many ordinary people can't tell the difference between moles and melanoma skin cancer, whereas the 5-year survival rate for melanoma if detected early can reach 99%. That's why the author wants to create a skin cancer classification system using the HAM10000 dataset which classifies skin cancer based on the condition of pigmented skin lesions into 7 classes of skin cancer, namely akiec, bcc, bkl, nv, mel, vasc, and df. The problem is that the dataset used has an unbalanced distribution of data between classes, so it is necessary to balance the quality of the data.

The classification architecture model was created using the CNN method. To balance the dataset, a resampling technique was used with the help of the imbalanced-learn Python library. In this study, several resampling techniques were tested to get the most optimal results. The resampling techniques used include random undersampler, cluster centroids, random oversampler, SMOTE, ADASYN, combined SMOTE with Tomek Link, and combined SMOTE with ENN.

Based on the results of training and testing of architectural models that have been designed and gone through trial and error to achieve the expected maximum quality. Among the three general resampling techniques, namely undersampling, oversampling, and hybrid, in general the combined resampling technique is the technique that gives the most optimal results. From several resampling techniques for the HAM10000, the combined resampling technique between SMOTE oversampling and ENN undersampling gave the most optimal results with a validation accuracy value is 0.986843, a validation precision is 0.958295, and a validation recall is 0.949213.

Keywords: resampling dataset, CNN, deep learning, skin cancer.