

ABSTRAK

Diabetes Melitus Tipe 2 (DMT2) merupakan penyakit metabolik kronis yang berkembang melalui interaksi faktor lingkungan dan faktor genetik. Susunan informasi DNA dapat memuat pola yang berkaitan dengan kerentanan terhadap DMT2, tetapi data tersebut bersifat simbolik, berdimensi tinggi, dan tidak dapat langsung digunakan sebagai masukan algoritma klasifikasi. Penelitian ini mengklasifikasikan susunan informasi DNA pada kelas DMT2 dan non-diabetes (NONDM) menggunakan variasi *k-mer* untuk membandingkan metode *Nu-Support Vector Classification* (NuSVC) dan *Extreme Gradient Boosting* (XGBoost).

Data awal terdiri atas 1.296 entri DMT2 dan 1.000 entri NONDM dalam format FASTA. Pemeriksaan data dilakukan melalui validasi karakter nukleotida, pengeluaran susunan informasi DNA dengan konflik label, dan penghapusan duplikasi identik. Setelah pembersihan, diperoleh 1.615 entri yang terdiri atas 629 DMT2 dan 986 NONDM. Data dibagi secara terstratifikasi menjadi 1.292 data latih dan 323 data uji sebelum pembentukan fitur. Susunan informasi DNA direpresentasikan menggunakan *k-mer* dengan $k = 3$, $k = 4$, dan $k = 5$, kemudian diberi bobot *Term Frequency-Inverse Document Frequency* (TF-IDF). Penyeimbangan kelas menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan hanya pada bagian pelatihan. Pemilihan nilai k dan model dilakukan melalui *10-fold cross-validation* pada data latih.

Hasil validasi silang memilih $k = 5$ untuk NuSVC dan XGBoost. XGBoost menjadi model terpilih dengan rata-rata *accuracy* sebesar 95,28%, AUC sebesar 0,9869, dan *log loss* sebesar 0,1391. Pada data uji, XGBoost memperoleh *accuracy* sebesar 94,43%, *precision* kelas DMT2 sebesar 95,00%, sensitivitas sebesar 90,48%, spesifisitas sebesar 96,95%, *F1-score* sebesar 92,68%, AUC sebesar 0,9915, dan *log loss* sebesar 0,1151. NuSVC memperoleh *accuracy* sebesar 89,16%, AUC sebesar 0,9606, dan *log loss* sebesar 0,2525.

Berdasarkan hasil tersebut, XGBoost dengan $k = 5$ memberikan kinerja klasifikasi yang lebih baik secara deskriptif dibandingkan NuSVC pada kumpulan data yang digunakan. Pemisahan data uji sebelum pembentukan TF-IDF dan penerapan SMOTE menjaga agar proses pemilihan model tidak menggunakan informasi dari data uji. Meskipun demikian, hasil penelitian masih terbatas pada tingkat entri susunan informasi DNA dan belum dapat diperlakukan sebagai alat diagnosis klinis maupun bukti biomarker genetik tanpa validasi biologis dan pengujian pada data eksternal.

Kata kunci: Diabetes Melitus Tipe 2; susunan informasi DNA; *k-mer*; TF-IDF; NuSVC; XGBoost.

ABSTRACT

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disease that develops through interactions between environmental and genetic factors. DNA information arrangements may contain patterns associated with susceptibility to T2DM; however, these data are symbolic, high-dimensional, and cannot be used directly as input for classification algorithms. This study classifies DNA information arrangements into T2DM and non-diabetes (NONDM) classes using variations of *k-mer* representations to compare the *Nu-Support Vector Classification* (NuSVC) and *Extreme Gradient Boosting* (XGBoost) methods.

The initial data consisted of 1,296 T2DM entries and 1,000 NONDM entries in FASTA format. Data examination included nucleotide-character validation, removal of entries with conflicting labels, and elimination of identical duplicates. After cleaning, 1,615 entries remained, consisting of 629 T2DM and 986 NONDM entries. The data were stratified into 1,292 training entries and 323 test entries before feature construction. DNA information arrangements were represented using *k-mers* with $k = 3$, $k = 4$, and $k = 5$, followed by *Term Frequency-Inverse Document Frequency* (TF-IDF) weighting. Class balancing using the *Synthetic Minority Over-sampling Technique* (SMOTE) was applied only to the training portion. The values of k and the model were selected through 10-fold cross-validation on the training data.

Cross-validation selected $k = 5$ for both NuSVC and XGBoost. XGBoost was selected as the primary model, with a mean accuracy of 95.28%, an AUC of 0.9869, and a log loss of 0.1391. On the independent test data, XGBoost achieved an accuracy of 94.43%, a T2DM precision of 95.00%, a sensitivity of 90.48%, a specificity of 96.95%, an F1-score of 92.68%, an AUC of 0.9915, and a log loss of 0.1151. NuSVC achieved an accuracy of 89.16%, an AUC of 0.9606, and a log loss of 0.2525.

Based on these results, XGBoost with $k = 5$ provided better descriptive classification performance than NuSVC on the dataset used in this study. Separating the test data before TF-IDF construction and SMOTE application ensured that model selection did not use information from the test data. Nevertheless, the findings remain limited to DNA-information-arrangement entries and should not be interpreted as a clinical diagnostic tool or evidence of genetic biomarkers without biological validation and evaluation on external data.

Keywords: Type 2 Diabetes Mellitus; DNA information arrangement; *k-mer*; TF-IDF; NuSVC; XGBoost.