ABSTRAK

Kekerasan yang terekam kamera pengawas dan tersebar di media sosial semakin marak, sementara deteksinya masih bergantung pada pemantauan manual yang rentan terhadap penurunan kewaspadaan manusia. Penelitian ini bertujuan merancang, melatih, dan mengevaluasi model deteksi kekerasan pada video yang akurat sekaligus ringan agar dapat diterapkan pada perangkat dengan sumber daya terbatas. Model dirancang sebagai arsitektur *hybrid MobileNetV3-Small* dan *Bidirectional Long Short-Term Memory* (*Bi-LSTM*) dengan mekanisme *attention* opsional; *MobileNetV3-Small* mengekstraksi fitur spasial tiap *frame*, sedangkan *Bi-LSTM* memodelkan dependensi temporal antar-*frame* secara dua arah. Model dilatih pada gabungan *dataset* RLVS dan RWF-2000 (±4.000 video) yang dibagi per sumber dengan rasio 70% latih, 15% validasi, dan 15% uji. Konfigurasi terbaik dicari secara otomatis menggunakan *Optuna* dengan *sampler Tree-structured Parzen Estimator* dan *Median Pruner* sebanyak 144 *trial*, lalu model final ditetapkan melalui prosedur dua tahap: evaluasi ulang pada data validasi dilanjutkan pelaporan satu kali pada data uji. Pada 600 data uji, model memperoleh akurasi 0,92 dengan pola kesalahan yang menguntungkan untuk deteksi kekerasan (*false negative* lebih kecil daripada *false positive*). Waktu inferensi rata-rata sekitar 12,1 ms per video menunjukkan tahap klasifikasi bukan penghambat penerapan *real-time*. Disimpulkan bahwa model yang diusulkan akurat sekaligus efisien pada perangkat terbatas. Penelitian lanjutan disarankan melakukan pengujian lintas-*dataset*, memperluas deteksi ke ragam jenis kekerasan, serta menerapkan model pada perangkat *edge* nyata.

Kata kunci: deteksi kekerasan, *MobileNetV3-Small*, *Bi-LSTM*, *Optuna*, *deep learning*

ABSTRACT

The prevalence of violence recorded by surveillance cameras and shared on social media is increasing, yet its detection heavily relies on manual monitoring, which is susceptible to human fatigue. This study aims to design, train, and evaluate a video violence detection model that is both accurate and lightweight for implementation on resource-constrained devices. The model utilizes a hybrid architecture comprising MobileNetV3-Small and Bidirectional Long Short-Term Memory (Bi-LSTM) with an optional attention mechanism. MobileNetV3-Small extracts spatial features from each frame, while the Bi-LSTM bidirectionally models the temporal dependencies between frames. The model was trained on a combined dataset of RLVS and RWF-2000 (±4,000 videos), distributed by source into 70% training, 15% *validation*, and 15% testing sets. Hyperparameter tuning was automated using Optuna with a Tree-structured Parzen Estimator sampler and Median Pruner across 144 trials. The final model was established through a two-stage procedure: re-evaluation on the *validation* set followed by a single evaluation on the test set. Tested on 600 videos, the model achieved an accuracy of 0.92, demonstrating a favorable error pattern for violence detection (fewer false negatives than false positives). Furthermore, an average inference time of 12.1 ms per video indicates that the classification stage does not hinder real-time applications. In conclusion, the proposed model is highly accurate and efficient for limited-resource environments. Future research should consider cross-dataset testing, broadening the scope to various forms of violence, and deploying the model on physical edge devices.

Keywords: violence detection, MobileNetV3-Small, Bi-LSTM, Optuna, deep learning