BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan pesat teknologi kecerdasan buatan telah melahirkan berbagai inovasi dalam bidang pengolahan citra digital, salah satunya adalah teknologi *deepfake*. Istilah *deepfake*, yang merupakan lakuran dari frasa "*deep learning*" dan "*fake*", merujuk pada teknik manipulasi media yang menggunakan algoritma pembelajaran mendalam untuk menghasilkan konten audio dan visual sintetis yang tampak otentik [1]. Teknologi ini umumnya memanfaatkan arsitektur *Generative Adversarial Networks* (GANs) untuk menciptakan representasi wajah yang sangat realistis dengan menggantikan identitas seseorang dalam sebuah video atau gambar [2].

Kemunculan teknologi *deepfake* pertama kali dipopulerkan oleh pengguna forum daring Reddit pada akhir tahun 2017. Pengguna tersebut mengaplikasikan metode *deep learning* untuk memanipulasi wajah dalam konten video [3]. Sejak saat itu, aksesibilitas terhadap teknologi ini terus meningkat seiring dengan berkembangnya berbagai aplikasi tingkat konsumen seperti FaceApp dan FaceSwap, yang memungkinkan pengguna awam untuk membuat konten *deepfake* dengan mudah.

Meskipun teknologi *deepfake* memiliki potensi aplikasi positif dalam industri hiburan, pendidikan, dan simulasi medis, dampak negatifnya terhadap masyarakat telah menjadi sorotan utama. Penggunaan teknologi ini dengan niat jahat dapat menghasilkan konten yang menyesatkan, menyebarkan misinformasi, dan mengancam integritas digital [4]. Salah satu kasus viral yang menunjukkan potensi destabilisasi sosial dan politik adalah video *deepfake* Presiden Ukraina, Volodymyr Zelenskyy, yang seolah-olah menyerah kepada Rusia pada tahun 2022 [5].

Mengingat ancaman serius yang ditimbulkan oleh *deepfake*, pengembangan sistem deteksi yang akurat dan andal menjadi sebuah kebutuhan yang mendesak. Pendekatan deteksi *deepfake* secara umum dapat dikategorikan menjadi tiga jenis: *naive detectors* yang menggunakan arsitektur CNN sederhana, *spatial detectors* yang mengeksplorasi artefak spasial, dan *frequency detectors* yang menganalisis domain frekuensi [6]. Namun, peningkatan kualitas *deepfake* yang dihasilkan oleh

teknik-teknik generasi terbaru menuntut pengembangan metode deteksi yang lebih canggih.

Arsitektur *deep learning* telah menunjukkan keunggulan signifikan dalam tugas deteksi *deepfake* jika dibandingkan dengan metode konvensional yang berbasis pada fitur rekayasa tangan. Keunggulan ini terletak pada kemampuan model untuk mempelajari representasi fitur secara otomatis langsung dari data, sehingga memungkinkan adaptasi yang lebih baik terhadap berbagai teknik manipulasi [7]. Berbagai arsitektur CNN termutakhir seperti ResNet [8], Xception [9], dan EfficientNet [10] telah diaplikasikan untuk deteksi *deepfake* dengan hasil yang menjanjikan.

1.1.1 Tantangan dan Keterbatasan Model Tunggal

Meskipun model tunggal menunjukkan performa yang baik, penelitian menunjukkan bahwa model tersebut sering kali memiliki keterbatasan dalam melakukan generalisasi terhadap teknik-teknik manipulasi baru yang terus berkembang. Variasi dalam metode pembuatan *deepfake*, kualitas set data, dan kondisi dunia nyata menuntut model deteksi untuk memiliki kemampuan generalisasi yang tinggi [3].

Penelitian terbaru menunjukkan bahwa berbagai metode deteksi *deepfake* memiliki performa yang bervariasi pada dataset yang berbeda. Metode POI-DeepFake yang menggunakan ResNet50 menunjukkan konsistensi performa yang baik di berbagai dataset dengan akurasi berkisar 81,10%-86,80% [11]. Sementara itu, pendekatan *hybrid* seperti DCPT yang menggabungkan CNN dengan *Vision Transformer* menunjukkan performa sangat baik pada dataset tertentu (92,11% pada FF++) namun kurang konsisten pada dataset lainnya (63,27% pada CelebDF) [12].

Penggunaan arsitektur *Xception* dalam beberapa penelitian seperti Si-Net, ISTVT, dan FAAF menunjukkan performa yang konsisten dan kompetitif, dengan akurasi berkisar 94,54%-99,85% tergantung pada dataset evaluasi [13, 14, 15]. Model *EfficientNet* juga menunjukkan keunggulan dalam hal efisiensi komputasi dengan performa yang kompetitif [16]. Namun, variabilitas performa ini mengindikasikan bahwa tidak ada satu arsitektur yang dominan untuk semua skenario, dan setiap model memiliki kelebihan dan kekurangan yang berbeda.

