

## ABSTRAK

### KLASIFIKASI GEN ESENSIAL PADA *DROSOPHILA MELANOGASTER* BERDASARKAN PROTEIN *SEQUENCE* MENGGUNAKAN METODE *LONG SHORT-TERM MEMORY (LSTM)*

Oleh

Dina Putri Aulia

Gen esensial adalah gen yang sangat diperlukan untuk mendukung kehidupan organisme seluler. Penghapusan gen esensial pada organisme akan mengakibatkan kematian dan infertilitas. Mengklasifikasikan gen esensial ada 2 cara, yaitu dengan teknik eksperimental dan komputasi, namun teknik eksperimental membutuhkan sumber daya yang besar. Pada penelitian ini mengusulkan metode *Long Short-Term Memory (LSTM)* untuk mengklasifikasikan gen esensial pada *Drosophila melanogaster* berdasarkan protein *sequence*. *Drosophila melanogaster* adalah salah satu organisme yang sering digunakan untuk analisis *science*. Tujuan penelitian ini, yaitu mengukur dan mengklasifikasi hasil penelitian menggunakan algoritma *Long Short-Term Memory (LSTM)* dan membandingkan hasilnya dengan penelitian terdahulu oleh Beder, et al, (2021). Data dibagi dengan 2 skenario, yaitu : 80% *training* 20% *validation* ; dan 90% *training* 10% data *validation*. Pada dataset CEG memiliki distribusi kelas yang tidak seimbang sehingga dilakukan proses *Random Undersampling (RUS)* untuk menyeimbangkan kelas. Evaluasi model dilakukan dengan metrik evaluasi, yaitu PR-AUC, ROC-AUC, *sensitivity* dan *specificity*. Hasil kinerja yang paling baik pada dataset OEG didapatkan pada skenario pembagian data 80% *training* dan 20% *validation*, dengan nilai *sensitivity* 81%, *specificity* 76%, ROC-AUC 79% , PR-AUC 82%. Hasil yang paling baik pada dataset CEG diperoleh dari pembagian data 80% *training* dan data 20% *validation* dengan nilai yang didapat untuk *sensitivity* 73%, *specificity* 50% ROC-AUC 61%, dan PR-AUC 45%.

**Kata Kunci** : Gen Esensial, Klasifikasi, *Drosophila melanogaster*, Protein *sequence*, LSTM.

## **ABSTRACT**

### **CLASSIFICATION OF ESSENTIAL GENES IN DROSOPHILA MELANOGASTER BASED ON PROTEIN SEQUENCE USING LONG SHORT-TERM MEMORY UNIT (LSTM)**

**Oleh**

**Dina Putri Aulia**

Essential genes are genes that are very necessary to support the life of cellular organisms. Deletion of essential genes in an organism will result in death and infertility. There are 2 ways to classify essential genes, specifically experimental and computational techniques, but experimental techniques require large resources. In this study, we propose the Long Short-Term Memory (LSTM) method to classify essential genes in *Drosophila melanogaster* based on protein sequences. *Drosophila melanogaster* is an organism that is often used for scientific analysis. The purpose of this research is to measure and classify research results using the Long Short-Term Memory (LSTM) algorithm and compare the results with previous research by Beder, et al, (2021). The data is divided into 2 scenarios, specifically 80% training 20% validation and 90% training 10% data validation. The CEG dataset has an unbalanced class distribution so a Random Undersampling (RUS) process is carried out to balance the classes. Model evaluation was carried out using evaluation metrics, specifically PR-AUC, ROC-AUC, sensitivity and specificity. The best performance results on the OEG dataset were obtained in the data sharing scenario of 80% training and 20% validation, with sensitivity values of 81%, specificity 76%, ROC-AUC 79%, PR-AUC 82%. The best results on the CEG dataset were obtained from dividing 80% training data and 20% validation data with values obtained for sensitivity 73%, specificity 50% ROC-AUC 61%, and PR-AUC 45%

**Keyword** : Gen Essential, Classification, *Drosophila melanogaster*, Protein sequence, LSTM.