ABSTRAK

KLASIFIKASI GEN ESENSIAL PADA *DROSOPHILA MELANOGASTER* DENGAN METODE *ADABOOST* DAN *XGBOOST* MENGGUNAKAN DATA *SEQUENCE* DNA DAN PROTEIN

Oleh

Gendis Ananda Putri

Gen esensial sangat penting untuk kelangsungan hidup dan fungsi normal suatu organisme. Studi ini bertujuan membandingkan kinerja algoritma *AdaBoost* dan *XGBoost* dalam mengklasifikasikan *Gen Esensial Seluler* (CEG) dan *Gen Esensial Organisme* (OEG) pada *Drosophila melanogaster*. Dataset diambil dari dataset CLEARER yang telah dikurasi oleh Beder et al. (2021), terdiri dari 11.547 gen untuk label CEG setelah proses pra-pemrosesan. Fitur diekstraksi menggunakan Komposisi Asam Amino (AAC) dari urutan protein, Komposisi *Tri-Nukleotida* (TNC) dan *Transformasi Fourier* (FT) dari urutan DNA, serta *PPI_degree*, sehingga menghasilkan 185 fitur awal. Sebanyak 45 fitur teratas dipilih menggunakan *Random Forest Gini Importance*. Untuk menangani ketidakseimbangan kelas yang parah (rasio 1:8,41 untuk CEG), teknik SMOTETomek diterapkan hanya pada data pelatihan setelah dilakukan pembagian stratifikasi 90:10. Model dilatih dengan validasi silang *10-fold* dan optimasi *threshold probabilitas*. Hasil menunjukkan bahwa *XGBoost* dengan SMOTETomek mencapai performa terbaik pada label CEG: akurasi 96,88%, skor F1 0,9498, MCC 0,9400, ROC-AUC 0,9888, dan PR-AUC 0,9869. Pada label OEG, *XGBoost* mencapai akurasi 94,98%. Sebagai perbandingan, AdaBoost memperoleh akurasi 85,00% pada CEG dan 84,15% pada OEG. Kesalahan klasifikasi umumnya terjadi akibat kemiripan pola urutan antara gen esensial dan non-esensial. Studi ini menunjukkan bahwa kombinasi *XGBoost* dengan SMOTETomek dan seleksi fitur berbasis *Gini Importance* secara efektif mengatasi ketidakseimbangan kelas dan memberikan performa yang kompetitif untuk klasifikasi gen esensial berbasis urutan pada *Drosophila melanogaster*.

Kata Kunci: Gen esensial, *AdaBoost*, *XGBoost*, *Drosophila melanogaster*, SMOTETomek

ABSTRACT

CLASSIFICATION OF ESSENTIAL GENES IN *DROSOPHILA MELANOGASTER* USING *ADABOOST* AND *XGBOOST* METHODS BASED ON DNA AND PROTEIN SEQUENCE DATA

By

Gendis Ananda Putri

Essential genes are critical for the survival and normal function of an organism. This study aimed to compare the performance of AdaBoost and XGBoost algorithms in classifying Cellular Essential Genes (CEG) and Organismal Essential Genes (OEG) in Drosophila melanogaster. The dataset was obtained from the curated CLEARER dataset by Beder et al. (2021), which, after preprocessing, consists of 11,547 genes for the CEG label. Features were extracted from protein sequences using Amino Acid Composition (AAC), from DNA sequences using Tri-Nucleotide Composition (TNC) and Fourier Transform (FT), and from PPI_degree, resulting in 185 initial features. The top 45 features were selected using Random Forest Gini Importance. To handle severe class imbalance (ratio 1:8.41 for CEG), the SMOTETomek technique was applied only to the training set after a 90:10 stratified split. Models were trained with 10-fold cross-validation and an optimized probability threshold. Results showed that XGBoost with SMOTETomek achieved the best performance on the CEG label: 96.88% accuracy, 0.9498 F1-score, 0.9400 MCC, 0.9888 ROC-AUC, and 0.9869 PR-AUC. On the OEG label, XGBoost reached 94.98% accuracy. In comparison, AdaBoost obtained 85.00% accuracy on CEG and 84.15% on OEG. Misclassifications mainly occurred due to sequence pattern similarities between essential and non-essential genes. This study demonstrates that combining XGBoost with SMOTETomek and Gini Importance-based feature selection effectively addresses class imbalance and provides competitive performance for sequence-based essential gene classification in Drosophila melanogaster.

Keywords: Essential genes, *AdaBoost*, *XGBoost*, *Drosophila melanogaster*, SMOTETomek