

ABSTRAK

Kecelakaan lalu lintas merupakan masalah global dengan tingkat keparahan yang dipengaruhi oleh berbagai faktor. Penelitian ini bertujuan membandingkan performa algoritma *XGBoost* dan *Autoencoder-XGBoost* untuk klasifikasi tingkat keparahan kecelakaan lalu lintas di Chicago menggunakan data tahun 2020–2025. Tahapan penelitian meliputi *preprocessing* data, *feature engineering*, *feature selection* berdasarkan *feature importance XGBoost*, penanganan ketidakseimbangan kelas menggunakan *SMOTE*, serta *hyperparameter tuning* menggunakan *Grid Search* dan *5-Fold Cross Validation*. Hasil penelitian menunjukkan bahwa penerapan *feature selection* dan *SMOTE* meningkatkan Recall kelas *Injury* dari 51,27% menjadi sekitar 71%. Model *Tuning XGBoost* menghasilkan performa terbaik dengan Akurasi sebesar 77,97%, Recall sebesar 71,04%, F1-Score sebesar 65,59%, dan ROC AUC sebesar 83,74%. Evaluasi menggunakan perbandingan performa data *train* dan *test*, *learning curve*, serta *5-Fold Cross Validation* menunjukkan bahwa model tidak menunjukkan indikasi *overfitting* yang signifikan dan mampu mempertahankan performa yang relatif konsisten pada data yang belum pernah dilihat sebelumnya. Sementara itu, penggunaan *Autoencoder* sebagai *feature extractor* belum mampu memberikan peningkatan performa yang konsisten dibandingkan *XGBoost*, karena fitur hasil *feature selection* telah cukup representatif untuk proses klasifikasi sehingga informasi tambahan dari representasi laten *Autoencoder* menjadi terbatas. Dengan demikian, *XGBoost* yang dikombinasikan dengan *feature selection* dan *SMOTE* lebih efektif dibandingkan *Autoencoder-XGBoost* untuk klasifikasi tingkat keparahan kecelakaan pada dataset yang digunakan dalam penelitian ini.

Kata kunci: *XGBoost*, *Autoencoder*, klasifikasi tingkat keparahan kecelakaan, *feature selection*, *SMOTE*.

ABSTRACT

Traffic accidents are a global problem, and their severity is influenced by various factors. This study aims to compare the performance of the XGBoost and Autoencoder-XGBoost algorithms for classifying the severity of traffic accidents in Chicago using data from 2020–2025. The research stages included data preprocessing, feature engineering, feature selection based on XGBoost feature importance, handling class imbalance using SMOTE, and hyperparameter tuning using Grid Search and 5-Fold Cross Validation. The results show that the application of feature selection and SMOTE improved the recall for the “Injury” class from 51.27% to approximately 71%. The tuned XGBoost model achieved the best performance with an accuracy of 77.97%, a recall of 71.04%, an F1-score of 65.59%, and an ROC AUC of 83.74%. Evaluation using a comparison of training and test set performance, learning curves, and 5-fold cross-validation showed that the model did not exhibit significant signs of overfitting and was able to maintain relatively consistent performance on previously unseen data. Meanwhile, the use of an autoencoder as a feature extractor has not been able to consistently improve performance compared to XGBoost, because the features resulting from feature selection are already sufficiently representative for the classification process, so the additional information provided by the autoencoder’s latent representation is limited. Thus, XGBoost combined with feature selection and SMOTE is more effective than the Autoencoder-XGBoost approach for classifying accident severity levels on the dataset used in this study.

Keywords: XGBoost, Autoencoder, accident severity classification, feature selection, SMOTE.