ABSTRAK

KLASIFIKASI *IMBALANCED DATA* UNTUK *POST-TRANSLATIONAL MODIFICATION* (PTM) PADA *SEQUENCES* PROTEIN LISIN MENGGUNAKAN *ADAPTIVE SYNTHETIC* (ADASYN) *SAMPLING* DENGAN METODE *LIGHT GRADIENT BOOSTING MACHINE* (LightGBM)

Oleh

ARDELLA DEAN AWALIA

Crotonylation merupakan salah satu jenis modifikasi pascatranslasi (PTM) berupa penambahan gugus asil pada asam amino lisin. Modifikasi ini memiliki kemampuan dalam mengatur ekspresi gen dan ditemukan terlibat pada beberapa penyakit, seperti depresi, ginjal kronis, hingga kanker. Identifikasi situs PTM menjadi hal krusial mengingat perannya dalam siklus sel. Metode machine learning untuk klasifikasi situs PTM dapat digunakan sebagai alternatif dalam mengenali situs PTM, namun membutuhkan data yang relatif banyak agar dapat memberikan hasil yang andal. Penelitian ini dilakukan menggunakan metode klasifikasi LightGBM dan teknik *oversampling* ADASYN untuk menangani ketersediaan data protein *crotonylation* yang cukup sedikit. Data yang digunakan diperoleh dari situs UniProt terdiri dari 159 data positif dan 847 data negatif. Ekstraksi fitur menggunakan *binary encoding, position weight amino acid, encoding based on grouped weight, k-nearest neighbors,* dan *pseudo-position specific scoring matrix* menghasilkan 833 fitur. 5-*fold cross-validation* digunakan pada proses *training* untuk mencari kombinasi *hyperparameter* terbaik. Hasil penelitian menunjukkan bahwa pembagian data menggunakan 90% data sebagai data latih dan 10% data sebagai data uji memberikan hasil tertinggi dengan nilai *accuracy* sebesar 96,04%, *sensitivity* sebesar 87,50%, *specificity* sebesar 97,65%, MCC sebesar 85,15%, dan AUC sebesar 98,90%.

Kata kunci: modifikasi pascatranslasi, *crotonylation,* ADASYN, LightGBM

ABSTRACT

CLASSIFICATION OF IMBALANCED DATA FOR POST-TRANSLATIONAL MODIFICATION ON LYSINE PROTEIN SEQUENCES USING ADAPTIVE SYNTHETIC SAMPLING WITH LIGHT GRADIENT BOOSTING MACHINE

By

ARDELLA DEAN AWALIA

Crotonylation is a type of post-translational modification (PTM). It is an addition of acyl group to the lysine residues. This modification has the ability to regulate gene expression and has been found to be involved in several diseases, such as depression, chronic kidney disease, and cancer. Identification of PTM sites is crucial considering their role in the cell cycle. Machine learning methods for classifying PTM sites can be used as an alternative for recognizing PTM sites, but they require relatively large amounts of data to provide reliable results. This research was carried out using the LightGBM classification method and the ADASYN oversampling technique to handle the limited availability of crotonylated protein sequences. The data was obtained from the UniProt website consisting of 159 positive data (crotonylated) and 847 negative data (noncrotonylated). Feature extraction by using binary encoding, position weight amino acid, encoding based on grouped weight, k-nearest neighbors, and pseudo-position specific scoring matrix produced 833 features. 5-fold cross validation was used in the training process to find best hyperparameter combinations. The results showed that the highest result was obtained by using 90% of the data as training data with the application of ADASYN oversampling (n_neighbors=9) and 10% of the data as test data with 96,04% accuracy, 87,50% sensitivity, 97.65% specificity, 85,15% MCC, and 98,90% AUC.

Keywords: post-translational modification, *crotonylation*, ADASYN, LightGBM