BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi *Artificial Intelligence* (AI) telah membawa pengaruh besar dalam pembuatan media digital. Teknologi ini mampu menghasilkan gambar dengan sangat realistis dan sulit dibedakan dengan gambar asli [1]. Salah satu bentuk pemanfaatan teknologi AI tersebut adalah *deepfake*, yaitu teknik yang digunakan untuk membuat, menggabungkan, atau memodifikasi gambar, video, maupun audio sehingga tampak menyerupai asli [2]. Fenomena ini di Indonesia terus mengalami peningkatan, seperti yang disampaikan pada siaran pers Komdigi tahun 2025 juga disebutkan bahwa kasus *deepfake* meningkat secara signifikan dan menjadi salah satu ancaman serius di ruang digital karena kontennya semakin sulit dibedakan dari media asli.

Pada September 2025, media sosial ramai dengan tren pembuatan foto bersama idola K-Pop, artis, hingga atlet menggunakan AI dengan gaya yang tidak pantas [3]. Fenomena tersebut menunjukkan bahwa teknologi AI kini tidak hanya dimanfaatkan untuk kebutuhan kreatif, tetapi juga berpotensi disalahgunakan untuk membuat gambar wajah sintetis yang manipulatif dan melanggar privasi. Salah satu kasus yang menarik perhatian publik adalah atlet timnas Indonesia Justin Hubner yang fotonya diedit menggunakan AI sehingga tampak sedang mencium seorang wanita. Selain Justin Hubner, beberapa artis seperti Adipati Dolken, Adhisty Zara, dan Tissa Biani juga menyampaikan keresahan mereka terhadap penggunaan AI yang menjadikan artis sebagai objek fantasi digital [3].

Maraknya penyebaran gambar hasil generatif AI menimbulkan berbagai permasalahan, seperti pencemaran nama baik, pelanggaran privasi, penipuan identitas, dan penyebaran disinformasi [2], yang penyebarannya meluas dan mudah diakses. Misalnya, pada tahun 2023 laporan terbaru menunjukkan lebih dari 500.000 video dan manipulasi suara telah teridentifikasi, kenaikan ini menggambarkan sekitar 550% sejak tahun 2019 [4]. Berdasarkan sebuah studi tentang kemampuan manusia dalam mengenali citra yang dihasilkan AI, ditemukan bahwa sekitar 39 – 40 % gambar yang sepenuhnya dihasilkan oleh AI tidak dapat diidentifikasi manusia [5] [6]. Kondisi ini menunjukkan adanya kebutuhan terhadap

sistem deteksi otomatis yang mampu mengklasifikasi gambar wajah asli dan palsu lebih akurat dan efisien.

Banyak metode yang digunakan untuk klasifikasi gambar hasil AI, misalnya CNN mencapai akurasi 92,98% pada dataset CIFAKE untuk gambar sintesis [7]. Penelitian sebelumnya mengembangkan CNN terbukti efektif dalam membedakan wajah asli dan palsu dan mampu mengenali fitur spasial yang kompleks pada gambar [8]. Adapun metode yang digunakan untuk mengklasifikasi *deepfake*, diantaranya penelitian sebelumnya banyak menggunakan arsitektur klasik seperti *Convolutional Neural Network* (CNN), *Deep Neural Network* (DNN), *Recurrent Neural Network* (RNN) dll [9].

