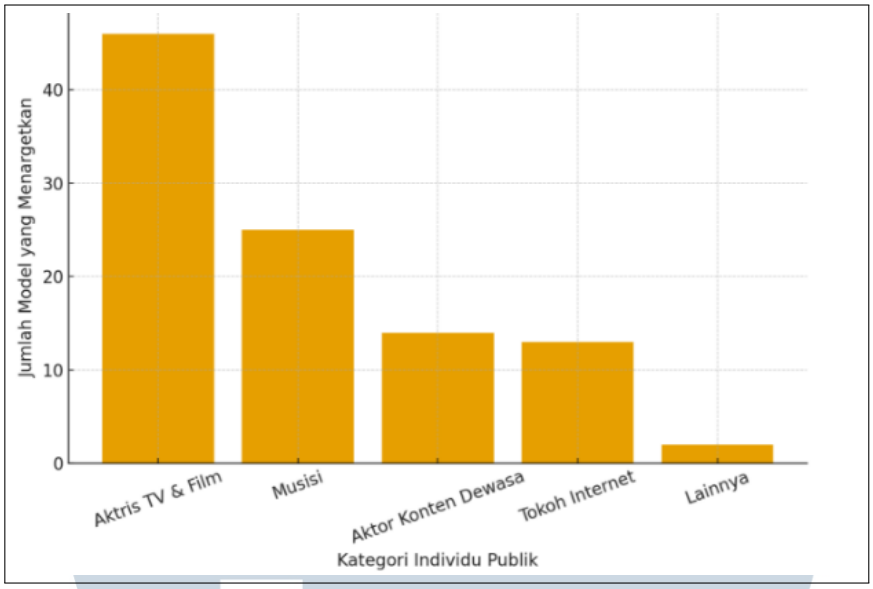
BAB 1
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi digital dan internet telah memberikan kemudahan dalam akses informasi dan komunikasi, namun sekaligus memunculkan berbagai ancaman dalam bidang cyber security, seperti serangan siber, penyebaran hoaks, black campaign, dan ujaran kebencian [1]. Tingginya volume informasi digital serta kemudahan distribusi konten melalui media sosial menyebabkan informasi palsu dapat menyebar secara luas dan sulit dikendalikan. Kondisi ini berpotensi menimbulkan kepanikan publik, memanipulasi opini masyarakat, serta menurunkan tingkat kepercayaan terhadap informasi digital.

Seiring dengan kemajuan kecerdasan buatan (artificial intelligence), muncul fenomena deepfake atau deephoax, yaitu teknologi yang memanfaatkan deep learning untuk memanipulasi gambar dan video sehingga tampak realistis dan sulit dibedakan dari konten asli. Teknologi ini meningkatkan kompleksitas ancaman digital karena konten manipulatif yang dihasilkan mampu melewati sistem deteksi tradisional dan menyesatkan publik melalui perubahan visual yang sangat meyakinkan. Fenomena penyalahgunaan deephoax tercermin dari analisis terhadap 100 model generatif deephoax paling banyak diunduh dengan label “celebrity” pada platform Civitai, yang menunjukkan bahwa sebagian besar model menargetkan figur publik terkenal, dengan rincian 46 model menargetkan perempuan di industri televisi dan film, 25 musisi, 14 aktor dalam konten dewasa, serta 13 tokoh internet seperti influencer di platform TikTok, Twitch, dan Instagram [2].



Gambar 1.1. Distribusi Target Deepfake (Deephoax) Berdasarkan Analisis 100 Model Civitai (2024)

Temuan ini menegaskan bahwa penyalahgunaan teknologi deephoax semakin meluas dan menimbulkan ancaman serius terhadap privasi individu, reputasi publik, serta stabilitas ekosistem informasi digital, sehingga pengembangan sistem deteksi otomatis menjadi kebutuhan penelitian yang sangat mendesak. Meningkatnya produksi konten visual sintetis akibat perkembangan kecerdasan buatan generatif (AI-generated) menimbulkan tantangan serius terhadap keamanan digital dan integritas informasi. Gambar deephoax berpotensi disalahgunakan untuk penipuan identitas, penyebaran hoaks, dan social engineering [3, 4]. Dalam konteks ini, pendekatan berbasis transfer learning dinilai memiliki potensi besar karena kemampuannya dalam mengekstraksi fitur visual kompleks yang sulit diidentifikasi secara manual, sehingga relevan untuk pengembangan sistem deteksi gambar deephoax yang andal.

