

ABSTRAK

PENDEKATAN *OVERSAMPLING* SMOTE UNTUK *IMBALANCED DATASET* AKSARA LAMPUNG DAN KLASIFIKASI MENGGUNAKAN SVM

oleh

EDO PRIYONO MUALIM

Klasifikasi gambar adalah pembelajaran *supervised* dengan mendefinisikan kelas gambar, kemudian melatih sebuah model menggunakan data training, validasi, dan testing untuk memprediksi label gambar tulisan tangan Aksara Lampung. Tujuan penelitian ini adalah untuk mengembangkan model *Support Vector Machine* (SVM) yang mampu mengklasifikasi kelas gambar secara akurat. Teknik klasifikasi SVM dipilih karena dapat memproses data dalam bentuk tulisan dan gambar, memiliki implementasi *kernel* sehingga memungkinkan data yang tidak dapat dibagi secara *linear* untuk dikonversikan ke data dimensi lebih tinggi. Hasil penelitian dapat digunakan sebagai panduan untuk penelitian selanjutnya dengan fokus topik yang lebih mendalam, terutama untuk subjek mengenai klasifikasi gambar. Penelitian ini menggunakan karakter tradisional dari Kota Lampung di Indonesia. Berbeda dengan alfabet latin umumnya, Bahasa Lampung memiliki 20 huruf induk, 12 anak huruf, dan tanda baca khusus. Dataset Lampung yang digunakan terdiri dari 18 huruf induk dan tersebar di 32140 gambar *grayscale* tanpa anak huruf dan tanda baca. Akan tetapi data tidak memiliki persebaran data yang seimbang. Penelitian ini menggunakan pendekatan oversampling *Synthetic Minority Oversampling Technique* (SMOTE) sebagai solusi masalah data tidak seimbang. Fitur gambar diekstraksi ke dalam *NumPy array*, dimanipulasi dan dianalisis menggunakan *library* dari bahasa pemrograman Python. Penelitian ini menghasilkan model SVM dengan akurasi terendah yaitu 92.03 dan akurasi tertinggi sebesar 95.90.

Kata Kunci : *Dataset* Lampung, Klasifikasi Gambar, *SMOTE* , Support Vector Machine

ABSTRACT

OVERSAMPLING USING SMOTE ON IMBALANCED DATASET OF LAMPUNG HANDWRITTEN IMAGES FOR SVM CLASSIFICATION

by

EDO PRIYONO MUALIM

Image classification is supervised learning by defining a set of target classes, then training a model using one partitioned dataset of training, validation, and testing which contains labeled data of Lampung handwritten images. The goal is to develop a model that could accurately classify these data with the right label using Support Vector Machine (SVM). SVM classification technique was chosen because it could process data in a form of text and images, and with the implementation of kernel, data does not have to be linearly separable since input data can be converted into high dimensional data. Information derived from this endeavor to develop an optimized classification model can also be used as references for future studies with a more specialized focus point, especially on similar subjects such as image classification. This research uses traditional characters from a city in Indonesia called Lampung. Unlike the conventional 26 Latin alphabets, Lampung has 20 characters, 12 diacritics, and unique punctuations. Lampung dataset consists of 18 Lampung characters, in the form of 32140 grayscale images without diacritics and punctuations. These data however are not evenly distributed among the 18 characters. This research makes use of oversampling approach using Synthetic Minority Oversampling Technique (SMOTE) as a solution for class imbalance found in this dataset. Lampung images features were extracted into NumPy array and are manipulated and analyzed using various libraries from Python programming language. This research yields an SVM model with the lowest accuracy of 92.03 and the highest accuracy of 95.90.

Keywords : Image Classification, Lampung Dataset, SMOTE, Support Vector Machine