Namun, efisiensi data dan model *transformer* selain ViT menjadi perbedaan signifikan dengan penelitian sebelumnya. Model *Vision Transformer* (ViT) umumnya memerlukan jumlah data pelatihan yang besar serta daya komputasi tinggi agar dapat mencapai performa optimal, sehingga kurang efisien ketika diaplikasikan pada *small dataset* atau data berukuran terbatas [10]. Sementara itu, penelitian yang mendeteksi dan mengklasifikasi keretakan jalan menunjukkan bahwa *Data Efficient Image Transformer* (DeiT) mampu bekerja secara lebih efektif pada dataset kecil dibandingkan ViT. Penelitian tersebut menggunakan subset data berjumlah sekitar 1.200 hingga 2.400 citra, dan DeiT terbukti menghasilkan performa yang lebih stabil serta akurat dengan akurasi tertinggi 99,75% dalam kondisi jumlah data yang terbatas [11]. Riset sebelumnya pada deteksi tulisan tangan palsu yang dihasilkan oleh AI, menunjukkan bahwa *Data-Efficient Image Transformer* (DeiT) efisien pada data yang lebih kecil sekitar 1.200 gambar dengan 4 kelas yang digunakan mencapai akurasi 98,46% [12]. Temuan-temuan tersebut mengindikasikan bahwa DeiT dirancang untuk tetap efektif pada skala data kecil hingga menengah (±1.000 - 5.000 citra), sehingga sesuai untuk penelitian ini yang menggunakan dataset dibawah 2.000 citra wajah.

Dari perbedaan tersebut munculah inovasi berbasis kecerdasan buatan (AI), untuk mengklasifikasi gambar generatif AI khusus pada wajah manusia. Salah satunya model DeiT memiliki fleksibilitas melalui beberapa varian model, seperti DeiT Base dan DeiT Tiny, yang dapat disesuaikan dengan kebutuhan komputasi tanpa mengorbankan akurasi [13]. DeiT menawarkan metode deteksi gambar yang

menunjukkan performa yang baik dalam tugas klasifikasi [14]. Oleh karena itu, penelitian ini menawarkan kebaruan pada aspek efisiensi data yang terbatas, adaptasi model DeiT varian Tiny, dan penerapan DeiT yang berfokus khusus dalam klasifikasi gambar asli atau palsu hasil generatif AI, yang sebelumnya belum banyak dieksplorasi. Penelitian ini menekankan pada efisiensi data serta evaluasi kemampuan model DeiT-Tiny dibandingkan dengan penelitian sebelumnya yang memiliki jumlah data besar [1] [15], [16].

Penelitian ini penting, karena penyebaran gambar wajah palsu hasil AI semakin meningkat dan berpotensi menimbulkan dampak sosial seperti penipuan identitas, penyebaran hoaks, hingga pelanggaran etika digital. Sistem deteksi yang efisien dan akurat sangat dibutuhkan untuk meningkatkan kepercayaan terhadap konten visual di era generatif AI. Sebagaimana dijelaskan dalam Al-Quran surat Al-Hujurat ayat (6) yang berbunyi: "*Wahai orang-orang yang beriman, jika datang kepada kamu orang fasik membawa suatu berita, maka periksalah dengan teliti agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu*." Diharapkan penelitian ini juga bisa dijadikan literatur dalam bidang akademik mengenai penerapan transformer yang lebih fleksibel di berbagai kondisi. Penelitian ini juga menggunakan metode pengembangan CRISP-DM sebagai kerangka kerja pembangunan sistem klasifikasi tersebut. Dengan itu penelitian akan berfokus pada ***"PENERAPAN MODEL DATA-EFFICIENT IMAGE TRANSFORMER (DEIT) UNTUK KLASIFIKASI GAMBAR WAJAH ASLI ATAU PALSU HASIL GENERATIF AI".***

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan dapat dirumuskan masalah dalam penelitian ini sebagai berikut:

1. Bagaimana menerapkan model *Data-Efficient Image Transformer* (DeiT) untuk klasifikasi gambar wajah asli atau palsu hasil generatif AI?
2. Bagaimana kinerja model *Data-Efficient Image Transformer* (DeiT) untuk deteksi gambar wajah asli atau palsu hasil generatif AI pada dataset kecil?

1.3 Batasan Masalah

Agar fokus penelitian terarah dan penelitian dapat mudah dipahami secara menyeluruh, serta mempersempit ruang lingkup penelitian, maka ditetapkan batasan-batasan masalah sebagai berikut:

1. Penelitian hanya berfokus pada klasifikasi gambar wajah manusia hasil generatif AI, bukan video, atau audio.
2. Data yang digunakan pada penelitian merupakan data gambar wajah asli pada manusia dan gambar palsu yang merupakan hasil generatif AI (bukan hasil manipulasi atau *editing* gambar).
3. Penelitian berupa klasifikasi gambar wajah asli pada manusia dan gambar palsu yang merupakan hasil generatif AI.
4. Pengujian dilakukan pada beberapa skenario yaitu augmentasi, *undersampling*, dan *oversampling*.
5. Penelitian hanya mengimplementasikan model dalam bentuk prototipe untuk pengujian fungsionalitas.