Berbagai penelitian sebelumnya telah mengeksplorasi penggunaan Convolutional Neural Networks (CNN) untuk mendeteksi konten visual hasil manipulasi, dengan model seperti VGG, ResNet, EfficientNet, dan Xception menunjukkan performa yang menjanjikan [5, 6]. Arsitektur Xception, yang menerapkan depthwise separable convolution, terbukti efektif dalam mengekstraksi fitur visual tingkat tinggi [7, 8]. Namun, sebagian besar penelitian masih terbatas pada pelaporan akurasi umum atau berfokus pada manipulasi berbasis video [9]. Oleh karena itu, penelitian ini melakukan evaluasi kinerja model transfer learning berbasis Xception menggunakan dataset gambar wajah dari Flickr-Faces-HQ

(FFHQ) dan gambar wajah hasil AI-generated, dengan metrik evaluasi berupa akurasi, precision, recall, f1-score, dan confusion matrix untuk menilai efektivitas serta kemampuan generalisasi model dalam mendeteksi gambar deephoax.

1.2 Rumusan Masalah

Berdasarkan latar belakang dan celah penelitian yang telah dipaparkan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana penerapan metode *transfer learning* menggunakan model Xception untuk mendeteksi gambar wajah *deephoax*?
2. Seberapa baik performa model Xception berbasis *transfer learning* dalam mendeteksi gambar wajah asli dan gambar wajah *deephoax*?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Menerapkan metode *transfer learning* menggunakan arsitektur Xception untuk mendeteksi gambar *deephoax* berbasis wajah.
2. Mengevaluasi kemampuan model Xception dalam mengekstraksi fitur visual dan membedakan gambar wajah asli dan gambar wajah *deephoax* pada dataset validation dan testing.

1.4 Urgensi Penelitian

Urgensi penelitian ini didasari oleh meningkatnya ancaman keamanan digital akibat perkembangan teknologi kecerdasan buatan generatif yang memungkinkan pembuatan gambar *deephoax* dengan tingkat realisme yang tinggi dan sulit dibedakan dari gambar asli. Penyebaran gambar *deephoax* berpotensi disalahgunakan untuk penipuan identitas, manipulasi opini publik, pencemaran nama baik, serta berbagai bentuk kejahatan siber lainnya, sehingga menimbulkan risiko serius terhadap privasi, kepercayaan publik, dan integritas informasi digital. Dalam konteks keilmuan, penelitian ini penting untuk mengevaluasi efektivitas pendekatan *transfer learning* berbasis arsitektur Xception dalam mendeteksi konten visual hasil manipulasi kecerdasan buatan, sekaligus mendukung pengembangan

sistem deteksi otomatis yang andal dan relevan dengan kebutuhan keamanan informasi di era digital.

1.5 Luaran Penelitian

Luaran dari penelitian ini meliputi penerapan model deteksi gambar *deephoax* berbasis *transfer learning* menggunakan arsitektur Xception yang telah dievaluasi secara komprehensif, penyusunan laporan penelitian yang mendokumentasikan metode penelitian dan hasil eksperimen, serta potensi publikasi artikel ilmiah pada jurnal atau prosiding di bidang *computer vision* dan *deep learning*. Selain itu, hasil penelitian ini juga dapat menjadi dasar pengembangan prototipe sistem deteksi gambar *deephoax* dan referensi untuk penelitian lanjutan, serta memiliki peluang untuk dikembangkan menjadi luaran berupa Hak Kekayaan Intelektual (HKI) pada tahap selanjutnya.