1.1.2 Potensi Ensemble Learning

Oleh karena itu, pendekatan *ensemble learning* yang menggabungkan beberapa model dengan karakteristik yang saling melengkapi menjadi strategi yang menjanjikan untuk meningkatkan akurasi dan keandalan sistem deteksi. *Ensemble learning* adalah sebuah teknik yang menggabungkan prediksi dari beberapa model dasar untuk menghasilkan keputusan akhir yang lebih akurat dan stabil dibandingkan model individual [17]. Dalam konteks deteksi *deepfake*, metode ansambel dapat menutupi kelemahan model-model individual dan meningkatkan kemampuan deteksi terhadap berbagai jenis manipulasi.

Penelitian terbaru menunjukkan bahwa ansambel dengan metode *weighted averaging* memberikan kinerja yang unggul dalam deteksi *deepfake* dibandingkan teknik ansambel konvensional. Metode ini menghitung kontribusi setiap model dasar berdasarkan kinerjanya pada set data validasi, sehingga model dengan akurasi lebih tinggi akan mendapatkan bobot yang lebih besar dalam pengambilan keputusan akhir. Pendekatan ini tidak hanya meningkatkan akurasi, tetapi juga memberikan interpretabilitas yang lebih baik mengenai kontribusi setiap model.

1.1.3 Pemilihan Arsitektur untuk Model Ensemble

Berdasarkan analisis literatur, kombinasi model-model dengan karakteristik arsitektur yang beragam dapat memberikan komplementaritas yang optimal untuk sebuah ansambel. Pemilihan arsitektur-arsitektur dalam penelitian ini didasarkan pada rekam jejak dan karakteristiknya yang saling melengkapi:

Custom CNN dirancang khusus untuk tugas deteksi dengan arsitektur hierarkis yang memanfaatkan karakteristik CNN dalam ekstraksi fitur lokal-ke-global. Model ini dapat mendeteksi inkonsistensi yang umum pada citra *deepfake* melalui pembelajaran fitur dari level rendah hingga level tinggi.

ResNet50 dengan mekanisme *residual learning* telah terbukti andal dalam berbagai literatur ilmiah dan kompetisi pengolahan citra [8]. Arsitektur ini menggunakan *shortcut connections* yang memungkinkan pelatihan jaringan yang sangat dalam tanpa mengalami masalah *vanishing gradient*.

Xception dengan *depthwise separable convolutions* menawarkan pendekatan ekstraksi fitur yang efisien dengan memisahkan operasi konvolusi spatial dan *channel-wise* [18]. Pendekatan ini mengurangi kompleksitas komputasi sambil mempertahankan kemampuan representasi fitur yang kuat.

EfficientNet dengan *compound scaling* menyeimbangkan kedalaman, lebar, dan resolusi input secara simultan, menghasilkan model yang efisien tanpa mengorbankan akurasi [10]. Strategi ini mengoptimalkan performa sistem secara menyeluruh.

Keragaman pendekatan ekstraksi fitur yang mereka tawarkan menjadi kunci utama dalam metode ansambel, di mana model-model dengan pola kesalahan yang berbeda cenderung dapat saling mengoreksi, sehingga berpotensi meningkatkan akurasi dan keandalan sistem deteksi secara keseluruhan.

Pemilihan set data yang tepat juga merupakan faktor krusial dalam pengembangan sistem deteksi *deepfake*. Set data ”140k Real and Fake Faces” dipilih dalam penelitian ini karena menyediakan keseimbangan yang baik antara data asli dan sintetis, dengan standardisasi format yang konsisten serta keragaman yang memadai untuk proses pelatihan dan evaluasi yang andal.

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan dan mengevaluasi sistem deteksi *deepfake* menggunakan metode *ensemble weighted averaging* yang menggabungkan empat arsitektur *deep learning* yang berbeda: Custom CNN, ResNet50, Xception, dan EfficientNet. Pendekatan ini diharapkan dapat memberikan kontribusi signifikan dalam peningkatan akurasi dan keandalan deteksi *deepfake*, serta memberikan wawasan mengenai efektivitas metode ansambel dalam domain *computer vision security*.

1.2 Rumusan Masalah

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

1. Bagaimanakah perbandingan kinerja metode *ensemble learning* dengan teknik *weighted averaging* dibandingkan dengan kinerja model-model individual dalam mengklasifikasikan citra *deepfake*?
2. Apakah metode ansambel menunjukkan kemampuan generalisasi pada skenario pengujian *cross-dataset* yang secara signifikan lebih unggul dibandingkan dengan kemampuan generalisasi masing-masing model tunggal penyusunnya?

