BAB I
PENDAHULUAN

1.1 Latar Belakang

Kasus kejahatan siber pada saat ini khususnya berbasis *Deepfake* kini berkembang semakin kompleks, dengan munculnya teknik manipulasi hibridia yang menggabungkan kecerdasan buatan dengan campur tangan manusia seperti proses penyuntingan lanjutan (*post-processing*) seperti penghalusan (*smoothing*) atau pengaburan (*blurring*) untuk menutupi jejak artefak visual kasar. Fenomena ini menjadi ancaman serius bagi keamanan siber karena dapat digunakan untuk menyebarkan disinformasi yang angat meyakinkan [1].

Secara teknis, proses penyuntingan pasca-produksi ini sering kali menghilangkan artefak frekuensi tinggi yang menjadi indikator utama pemalsuan. Hal ini dikonfirmasi oleh [2], yang menyatakan bahwa operasi pengolahan citra konvensional dapat menghancurkan residu forensik, menyebabkan detektor standar gagal membedakan konten asli dan palsu. Akibatnya, metode deteksi konvensional kehilangan akurasi secara signifikan saat menghadapi video yang telah melalui tahap penyempurnaan visual oleh manusia [3].

Untuk mengatasi kendala tersebut, penelitian ini mengembangkan *Forensic Discriminator* menggunakan arsitekktur XceptionNet. Pemilihan model ini merujuk pada standar *benchmark* FaceForensics++ oleh [4], yang membuktikan bahwa XceptionNet memiliki performa kompetitif dan efisiensi perameter yang lebih baik dibandingkan arsitektur VGG19 atau ResNet50 [5]. XceptionNet terbukti efektif dalam mengekstraksi fitur spasial wajah secara cepat, menjadikannya fondasi yang ideal untuk sistem deteksi efisien.

Kebaruan penelitian ini terletak pada integrasi strategi *Defense Augmentation* yang mensimulasikan efek *post-processing* manusia (*Gaussian, Median, Bilateral*). Pendekatan ini mengadopsi strategi Stanciu & Lonescu [6], yang membuktikan bahwa augmentasi gangguan (*corruption*) agresif adalah kunci meningkatkan generalisasi model (*robustness*) terhadap serangan visual tak

dikenal. Dengan demikian, model diharapkan mampu mendeteksi jejak manipulasi halus sekalipun tekstur wajah telah disamarkan.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan sebelumnya, perumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana efektivitas kinerja arsitektur *XceptionNet* dalam mengklasifikasikan video *Deepfake* dan *Real* pada dataset campuran yang mengandung variasi berupa manipulasi *post-processing* (*smoothing* dan *blurring*)?
2. Apakah strategi distribusi data latih yang menggabungkan *deepfake* mentah (*raw*) dan *deepfake* termanipulasi efektif dalam meningkatkan ketahanan (*robustness*) model terhadap upaya penyembunyiaan jejak digital?
3. Bagaimana kapabilitas generalisasi model ketika diuji terhadap data video eksternal (*wild data*) yang memiliki karakteristik kompresi dan metode generasi yang tidak diketahui?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan tidak menyimpang dari tujuan utama, peneliti menetapkan beberapa batasan masalah sebagai berikut:

1. Dataset Penelitian: Dataset yang digunakan merupakan gabungan dari *FaceForensics++*, Celeb-DF, dan DFDC. Data dikelompokkan menjadi dua kelas utama, yaitu Kelas *Real* (video asli tanpa manipulasi) dan Kelas *Fake* (gabungan antara video *deepfake* mentah dan video *deepfake* yang telah dimanipulasi).
2. Fokus Manipulasi: Simulasi *post-processing* yang diterapkan pada 50% data *Fake* (Khususnya dataset *FaceForensics++* dan Celeb-DF) menggunakan teknik *Gaussian Blur, Bilateral Filter,* dan *Median Blur,* untuk melatih ketahanan model terhadap upaya penyamaran artefak visual.
3. Metode Algoritma: Sistem klasifikasi dibangun menggunakan algoritma *Deep Learning* dengan arsitektur *XceptionNet* yang menerapkan teknik

Transfer Learning. Penelitian ini tidak membandingkan efektivitas seluruh jenis arsitektur CNN yang ada, melainkan fokus pada optimalisasi XceptionNet sebagai discriminator forensik.