1.6 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut.

1. Memberikan kontribusi ilmiah dalam bidang *computer vision* dan *deep learning*, khususnya terkait evaluasi kinerja model *transfer learning* berbasis Xception untuk deteksi gambar *deephoax*.
2. Menjadi referensi bagi penelitian selanjutnya dalam pengembangan dan perbandingan model deteksi *deephoax* dengan arsitektur deep learning lainnya.

1.7 Batasan Penelitian

Agar penelitian ini lebih terarah dan sesuai dengan tujuan yang telah ditetapkan, maka batasan penelitian yang digunakan adalah sebagai berikut.

1. Penelitian difokuskan pada deteksi gambar *deephoax* berbasis wajah (face images).
2. Model yang digunakan adalah arsitektur Xception dengan pendekatan *transfer learning* berbasis *feature extraction*.

3. Dataset yang digunakan terdiri dari 30.000 gambar wajah asli dari Flickr-Faces-HQ (FFHQ) dan 30.000 gambar wajah hasil generasi kecerdasan buatan sebagai representasi *deephoax*.
4. Penelitian ini tidak membahas deteksi *deephoax* pada media video maupun audio.
5. Pengujian model hanya dilakukan menggunakan dataset *validation* dan *testing*, tanpa pengujian tambahan pada data dunia nyata (*real-world data*).
6. Evaluasi performa model dibatasi pada penggunaan confusion matrix dan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

BAB 2
TINJAUAN PUSTAKA

2.1 Deephoax

Deepfake merupakan teknologi manipulasi konten digital berbasis kecerdasan buatan, khususnya *deep learning*, yang digunakan untuk menghasilkan media sintetis berupa gambar, video, maupun audio yang tampak sangat realistis namun sebenarnya palsu [10]. Teknologi ini memanfaatkan berbagai arsitektur jaringan saraf seperti *Generative Adversarial Networks* (GAN), *autoencoder*, serta *Convolutional Neural Networks* (CNN) untuk merekayasa wajah, ekspresi, dan atribut visual lainnya sehingga menyerupai data asli dengan tingkat kemiripan yang tinggi [11]. Kemampuan deepfake dalam menghasilkan konten visual yang meyakinkan menjadikannya semakin sulit dibedakan dari konten asli oleh persepsi manusia.

Istilah *deephoax* dalam penelitian ini digunakan untuk menggambarkan pemanfaatan teknologi deepfake dalam konteks penyebaran informasi palsu atau manipulatif [12]. Berbeda dengan hoaks digital konvensional yang umumnya berbasis teks atau manipulasi visual sederhana, deephoax mengandalkan proses otomatis berbasis kecerdasan buatan untuk menghasilkan manipulasi visual dengan tingkat realisme yang tinggi. Oleh karena itu, pada penelitian ini istilah *deephoax* digunakan untuk menekankan aspek penyalahgunaan teknologi deepfake sebagai sarana penipuan visual dan disinformasi digital. Penggunaan istilah ini relevan dengan tujuan penelitian, yaitu mendeteksi gambar wajah hasil *ai generated* yang berpotensi menyesatkan publik dan mengancam keamanan informasi, privasi individu, serta kepercayaan terhadap konten digital.

2.1.1 Deteksi Gambar Deephoax

Untuk mengatasi ancaman yang ditimbulkan oleh teknologi deepfake dan deephoax, berbagai penelitian telah mengembangkan sistem deteksi otomatis berbasis deep learning. Pada domain visual, pendekatan yang umum digunakan adalah Convolutional Neural Networks (CNN), yang mampu mengekstraksi fitur visual tingkat rendah hingga tingkat tinggi dari gambar wajah. Model CNN dapat mempelajari pola-pola visual tertentu yang sering muncul pada gambar hasil manipulasi, seperti ketidakkonsistenan tekstur kulit, artefak pada area mata dan