BAB I
PENDAHULUAN

A. Latar Belakang

Seperti namanya, ”*deepfake*” secara etimologi adalah gabungan dari “*deep learning*” dan “palsu” yang melibatkan pengubahan video secara digital untuk menciptakan representasi hiper-realistis dari individu yang seolah-olah mengatakan dan melakukan hal-hal yang sebenarnya tidak pernah terjadi [1]. Istilah ini umumnya digunakan untuk merujuk pada manipulasi media yang sudah ada (gambar, video, dan/atau audio) atau pembangkitan media baru (sintetis) menggunakan pendekatan berbasis *deep learning*. Data *deepfake* yang paling sering dibahas meliputi gambar wajah palsu, rekaman suara palsu, dan video palsu yang menggabungkan gambar palsu dan rekaman suara palsu. Meskipun kata “palsu” dalam istilah ini menunjukkan media yang dimanipulasi atau disintesis, terdapat banyak aplikasi yang tidak berbahaya dari teknologi *deepfake*, misalnya untuk hiburan dan seni kreatif. Dengan pertimbangan ini, istilah lain “deep synthesis” telah diusulkan sebagai alternatif yang lebih netral [2]. Namun, istilah baru ini belum banyak diketahui secara luas.

Istilah *deepfake* pertama kali muncul pada tahun 2017 dan menjadi sangat populer setelah kemunculannya. Semua bermula dari salah satu pengguna aplikasi *Reddit* dengan *username* /u/deepfakes, pencipta subreddit /r/deepfakes yang kemudian diblokir, mendedikasikannya untuk video porno sintetis, dan repositori kode open-source GitHub untuk sistem *deepfake* yang banyak direplikasi dan di-fork di mana pengguna tersebut mengembangkan GAN menggunakan TensorFlow, perangkat lunak mesin pencari gratis dari Google untuk menempelkan wajah seseorang ke tubuh wanita lain dalam film pornografi [3]. Munculnya *deepfake* semacam itu kemudian kembali menjadi perbincangan publik yang hangat ketika sebuah video dari Mark Zuckerberg muncul pada Juni 2019 yang membahas kekuatan Facebook untuk mengendalikan semua data penggunanya yang ternyata adalah video yang diedit secara masif, padahal video aslinya telah diunggah

sebelumnya oleh Zuckerberg sendiri pada September 2017 ketika dia membahas campur tangan Facebook dalam pemilihan umum di Rusia [4].

Sejauh ini penggunaan teknologi *deepfake* AI generatif yang paling umum adalah untuk membuat gambar intim dan pornografi non-konsensual. Sebuah tinjauan *deepfake* online tahun 2019 menemukan bahwa 96% adalah pornografi [5]. Analisis tahun 2023 tentang *deepfake* non-konsensual menemukan bahwa setidaknya 244.625 video telah ditambahkan ke situs web teratas yang disiapkan untuk menghosting video porno *deepfake* dalam tujuh tahun sebelumnya, 113.000 di antaranya ditambahkan pada tahun 2023, menandai peningkatan 54% dibandingkan tahun sebelumnya [6]. Pada tahun 2025, Motif penyalahgunaan *deepfake* juga semakin bervariasi, dengan kategori terbesar pada konten eksplisit non-konsensual (32%), penipuan finansial (23%), manipulasi politik (14%), dan misinformasi (13%). Serangan tidak hanya menargetkan figur publik, tetapi juga masyarakat umum, termasuk perempuan, pelajar, dan anak di bawah umur. Secara global, insiden paling banyak terjadi di Amerika Utara (38%), diikuti Asia (27%) dan Eropa (21%). Bahkan 63% kasus melibatkan elemen lintas negara, sehingga menambah kompleksitas dalam proses penegakan hukum dan pencegahan. [7].

Perkembangan teknologi *deepfake* yang menjamur dapat menimbulkan ancaman epistemik, seperti menggiring opini masyarakat untuk lebih mudah mempercayai akan informasi yang mereka konsumsi yang sebenarnya tidak nyata dan tidak pernah terjadi [8]. Menjamurnya teknologi seperti *deepfake* dapat menimbulkan dampak negatif , yakni munculnya tindak pidana baru [9]. Salah satu fenomena atau kasus nyata yang muncul akibat perkembangan teknologi ini risiko *"People Defense"*, di mana terdakwa mengklaim bukti yang didakwakan kepada tersangka adalah palsu, seperti pada kasus *People v. Foreman*. Kondisi tersebut berpotensi menimbulkan krisis kredibilitas bukti digital di pengadilan. [10]. Di Korea Selatan, kasus manipulasi wajah guru perempuan ke dalam konten pornografi meningkat drastis, dengan data menunjukkan lonjakan lebih dari 670% antara 2021–2024. Survei pada 2024 mencatat lebih dari 2.492 kasus manipulasi foto ilegal di lingkungan sekolah, melibatkan ratusan guru dan siswa sebagai korban [11]. Penyalahgunaan dalam dalam bentuk politik, konflik Rusia-Ukraina yang

sedang berlangsung, video *deepfake* telah muncul sebagai alat propaganda perang. Dalam satu kasus, sebuah video palsu Presiden Ukraina Volodymyr Zelenskyy muncul secara online, yang secara salah menyatakan penyerahan Ukraina [12].