1.3 Batasan Masalah

Untuk menjaga agar penelitian ini tetap fokus dan terarah, ditetapkan beberapa batasan masalah sebagai berikut:

1.3.1 Batasan Set Data dan Pra-pemrosesan

1. Penelitian ini hanya menggunakan set data ”140k Real and Fake Faces” yang terdiri dari 140.000 citra wajah berformat JPEG dengan resolusi 256×256 piksel.
2. Set data dibagi menjadi tiga bagian dengan rasio 100.000 untuk pelatihan, 20000 untuk validasi, dan 20000 untuk pengujian.
3. Proses *preprocessing* data terbatas pada normalisasi nilai piksel (penyekalaan ulang ke rentang 0–1) dan augmentasi data berupa pembalikan horizontal pada data pelatihan.
4. Penelitian ini berfokus pada deteksi *deepfake* berbasis citra statis dan tidak mencakup analisis pada sekuens video.

1.3.2 Batasan Arsitektur Model

1. Model ansambel terdiri dari empat arsitektur: Custom CNN, ResNet50, Xception, dan EfficientNet-B4.
2. Model ResNet50, Xception, dan EfficientNet memanfaatkan mekanisme *transfer learning* dengan menggunakan bobot pra-terlatih dari set data ImageNet.
3. Arsitektur Custom CNN dirancang dengan 4 blok konvolusional yang diikuti oleh lapisan terhubung penuh.
4. Metode ansambel yang digunakan terbatas pada *weighted averaging* yang bobotnya ditentukan berdasarkan akurasi validasi.

1.3.3 Batasan Evaluasi

1. Metrik yang digunakan untuk evaluasi kinerja meliputi Akurasi, Presisi, Perolehan (*Recall*), dan Skor-F1

2. Evaluasi kinerja model dilakukan melalui dua skenario pengujian utama:

   - **Pengujian internal:** Menggunakan test set yang berasal dari partisi dataset utama ”140k Real and Fake Faces” untuk mengukur performa model pada data dengan distribusi serupa.
   - **Pengujian Generalisasi (Cross-Dataset):** Menggunakan dataset eksternal ”DeepFakeFace” untuk menguji kemampuan generalisasi dan robustisitas model terhadap data deepfake yang dibuat dengan teknik berbeda dan tidak pernah dilihat sebelumnya.

1.3.4 Batasan Teknis

1. Proses pelatihan model dilakukan menggunakan platform Google Colab Pro+ dengan GPU Nvidia T4 dan RAM 16 GB.
2. Implementasi model menggunakan kerangka kerja TensorFlow/Keras dengan bahasa pemrograman Python 3.8+.
3. Jumlah maksimum epoch pelatihan adalah 15, dengan mekanisme penghentian dini pada patience 3.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, penelitian ini memiliki tujuan umum dan khusus sebagai berikut:

1. **Tujuan Umum:**
   Mengembangkan dan mengevaluasi sebuah sistem deteksi *deepfake* berbasis *ensemble learning* untuk menghasilkan metode klasifikasi yang tidak hanya akurat tetapi juga memiliki kemampuan generalisasi yang andal.
2. **Tujuan Khusus:**
   Secara spesifik, penelitian ini bertujuan untuk:
   (a) Membandingkan kinerja metode *ensemble learning* dengan teknik *weighted averaging* terhadap kinerja masing-masing model individual (Custom CNN, ResNet50, Xception, dan EfficientNet) dalam mengklasifikasikan citra *deepfake*.

(b) Menguji apakah pendekatan ansambel menunjukkan kemampuan generalisasi yang secara signifikan lebih unggul dibandingkan dengan setiap model individual ketika dihadapkan pada skenario pengujian *cross-dataset*.

3. **Indikator Evaluasi dan Metode Pengukuran**
   Pencapaian tujuan penelitian ini akan dievaluasi menggunakan indikator-indikator berikut:

   (a) **Peningkatan Akurasi**: Diukur melalui perbandingan nilai akurasi ensemble dengan model individual terbaik.

   (b) **Konsistensi Metrik**: Evaluasi menggunakan empat metrik utama (Accuracy, Precision, Recall, F1-Score) pada dataset pengujian.

   (c) **Kemampuan Generalisasi**: Diukur melalui pengujian cross-dataset menggunakan dataset eksternal (DeepFakeFace) dengan membandingkan penurunan performa ensemble vs model individual.

   (d) **Komplementaritas Model**: Dianalisis melalui confusion matrix dan distribusi bobot ensemble untuk memvalidasi kontribusi setiap model.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat pada berbagai aspek sebagai berikut:

1.5.1 Manfaat Teoritis

1. **Pengembangan Metode Ensemble**: Memberikan kontribusi pada pengembangan teknik *weighted averaging* untuk meningkatkan akurasi sistem deteksi *deepfake* secara signifikan.

