INTISARI

Saat ini, jumlah data yang tersedia terus mengalami peningkatan. Ketersediaan data dalam jumlah yang besar menjadi salah satu aspek penting dalam proses latih model *Deep Learning*. Data yang memiliki jumlah distribusi kelas yang seimbang merupakan data yang ideal. Namun pada praktik di kehidupan nyata, data yang tersedia tidak selalu memiliki distribusi kelas yang seimbang. Dampaknya, model akan lebih cenderung pada kelas mayoritas dan memperbesar peluang kesalahan klasifikasi pada kelas minoritas. Karakteristik yang ditimbulkan tersebut dapat berakibat kepada performa model yang buruk pada kelas minoritas.

Pada penelitian ini, penulis akan melakukan penelitian mengenai metode-metode penyelesaian kasus data imbalance pada salah satu arsitektur *Deep Learning*, yaitu *Convolutional Neural Network* (CNN). Penulis akan melakukan komparasi terhadap metode- metode penyelesaian kasus data imbalance yang diaplikasikan pada model CNN, yaitu *oversampling*, *undersampling*, *focal loss*, dan *cost-sensitive learning*. Komparasi ini dilakukan dengan menguji performa skor F1 model CNN pada dataset yang dibuat menjadi tidak seimbang. Set distribusi data yang dibuat berasal dari dataset CIFAR-10 dengan mengurangi jumlah sampel pada kelas minoritas menjadi set data distribusi dengan rasio $\rho \in \{3, 10, 200\}$ dan fraksi kelas minoritas $\mu \in \{0.1, 0.5, 0.9\}$.

Hasil pengujian menunjukkan bahwa metode-metode penyelesaian kasus data imbalance secara umum mampu meningkatkan performa klasifikas model *Convolutional Neural Network* (CNN) pada beberapa distribusi set data yang tidak seimbang. Metode yang paling efektif dalam meningkatkan performa model pada kasus data imbalance pada penelitian ini adalah metode *cost-sensitive learning* yang mengalami peningkatan performa skor F1 model pada semua kasus distribusi dibandingkan dengan model basis. Hasil pengujian statistik juga menunjukkan bahwa *cost-sensitive learning* secara signifikan meningkatkan performa model dibandingkan dengan model lainnya.

Kata kunci: *data imbalance*, *cost-sensitive learning*, *focal loss*, *oversampling*, *undersampling*.

ABSTRACT

Currently, the available amount of data continues to increase. The availability of a large amount of data is an important aspect in the process of training Deep Learning models. Ideally, the data should have a balanced class distribution. However, in real-life practice, the available data does not always have a balanced class distribution. As a result, the model will tend to favor the majority class and increase the chances of misclassification errors in the minority class. These characteristics can lead to poor model performance on the minority class.

In this study, the author will conduct research on methods for addressing data imbalance issues in one of the Deep Learning architectures, namely Convolutional Neural Network (CNN). The author will compare the methods for addressing data imbalance issues applied to CNN models, including oversampling, undersampling, focal loss, and cost-sensitive learning. The comparison will be done by evaluating the F1 score performance of the CNN model on an imbalanced dataset. The data distribution sets used in the study are derived from the CIFAR-10 dataset by reducing the number of samples in the minority class, resulting in data distribution sets with a ratio $\rho \in 3, 10, 200$ *and a minority class fraction* $\mu \in 0.1, 0.5, 0.9$.

The test results show that the methods for addressing data imbalance issues generally improve the classification performance of Convolutional Neural Network (CNN) models on several imbalanced data distribution sets. The most effective method for improving model performance in data imbalance cases in this study is cost-sensitive learning, which enhances the F1 score performance of the model in all distribution cases compared to the baseline model. The statistical test results also indicate that cost-sensitive learning significantly improves model performance compared to other models.

Keywords : *data imbalance, cost-sensitive learning, focal loss, oversampling, undersampling.*