4. Lingkup Analisis: Analisis dilakukan secara spasial berbasis *frame* (citra wajah) yang telah dinormalisasi ke resolusi 299x299 piksel. Penelitian ini tidak membahas analisis temporal (gerakan antar-*frame*) maupun analisis audio (suara) dalam video.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Menganalisis Kinerja Arsitektur XceptionNet pada *Data Post-Processing*: Mengukur dan menganalisis performa model XceptionNet (berdasarkan metrik *accuracy, loss, precision, recall,* dan *F1-score*) dalam mendeteksi serta mengklasifikasikan video *deepfake* yang telah mengalami manipulasi *post-processing* spesifik (seperti *smoothing* dan *bluring*).
2. Mengevalusai Efektivitas Strategi Data *Augmentation*: Membuktikan apakah strategi distribusi data latih yang menggabungkan *deepfake* mentah (*raw*) dan *deepfake* termanipulasi mampu meningkatkan sensitivitas model dalam mengenali anomali tekstur wajah yang telah disamarkan, serta mencegah terjadinya *overfitting* pada artefak tertentu.
3. Menguji Robustitas Model pada Skenario *Cross-Dataset*: Mengetahui tingkat keandalan dan generalisasi model ketika dihadapkan pada data video liar (*wild data*) yang memiliki karakteristik kompresi dan metode generasi yang berbeda dari data pelatihan.

1.5 Manfaat Penelitian

Berdasarkan tujuan yang ingin dicapai, penelitian ini dihaarapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat Teoritis

   a. Pengayaan Literatur Forensik Digital: Memberikan kontribusi akademis terkait efektivitas arsitektur XceptionNet dalam mendeteksi video *deepfake* yang telah mengalami upaya penyembunyian jejak (*anti-forensics*) melalui manipulasi *post-processing* manusia.

   b. Validasi Strategi Augmentasi: Membuktikan secara empiris bahwa pendekatan *data augmentation* berbasis simulasi efek visual (*blurring* dan *smoothing*) dapat meningkatkan ketahanan (*robustness*) model *Deep Learning* terhadap serangan manipulasi visual sederhana.

2. Manfaat Praktis

   a. Bagi Praktisi Keamanan Siber dan Forensik: Menyediakan modul detektor (*discriminator*) yang mampu mengenali video manipulasi "hibridia" (hasil AI yang disempurnakan manusia), yang selama ini sering lolos dari deteksi algoritma standar.

   b. Bagi Masyarakat dan Media: Memberikan landasan teknis bagi pengembangan alat verifikasi konten yang lebih handal, guna meminimalisir penyebaran disinformasi atau berita bohong (*hoax*) berbasis video berkualitas tinggi di media sosial.

   c. Bagi Pengembangan Sistem Deteksi Terintegrasi: Sebagai modul validasi visual yang dapat diintegrasikan dengan sistem deteksi lain (seperti deteksi audio atau temporal) untuk menciptakan ekosistem keamanan konten digital yang komprehensif.

1.6 Sistematika Penulisan

Untuk memberikan gambaran yang jelas dan terstruktur mengenai penyusunan skripsi ini, penulis menguraikan sistematika penulisan ke dalam lima bab utama sebagai berikut:

BAB I PENDAHULUAN, Bab ini berisi gambaran umum mengenai dasar penelitian yang meliputi latar belakang masalah mengenai ancaman

deepfake dan tantangan manipulasi *post-processing*, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan yang digunakan dalam penyusunan laporan skripsi ini.

BAB II TINJAUAN PUSTAKA, Bab ini memuat uraian tentang penelitian terdahulu yang relevan sebagai bahan perbandingan, serta landasan teori yang menjadi kerangka acuan dalam pemecahan masalah. Teori-teori yang dibahas meliputi konsep *Deep Learning, Convolutional Neural Network* (CNN), arsitektur *XceptionNet, Transfer Learning, FaceForensics*, serta teori mengenai teknik manipulasi citra digital (*smoothing* dan *blurring*).

BAB III METODE PENELITIAN, Bab ini menjelaskan tahapan dan metodologi yang dilakukan penulis dalam melaksanakan penelitian. Di dalamnya terdapat tinjauan umum tentang dataset yang digunakan, skenario simulasi *post-processing* (augmentasi data), alur pra-pemrosesan data (*preprocessing*), serta perancangan arsitektur model *Forensic Discriminatos* berbasis XceptionNet yang akan dibangun.

BAB IV HASIL DAN PEMBAHASAN, Bab ini merupakan inti dari penelitian yang memuat proses implementasi sistem menggunakan bahasa pemrograman Python, proses pelatihan (*training*) model, serta pengujian sistem. Bab ini juga menyajikan analisis mendalam mengenai hasil evaluasi kinerja model berdasarkan metrik akurasi, *loss*, presisi, dan *recall*, khusunya dalam mendeteksi video yang telah mengalami manipulasi *visual*.

BAB V PENUTUP, Bab ini berisi kesimpulan akhir yang diperoleh dari hasil penelitian dan analisis yang telah dilakukan, serta saran-saran konstruktif untuk pengembangan penelitian selanjutnya agar sistem deteksi *deepfake* dapat menjadi lebih optimal di masa mendatang.