Teknologi *deepfake* juga merambat sampai ke Indonesia. Fenomena penipuan berbasis *deepfake* di Indonesia menunjukkan eskalasi signifikan dengan peningkatan kasus sebesar 1550% antara tahun 2022 hingga 2023. Lonjakan ini menegaskan bahwa teknologi *deepfake* selain membawa inovasi, juga menghadirkan ancaman serius terhadap keamanan digital [13]. Kasus *deepfake* yang terjadi di Indonesia salah satunya adalah video manipulatif untuk menawarkan bantuan pemerintah palsu menggunakan wajah Presiden Prabowo Subianto dengan syarat transfer dana melalui WhatsAp sekitar 100 orang tersebar di 20 provinsi, Kerugian diperkirakan sekitar Rp 95 juta rupiah dalam kurun waktu empat bulan sebelum tertangkap [14]. Penyimpangan lainnya juga terjadi di Universitas Udayana (Unud), Bali, seorang mahasiswa berinisial SLKDP diduga manipulasi data pribadi mahasiswi menjadi konten vulgar menggunakan *deepfake* di Telegram, dengan mengambil tangkapan layar dari akun Instagram korban tanpa izin, selain melanggar privasi ini juga mempengaruhi mentalitas korban, menghancurkan hubungan mereka, dan menodai martabat korban [15]. Seperti halnya video *deepfake* lainnya menjelang pemilu Presiden 2029, diperkirakan akan terancam *deepfake* dan isu SARA [16]. Untuk mengantisipasi tantangan tersebut, penyelenggara pemilu memperkuat kolaborasi dengan Kementerian Komunikasi dan Digital (Komdigi) serta Badan Siber dan Sandi Negara (BSSN) [17].

Dengan besarnya dampak negatif yang dihasilkan teknologi *Deepfake*, berbagai penelitian telah diilakukan untuk mendeteksi *deepfake*, mulai dari analisis tekstur wajah, pola pergerakan mata dan bibir, hingga deteksi artefak visual yang tidak alami [18]. Deteksi video *deepfake* dapat diklasifikasikan menjadi tiga kategori berdasarkan pendekatannya: Metode *Feature-Based (Handcrafted Features), Deep learning-Based Model,* dan *Hybrid Methods (Hybrid Feature-Based* dan *Deep learning Approaches)* [19]. Penelitian untuk mendeteksi *deepfake* menggunakan *Convolutional Neural Network* (CNN) yang dioptimalkan dengan arsitektur *EfficientNet-B6*. Hasil penelitian menunjukkan bahwa model CNN

EfficientNet mampu mencapai akurasi sekitar 90%, dengan nilai precision, recall, dan F1-score secara real-time melalui platform seperti Flask [20]. Penelitian dengan metode serupa yang sama berfokus pada pengembangan sistem deteksi *deepfake* berbasis CNN. Penelitian yang dilakukan pada dataset publik Celeb-DF. Menunjukkan bahwa model mencapai performa tertinggi, dengan akurasi 93,2% dan AUC 0,93. Model CNN menawarkan keseimbangan yang praktis antara akurasi dan biaya komputasi, menjadikannya cocok untuk digunakan di dunia nyata, terutama pada perangkat dengan sumber daya terbatas [21]. Penelitian yang dilakukan oleh Anlin mengusulkan sistem deteksi *deepfake* berbasis *transfer learning* dengan memanfaatkan FaceNet512 untuk menghasilkan sampel wajah, kemudian diklasifikasikan menggunakan MobileNetV2. Hasil pengujian menunjukkan akurasi rata-rata mencapai 82,4% dengan nilai AUC sebesar 0,82. Pendekatan ini terbukti lebih efisien dibandingkan metode analisis *full-frame video* karena hanya berfokus pada fitur wajah, sehingga lebih cepat, ringan, dan tetap akurat [22].

Berdasarkan uraian pada latar belakang tersebut, penelitian ini mengusulkan sebuah sistem ”DEFEND” atau ”*Deepfake Efficient Detector*”. Penelitian ini bertujuan untuk mengukur kemampuan dan tingkat akurasi metode *transfer learning* pada arsitektur *EfficientNet* dalam mengklasifikasi dan mengenali video asli serta video *deepfake* yang banyak beredar di media sosial. Dataset *FaceForensics++* digunakan sebagai data latih dan data uji untuk memastikan model mampu mendeteksi manipulasi artefak visual secara akurat. Penelitian ini hadir dengan mengimplementasikan model *EfficientNet-B0* yang dirancang melalui teknik *compound scaling* untuk menyeimbangkan kedalaman, lebar, dan resolusi jaringan secara efisien. Penggunaan *EfficientNet* diharapkan mampu memberikan akurasi yang kompetitif namun dengan beban komputasi (FLOPs) yang jauh lebih rendah dibandingkan model konvensional. Hasil penelitian diharapkan dapat berkontribusi pada peningkatan akurasi deteksi *deepfake* serta membantu upaya meminimalisir mitigasi penyebaran konten manipulatif yang dapat menimbulkan dampak negatif yang di masyarakat.

