ABSTRAK

Kemajuan kecerdasan buatan telah mendorong kemunculan *deepfake*, yaitu manipulasi citra yang sangat realistis dengan memanfaatkan teknologi *Generative Adversarial Networks*. Penyalahgunaan *deepfake* menimbulkan ancaman serius seperti penyebaran disinformasi dan penipuan identitas. Namun, model *Convolutional Neural Networks* konvensional sering kali tidak dapat mempertahankan akurasi tinggi saat mengklasifikasikan citra manipulasi generasi terbaru. Arsitektur mutakhir ConvNeXt memiliki kemampuan ekstraksi fitur yang lebih baik, namun rentan mengalami *overfitting* selama masa pelatihan. Oleh karena itu, penelitian ini mengevaluasi performa ConvNeXt dalam mengklasifikasikan citra wajah asli dan wajah *deepfake* dengan menerapkan *hyperparameter tuning* menggunakan metode *Grid Search* serta teknik regularisasi untuk mencegah terjadinya *overfitting*.

Penelitian ini menggunakan dataset publik *Kaggle* berisi 12.890 citra, yang terdiri atas 5.890 wajah asli dan 7.000 wajah *deepfake* yang dihasilkan oleh arsitektur *StyleGAN* dan StyleGAN2. Pra-pemrosesan diawali dengan pembagian proporsi data menjadi 80% data latih, 10% validasi, dan 10% uji. Citra kemudian diubah ukurannya menjadi 224x224 piksel, dinormalisasi, dan diaugmentasi melalui pembalikan horizontal serta rotasi acak khusus pada himpunan data latih. Penelitian ini menggunakan arsitektur ConvNeXt-Base melalui pendekatan *transfer learning* dengan strategi *full fine-tuning*. Lapisan klasifikasi akhir dimodifikasi menjadi satu neuron keluaran untuk keperluan klasifikasi biner. Pelatihan model menggunakan fungsi kerugian *BCEWithLogitsLoss* yang menggabungkan operasi *Sigmoid* dan *Binary Cross-Entropy* dalam satu komputasi serta *optimizer* AdamW, dipadukan dengan *Grid Search* untuk mencari kombinasi *learning rate* dan *batch size* yang paling optimal, serta penerapan mekanisme *early stopping*.

Hasil pengujian membuktikan bahwa implementasi ConvNeXt dengan pendekatan tersebut dapat mencegah *overfitting* dan memberikan hasil klasifikasi yang sangat optimal. Konfigurasi terbaik berdasarkan pengujian *Grid Search* dicapai pada penggunaan *learning rate* $1\text{x}10^{-5}$, *batch size* 16, dan tingkat augmentasi rotasi 30 derajat. Eksperimen perbandingan menunjukkan bahwa penerapan augmentasi menghasilkan akurasi validasi 99,92% dengan validation loss 0,0025, lebih baik dibandingkan pelatihan tanpa augmentasi yang menghasilkan akurasi validasi 99,69% dengan validation loss 0,0150, membuktikan kontribusi augmentasi terhadap kemampuan generalisasi model. Evaluasi model pada himpunan data uji mendapatkan capaian akurasi 99,53%, presisi kelas Real 0,9966 dan kelas Fake 0,9943, nilai *recall* kelas Real 0,9932 dan kelas Fake 0,9971, F1-*Score* 99,53% (*macro avg*), serta nilai *Area Under the Curve* mencapai 0,9952. Penelitian ini menunjukkan tingkat kesalahan prediksi yang sangat minim, di mana hanya terdapat 2 kesalahan prediksi pada kelas *Fake* (*False Positive*) dan 4 kesalahan prediksi pada kelas wajah asli (*False Negative*), dari total 1.289 citra uji.

Kata Kunci: ConvNeXt, Deepfake, Grid Search, Klasifikasi Citra, Overfitting

ABSTRACT

The advancement of artificial intelligence has driven the emergence of deepfakes, highly realistic image manipulations utilizing Generative Adversarial Networks technology. The misuse of deepfakes poses serious threats such as the spread of disinformation and identity fraud. However, conventional Convolutional Neural Networks models often fail to maintain high accuracy when classifying new generations of manipulated images. The state-of-the-art ConvNeXt architecture possesses better feature extraction capabilities but is prone to experiencing overfitting during the training phase. Therefore, this research evaluates the performance of ConvNeXt in classifying real and deepfake face images by applying hyperparameter tuning using the Grid Search method and regularization techniques to prevent overfitting.

This research utilized a public Kaggle dataset containing 12,890 images, consisting of 5,890 real faces and 7,000 deepfake faces generated by the *StyleGAN* and *StyleGAN2* architectures. Pre-processing began by dividing the data proportion into 80% training data, 10% validation, and 10% test data. The images were then resized to 224x224 pixels, normalized, and augmented through horizontal flipping as well as random rotation specifically on the training dataset. This study employed the ConvNeXt-Base architecture through a transfer learning approach with a full fine-tuning strategy. The final classification layer was modified into a single output neuron for binary classification purposes. The model training utilized the *BCEWithLogitsLoss* function which integrates the *Sigmoid* operation and Binary Cross-Entropy into a single, numerically stable computation alongside the AdamW optimizer, combined with Grid Search to find the most optimal combination of learning rate and batch size, alongside the implementation of an early stopping mechanism.

The test results proved that the implementation of ConvNeXt using this approach successfully prevented overfitting and provided highly optimal classification results. The best configuration based on the Grid Search evaluation was achieved using a learning rate of $1\text{x}10^{-5}$, a batch size of 16, and a rotation augmentation degree of 30. A comparative experiment demonstrated that training with augmentation yielded a validation accuracy of 99.92% with a validation loss of 0.0025, outperforming training without augmentation which achieved 99.69% validation accuracy with a validation loss of 0.0150, confirming the contribution of augmentation to the model's generalization capability. The model evaluation on the test dataset achieved an accuracy of 99.53%, precision of 0.9966 (Real) and 0.9943 (Fake), recall of 0.9932 (Real) and 0.9971 (Fake), macro avg F1-Score of 99.53%, and an Area Under the Curve value of approximately 0.9952. This research demonstrated a very minimal prediction error rate, with only 2 misclassifications in the Fake class (False Positive) and 4 in the real face class (False Negative) out of 1,289 test images.

Keywords: *ConvNeXt, Deepfake, Grid Search, Image Classification, Overfitting*