ABSTRACT

Multiclass skin lesion classification based on dermoscopic images still faces the challenge of class imbalance, which may cause the model to recognize majority classes more dominantly while performing less optimally on minority classes. Class imbalance refers to a condition in which the number of samples across classes is not evenly distributed. This study aims to analyze the effect of data distribution scenarios and loss functions in handling class imbalance for multiclass skin lesion classification using EfficientNetB0. The dataset used in this study is HAM10000, obtained from Kaggle. HAM10000 consists of 10,015 dermoscopic images from seven skin lesion diagnosis classes. The experiments were designed into 16 experimental scenarios obtained from the combination of four data distribution scenarios, namely original, oversampling, undersampling, and over-undersampling, and four loss functions, namely categorical cross-entropy, weighted cross-entropy, focal loss, and focal loss with class weights. Model performance was evaluated using balanced accuracy as the main metric, with G-Mean, macro precision, macro-F1, and overall accuracy as supporting metrics. The results show that changes in data distribution had a greater effect on model performance than changes in loss function. This is indicated by the range of average balanced accuracy across data distribution scenarios, which was 0.0590, higher than the range across loss functions, which was 0.0196. The best scenario on the validation data was over-undersampling with focal loss, achieving a balanced accuracy of 0.7126. The two closest validation performances were obtained by undersampling with weighted cross-entropy, with a balanced accuracy of 0.7102, and over-undersampling with categorical cross-entropy, with a balanced accuracy of 0.7082. In the final evaluation using the test data, the over-undersampling scenario with focal loss achieved a balanced accuracy of 0.7116, indicating relatively consistent performance compared with the validation result. These findings indicate that data distribution plays a more dominant role in handling class imbalance, while the loss function still affects performance when combined with an appropriate training data composition.

Keywords : class imbalance, data distribution, EfficientNetB0, HAM10000, loss function, skin lesion classification.