BAB 1. PENDAHULUAN

1.1 Latar Belakang

Citra sintesis (Citra *generative AI*) atau seringkali juga disebut *deepfake* adalah sebuah citra yang merupakan representasi visual bukan dari kamera nyata melainkan citra yang dihasilkan menggunakan teknik *deep learning* berupa model *generative AI*, dimana model dilatih untuk meniru citra asli berupa citra yang berasal dari kamera nyata untuk membuat citra baru yang serupa dengan citra asli (Angeline & Kusniyati, 2024). Proses pembuatan citra sintesis ini dilakukan menggunakan model generatif berupa proses mengubah input teks, sketsa, atau sumber lain seperti citra kedalam citra yang baru. Ini menjadi isu penting yang telah menarik perhatian komunitas penelitian untuk mengatasi tantangan tingkat tinggi dalam menghasilkan citra yang fotorealistik (S. Baraheem dkk., 2023). Tak hanya itu, kemunculan citra sintesis juga menjadi tantangan baru pada teknologi biometrik, salah satunya sistem pengenalan wajah (*face recognition*). Oleh karena itu, sebuah sistem deteksi wajah palsu (*fake face detection*) menjadi komponen penting yang harus diintegrasikan ke dalam sistem deteksi keamanan era saat ini. Teknik pengenalan wajah (*face recognition*) telah diteliti sejak tahun 1990-an dengan berbagai metode tradisional berupa analisis komponen wajah, hingga metode untuk mendeteksi fitur tekstur lokal pada sebuah citra wajah seperti *Local Binary Pattern* (LBP) yang merupakan teknik non parametik yang mendeskripsikan fitur tekstur lokal pada sebuah citra. Proses deteksi untuk pengenalan wajah (face recognition) berupaya untuk mengabaikan pengaruh seperti pencahayaan, bayangan, serta postur untuk identifikasi wajah seseorang. Sebaliknya deteksi wajah citra sintesis justru berfokus pada pengaruh tersebut untuk mendeteksi citra tersebut sintesis atau asli. (Favorskaya & Yakimchuk, 2021).

Tantangan teknis untuk mengembangkan sistem deteksi *deepfake* menjadi semakin krusial di era digital yang telah berkembang ini. Dengan kemunculan

media sosial yang menjadi platform penting untuk berkomunikasi dan berinteraksi di era digital seiring dengan popularitasnya yang meningkat menyebabkan kekhawatiran tentang potensi penyalahgunaan AI di media sosial. Salah satu masalah utama adalah kurangnya kerangka kerja yang jelas untuk memastikan bahwa AI di media sosial dapat dipercaya karena dapat menyebabkan berbagai akibat negatif, seperti penyebaran informasi yang salah, diskriminasi, dan pelanggaran privasi (Lewis & Moorkens, 2020). Menurut data *Statista Market Insight* pada tahun 2022, penggunaan alat *generative AI* di Indonesia meningkat, menunjukkan bahwa citra sintesis atau *deepfake* dapat tersebar dan semakin populer di media sosial. Hal ini memperkuat kekhawatiran bahwa citra sintesis dapat disalahgunakan untuk berbagai tujuan tindak kriminal. Misalnya, citra sintesis bisa digunakan untuk menyebarkan disinformasi dengan membuat konten palsu atau menyesatkan yang dapat memanipulasi opini publik dan menimbulkan perselisihan di media sosial. Selain itu, citra dapat digunakan dalam penipuan dan pencurian identitas dengan membuat profil palsu atau meniru identitas orang lain untuk tujuan penipuan atau pencurian informasi pribadi (Blauth dkk., 2022).

Penggunaan model generatif seperti *Generative Adversarial Networks* (GANs) adalah contoh lain dari penggunaan AI dalam menggunakan sintesis citra. Studi yang dilakukan oleh Yu, Davis, dan Fritz (2019) menemukan bahwa *Generative Adversarial Networks* (GANs) dapat digunakan untuk membuat citra palsu yang sangat percaya diri yang sulit dibedakan dari citra asli. Dalam artikel jurnal berjudul "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints" mereka menjelaskan bagaimana *Generative Adversarial Networks* (GANs) dapat digunakan untuk membuat identitas palsu, menyebarkan informasi yang salah tentang citra, dan bahkan melakukan penipuan visual yang kompleks. Studi mereka menunjukkan bahwa mempelajari dan menganalisis sidik jari unik yang dibuat oleh *Generative Adversarial Networks* (GANs) sangat penting untuk mengembangkan teknik deteksi yang lebih baik (Yu dkk., 2019).

Untuk mengatasi masalah ini, berbagai penelitian yang telah dilakukan untuk mengembangkan teknik yang dapat diandalkan untuk mendeteksi citra hasil AI. Misalnya, pada penelitian oleh Wang dkk, yang mengembangkan LBP-Net dan menghasilkan akurasi cukup tinggi dengan didukung oleh dataset yang cukup besar sehingga menghasilkan akurasi pada model LBP-Net mencapai akurasi cukup tinggi dengan berbagai teknik augmentasi tertentu (Wang dkk., 2021). Tak hanya itu, penelitian oleh Angelina dengan mengembangkan serta membandingkan berbagai model transfer learning CNN seperti VGG-Net sebagai salah satu model pre-trained yang populer dalam bidang *computer vision* mendapatkan hasil yang cukup tinggi juga meskipun dengan data yang terbatas pada berbagi model yang telah digunakan (Angeline & Kusniyati, 2024). Namun, belum ada penelitian yang secara langsung membandingkan pengaruh pra-pemrosesan *Local Binary Pattern* (LBP) pada arsitektur *transfer learning* salah satunya VGG-Net, sehingga diperlukan analisis mendalam terkait pengaruh LBP pada performa model CNN.