2. **Validasi Komplementaritas Model**: Membuktikan secara empiris bahwa kombinasi arsitektur CNN yang beragam dapat meningkatkan robustisitas deteksi dengan mengurangi *false negative* dan *false positive*.

3. **Framework Evaluasi**: Menyediakan kerangka evaluasi komprehensif untuk sistem deteksi *deepfake* yang mencakup pengujian *cross-dataset* untuk mengukur kemampuan generalisasi.

1.5.2 Manfaat Praktis

1. **Peningkatan Akurasi Deteksi**: Sistem ensemble yang dihasilkan mencapai akurasi 99,64%, meningkatkan kemampuan deteksi *deepfake* untuk implementasi pada platform media sosial dan sistem verifikasi berita.
2. **Optimalisasi Computational Trade-off**: Memberikan keseimbangan antara akurasi tinggi dan efisiensi komputasi melalui kombinasi model yang ter-optimasi.
3. **Aplikabilitas Industri**: Menyediakan solusi praktis yang dapat diintegrasikan dalam sistem moderasi konten *real-time* dan analisis forensik digital.

1.5.3 Manfaat Akademis

1. **Referensi dan Tolok Ukur (*Benchmark*)**: Menjadi referensi dan menyediakan tolok ukur kinerja (*benchmark*) untuk implementasi metode *ensemble weighted averaging* pada tugas deteksi *deepfake*, yang dapat dimanfaatkan oleh komunitas akademik untuk penelitian selanjutnya.
2. **Dasar Pengembangan Lanjutan**: Menyediakan dasar dan wawasan untuk pengembangan metode deteksi *deepfake* yang lebih maju, khususnya dalam eksplorasi arsitektur yang komplementer dan teknik ansambel yang lebih canggih.

1.5.4 Manfaat Sosial

1. **Literasi Media**: Berkontribusi pada upaya peningkatan kemampuan masyarakat untuk mengidentifikasi konten manipulatif di era digital.
2. **Integritas Informasi**: Mendukung terjaganya kebenaran dan kepercayaan dalam ekosistem informasi digital melalui teknologi deteksi yang canggih.
3. **Pengembangan AI yang Etis**: Memberikan contoh penggunaan kecerdasan buatan untuk tujuan defensif dan protektif, menyeimbangkan kemajuan teknologi generatif dengan kapabilitas untuk memitigasi risikonya.

1.6 Sistematika Penulisan

Laporan penelitian ini disusun secara sistematis dan logis untuk memberikan pemahaman yang komprehensif mengenai penelitian yang dilakukan. Struktur penulisan terdiri dari lima bab utama dengan rincian sebagai berikut:

- **Bab I - PENDAHULUAN** Bab ini menguraikan latar belakang yang menjelaskan urgensi pengembangan sistem deteksi *deepfake*, rumusan masalah, batasan-batasan penelitian, tujuan yang ingin dicapai, manfaat teoretis dan praktis, serta sistematika penulisan laporan.
- **Bab II - TINJAUAN PUSTAKA** Bab ini menyajikan tinjauan pustaka dan landasan teori yang relevan, mencakup konsep fundamental kecerdasan buatan dan *deep learning*, teknologi *deepfake* dan *Generative Adversarial Networks*, arsitektur *deep learning* yang digunakan (CNN, ResNet, Xception, EfficientNet), teori *ensemble learning* dan *weighted averaging*, serta metrik evaluasi yang komprehensif.
- **Bab III - METODOLOGI PENELITIAN** Bab ini menjelaskan metodologi penelitian secara terperinci, meliputi desain penelitian dan alur kerja eksperimen, karakteristik dan proses pra-pemrosesan set data, arsitektur dan konfigurasi setiap model individual, implementasi metode *ensemble weighted averaging*, prosedur pelatihan dan penalaan hiperparameter, serta lingkungan komputasi dan reprodusibilitas.
- **Bab IV - HASIL DAN PEMBAHASAN** Bab ini menyajikan hasil eksperimen beserta pembahasan yang komprehensif. Cakupannya meliputi analisis kinerja setiap model *deep learning* secara individual, hasil implementasi sistem ansambel, perbandingan kuantitatif antara pendekatan individual dan ansambel, analisis kontribusi dan komplementaritas antar model, interpretasi hasil dalam konteks metode termutakhir, serta pembahasan mengenai limitasi dan implikasi dari temuan penelitian.
- **Bab V - KESIMPULAN DAN SARAN** Bab ini berisi kesimpulan penelitian yang menjawab rumusan masalah berdasarkan hasil eksperimen, menguraikan kontribusi ilmiah yang dihasilkan, mengidentifikasi keterbatasan penelitian, serta memberikan saran untuk pengembangan penelitian selanjutnya dan implementasi praktis dalam aplikasi dunia nyata.