B. Perumusan Masalah

Berdasarkan latar belakang penelitian yang telah diuraikan di atas, maka dapat disimpulkan bahwa rumusan masalah penelitian ini adalah untuk mengetahui tingkat akurasi yang dihasilkan metode *transfer learning* model *EffientNet* dalam mengklasifikasikan keaslian wajah pada video.

C. Batasan Penelitian

Dalam penelitian ini, penulis telah menetapkan beberapa batasan terkait penelitian ini sebagai berikut :

1. Penellitian ini berfokus pada metode *transfer learning* dengan menggunakan arsitektur *EfficientNet-B0* dilatih *(pre-trained)* pada dataset *ImageNet*.
2. Analisis deteksi difokuskan hanya pada area wajah (*Face Region)* dan diekstraksi menjadi *frames*/video. Penelitian ini tidak menganalisis komponen audio (suara), gestur, atau bagian tubuh lainnya untuk menentukan keaslian video. Dengan format input berupa video ".mp4". Durasi video rata-rata 10-50 detik, menggunakan berbagai variasi *frame fate* (seperti 24, 25, 30, 60, hingga 120 fps). Proses pengambilan sampel *frames* dilakukan sebanyak 20 *frames* per video sebagai representasi visual dari seluruh durasi video.
3. Dataset yang digunakan pada penelitian ini berasal dari *FaceForensics++*, https://niessnerlab.org/projects/roessler2018faceforensics.html, dengan rasio 224x224 piksel, yang dibagi dibagi menjadi data latih (70%), data validasi (20%), dan data uji (10%). Dataset ini berjumlah 300 video (terdiri dari 150 video "Asli", "0" dan 150 video "*Deepfake*", "1"). Model hanya bisa melakukan analisis pada video, jika terdapat minimal 1 wajah dan maksimal 3 wajah pada video yang di input/diunggah.
4. Model dikembangkan menggunakan dataset FaceForensics++, namun untuk menguji keandalan sistem "DEFEND", dilakukan pengujian *Trial and Error* menggunakan video dari internet (*In the wild*). Pengujian ini mencakup perbandingan berbagai resolusi video (seperti 360p, 480p, 720p, hingga 1080p) untuk menganalisis sejauh mana kualitas resolusi memengaruhi tingkat akurasi model dalam mengidentifikasi artefak manipulasi wajah.

5. *Output* akhir dari yang diharapkan dari model ini adalah mampu memberikan klasifikasi biner ”0” atau ”1”, dalam bentuk probabilitas keyakinan ”Asli” atau ”*Deepfake*” untuk hasil akhir dari identifikasi.

D. Tujuan Penelitian

Berdasarkan latar belakang penelitian yang telah diuraikan di atas, maka dapat disimpulkan bahwa tujuan penelitian ini adalah untuk mengetahui tingkat akurasi yang dihasilkan metode *transfer learning* model *EffientNet* dalam mengklasifikasikan keaslian wajah pada video.

E. Manfaat Penelitian

Manfaat dari penelitian yang penulis lakukan adalah sebagai berikut :

1. Bagi penulis, penelitian dan penulisan karya ilmiah ini memberikan pengalaman dan pengetahuan baru yang berharga, khususnya dalam memahami dan menerapkan metode *transfer learning* Model Arsitektur *EfficientNet* dalam mengidentifikasi keaslian video. Penelitian ini menawarkan solusi praktis dalam mengklasifikasikan unsur *Deepfake* guna menghindari terjadinya kesalahpahaman dalam penyebaran informasi.
2. Bagi perguruan tinggi dan akademisi, Penelitian ini diharapkan dapat memberikan kontribusi dalam penggunaan metode *transfer learning* model *EfficientNet* sebagai Algoritma yang mengidentifikasi Keaslian Video Asli dan Video *Deepfake* dapat digunakan sebagai acuan untuk penelitian yang akan datang di masa depan serta memperbaharui dan melengkapi penelitian sebelumnya yang sudah ada sebelumnya.
3. Bagi Masyarakat secara umum, penelitian ini diharapkan bisa meningkatkan kesadaran masyarakat terhadap ancaman penyalahgunaan dari teknologi *Deepfake* dan memberikan upaya mengenali penyebaran informasi palsu yang berbasis video di media sosial. Hasil dari penelitian ini jika memungkinkan akan dipublikasikan dalam bentuk, artikel ilmiah yang dapat diakses secara terbuka dalam bentuk jurnal untuk meningkatkan literasi digital masyarakat terhadap risiko konten *Deepfake* yang disalahgunakan.