

## ABSTRAK

**Arya Putra, I Putu** (2024) *Inisiasi Pembentukan Dataset Butterfly Fish Menggunakan Sintesa Data Smote Pada Fitur Hsv (Hue, Saturation, Value)*. Tesis, Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

Tesis ini sudah disetujui dan diperiksa oleh Pembimbing I: Dr. I Gede Aris Gunadi, S.Si. M.Kom. dan Pembimbing II: Dr. I Made Agus Wirawan, S.Kom., M.Cs.

*Kata-kata kunci : SMOTE, HSV (Hue, Saturation, Value), spesies butterfly fish.*

Perairan Pulau Bali merupakan salah satu perairan dengan keanekaragaman hayati tinggi, terutama dalam hal keragaman spesies ikan karang. Butterfly fish adalah salah satu spesies ikan karang yang memiliki nilai ekonomi dan estetika yang bagus, terutama bagi pecinta ikan hias. Identifikasi spesies secara tradisional membutuhkan waktu dan keahlian, sehingga diperlukan model machine learning untuk mempercepat dan mempermudah proses identifikasi. Penelitian ini menginisiasi pembentukan dataset citra butterfly fish dengan data yang diperoleh melalui akuisisi langsung di perairan Bali. Ekstraksi fitur dilakukan menggunakan metode HSV (Hue, Saturation, Value) untuk menangkap karakteristik warna dari citra, sedangkan SMOTE diterapkan untuk mengatasi ketidakseimbangan kelas dalam dataset. Hasil penelitian menunjukkan preprosesing efektif pada data terkondisi, yang berhasil mencapai performa terbaik pada model KNN maupun Decision Tree. Data terkondisi memiliki citra dengan objek yang sudah terkondisi berada di tengah citra, tentu ini memudahkan model preproses dalam cropping pada proses preprosesing. SMOTE pada data terkondisi juga efektif menyeimbangkan kelas dan meningkatkan hasil akurasi. Sebaliknya, pada data tidak terkondisi, preprocessing justru menurunkan performa model, karena objek dalam citra tidak berada di tengah akibat objek bergerak bebas. Namun, penerapan SMOTE langsung pada data tidak terkondisi tetap memberikan dampak positif signifikan. Eksperimen membandingkan performa dataset terkondisi dan tidak terkondisi menggunakan KNN dan Decision Tree, menunjukkan bahwa model terbaik dicapai pada konfigurasi "Data Original dengan SMOTE tanpa preprosesing" yang diuji dengan KNN ( $k=3$ , 10-fold cross-validation). Model ini memberikan hasil tertinggi pada semua metrik evaluasi, melebihi performa pada dataset terkondisi, yang disebabkan karena jumlah data dari dataset tidak terkondisi jauh lebih banyak ketimbang data tidak terkondisi, dengan demikian model mesin learning bisa lebih banyak belajar. Hasil juga menunjukkan bahwa peningkatan jumlah  $K$  pada KNN cenderung menurunkan akurasi karena bertambahnya data tetangga dari kelas berbeda. Penelitian ini memberikan kontribusi awal dalam pembentukan dataset optimal untuk identifikasi spesies butterfly fish di Bali yang di publish pada <https://data.mendeley.com/datasets/hsmjkd5sm/2>

**Arya Putra, I Putu (2024).** *Initiating the Development of a Butterfly Fish Dataset Using SMOTE Data Synthesis on HSV (Hue, Saturation, Value) Features.* Thesis, Computer Science, Postgraduate Program, Universitas Pendidikan Ganesha.

This thesis has been approved by Supervisor I: Dr. I Gede Aris Gunadi, S.Si., M.Kom., and Supervisor II: Dr. I Made Agus Wirawan, S.Kom., M.Cs.

*Keywords: SMOTE, HSV (Hue, Saturation, Value), butterfly fish species.*

The waters around Bali Island are known for their rich biodiversity, particularly in the variety of coral reef fish species. Butterfly fish are one of these coral reef species, valued for both their economic and aesthetic appeal, especially among ornamental fish enthusiasts. Traditional methods of species identification require significant time and expertise, prompting the need for machine learning models to expedite and simplify the identification process. This research initiated the development of a butterfly fish image dataset using data directly acquired from Bali's waters. Feature extraction was conducted using the HSV (Hue, Saturation, Value) method to capture the color characteristics of the images, while SMOTE was applied to address class imbalance in the dataset. The results demonstrated effective preprocessing on conditioned data, achieving optimal performance with both KNN and Decision Tree models. Conditioned data refers to images where objects are centrally positioned, facilitating cropping during preprocessing. SMOTE applied to conditioned data effectively balanced the classes and improved accuracy. Conversely, for unconditioned data, preprocessing negatively impacted model performance, as objects in the images were not centrally positioned due to their free movement. However, the direct application of SMOTE to unconditioned data still resulted in significant positive impacts. Experiments comparing the performance of conditioned and unconditioned datasets using KNN and Decision Tree models revealed that the best configuration was achieved with the "Original Data with SMOTE without Preprocessing" setup, evaluated using KNN ( $k=3$ , 10-fold cross-validation). This configuration yielded the highest scores across all evaluation metrics, surpassing the performance of conditioned datasets. This outcome is attributed to the larger volume of data in the unconditioned dataset, allowing the machine learning model to learn more effectively. The results also indicated that increasing the value of  $k$  in KNN tends to reduce accuracy due to the inclusion of more neighbors from different classes. This study provides an initial contribution to the development of an optimal dataset for butterfly fish species identification in Bali. The dataset is publicly accessible at <https://data.mendeley.com/datasets/hsmjkd5sm/2>.