Berdasarkan penelitan sebelumnya, terbukti kuat bahwa *deep learning* khususnya dapat menjadi solusi yang efektif untuk melakukan pembelajaran yang lebih akurat dalam mendeteksi citra hasil *AI generated*. *Deep learning* merupakan bidang studi terkini dalam pembelajaran mesin yang memanfaatkan lapisan tersembunyi dari jaringan neural buatan dan abstraksi model tingkat tinggi pada basis data yang sangat besar (Shinde & Kalpana, 2023).

Menjawab tantangan tersebut, penelitian ini mengusulkan pengembangan dan evaluasi sebuah sistem deteksi citra wajah hasil AI menggunakan arsitektur deep learning VGG-16 dengan menganalisis performa model secara mendalam, termasuk pengaruh fitur tekstur *Local Binary Pattern* (LBP). Hal ini bertujuan untuk menghasilkan deteksi citra wajah citra hasil AI di media sosial yang lebih akurat dan dapat digeneralisasi.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan diatas, maka digunakan analisis tekstur wajah menggunakan *Local Binary Pattern* (LBP). Oleh karena itu, rumusan masalah dalam penelitian ini adalah:

a. Bagaimana pengaruh *Local Binary Pattern* (LBP) sebagai analisis tekstur wajah saat *preprocessing* terhadap performa dan akurasi model ?
b. Bagaimana tingkat akurasi performa model dengan pendekatan *deep learning VGG-16* dalam mendeteksi berbagai citra wajah hasil *AI generated* ?
c. Bagaimana mengevaluasi hasil dengan pendekatan *deep learning VGG-16* dalam mendeteksi berbagai citra wajah hasil *AI generated* ?
d. Bagaimana model *deep learning* yang telah dilatih dapat diimplementasikan menjadi sebuah sistem aplikasi yang fungsional unntuk mendeteksi citra wajah sintesis pada platform media sosial dengan kualitas yang tinggi?

1.3 Tujuan

Berdasarkan rumusan masalah diatas, dapat disimpulkan tujuan dari penelitan ini sebagai berikut:

a. Menganalisis dan menyimpulkan pengaruh *Local Binary Pattern* (LBP) sebagai analisis tekstur wajah saat *preprocessing* terhadap performa dan akurasi model.
b. Menganalisis tingkat akurasi performa model dengan pendekatan *deep learning VGG-16* dalam mendeteksi berbagai citra wajah hasil *AI generated.*
c. Mengevaluasi hasil dengan pendekatan *deep learning VGG-16* dalam mendeteksi berbagai citra wajah hasil *AI generated*.
d. Mengimplementasikan model *deep learning* yang telah dilatih dapat diimplementasikan menjadi sebuah sistem aplikasi yang fungsional unntuk mendeteksi citra wajah sintesis pada platform media sosial dengan kualitas yang tinggi.

1.4 Manfaat

Hasil penelitian ini diharapkan mampu memberikan manfaat sebagai berikut:

a. Membantu masyarakat umum untuk mengenali antara citra palsu berupa *AI generated* dan citra asli pada konten media sosial. Sehingga luaran dari sistem ini dapat memberikan kontribusi terhadap bidang keamanan pada

konten media sosial dengan mengurangi penyebaran informasi yang salah.

b. Penelitian ini juga diharapkan dapat menjadi alat pendukung upaya pencegahan penipuan visual dan berbagai kejahatan AI lainnya yang dapat merugikan masyarakat secara luas di era internet saat ini dengan meningkatkan kemampuan untuk mengidentifikasi dan mencegah penyalahgunaan identitas palsu pada konten medisa sosial.

1.5 Batasan Masalah

Agar penelitian ini lebih terfokus dan terarah, maka ditetapkan beberapa batasan masalah sebagai berikut :

a. Dataset citra sintesis atau *deepfake* yang digunakan dalam penelitian ini berasal dari satu sumber, yaitu model generatif *Meta AI* pada *WhatsApp*. Penelitian ini tidak mencakup perbandingan deteksi berbagai model lain seperti *Midjourney*, *DALL-E*, atau *StyleGAN*.
b. Untuk menjaga konsistensi agar model dapat fokus pengenalan artefak atau ciri citra sintesis, dataset citra asli dan citra sintesis atau *deepfake* yang dikumpulkan berfokus pada satu tokoh publik, yaitu Cristiano Ronaldo.
c. Penelitian ini berfokus pada penggunaan arsitektur VGG-16 sebagai model dasar (*base model*). Perbandingan performa dengan arsitektur *transfer learning* lainnya tidak dilakukan dalam penelitian ini.
d. Sistem yang dikembangkan hanya melakukan klasifikasi biner, yaitu membedakan antara citra asli dan palsu atau sintesis. Sistem tidak dirancang untuk mengidentifikasi jenis manipulasi berupa model generative AI spesifik yang menghasilkan citra tersebut.
e. Analisis pengaruh fitur hanya terbatas pada perbandingan antara citra *grayscale* dan citra yang telah melalui transformasi *Local Binary Pattern* (LBP).
f. Model yang dihasilkan diimplementasikan menjadi sebuah aplikasi *mobile* berbasis Android dan tidak mencakup platform lain seperti *website* atau *desktop*.