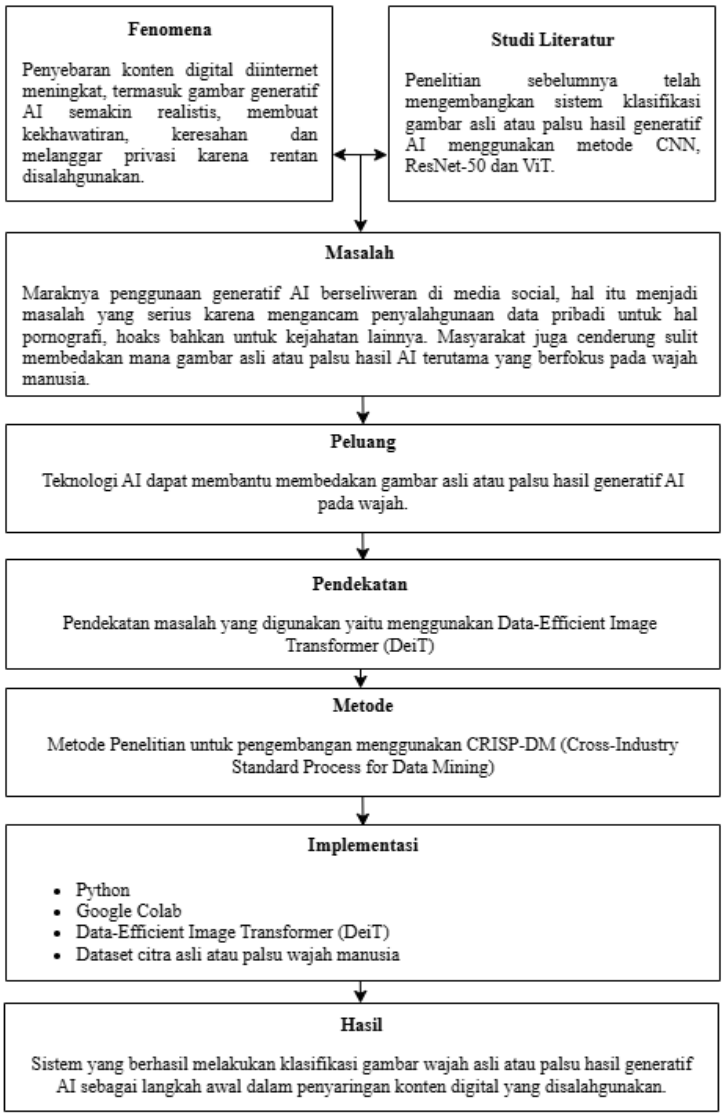
1.4 Tujuan

1. Menerapkan model *Data-Efficient Image Transformer* (DeiT) untuk klasifikasi gambar wajah asli atau palsu hasil generatif AI.
2. Mengetahui kinerja model *Data-Efficient Image Transformer* (DeiT) untuk deteksi gambar wajah asli atau palsu hasil generatif AI pada dataset kecil.

1.5 Manfaat

1. Penelitian ini diharapkan membantu dalam literatur ilmiah terutama pada penerapan *Data-Efficient Image Transformer* (DeiT) untuk klasifikasi gambar wajah asli atau palsu hasil generatif AI.
2. Memberikan referensi terkait cara penerapan CRISP-DM untuk membuat sistem klasifikasi gambar wajah asli atau palsu hasil generatif AI.
3. Membantu masyarakat untuk lebih peduli terhadap ancaman gambar hasil AI yang semakin meluas.

1.6 Kerangka Pemikiran



Gambar 1.1 Kerangka Pemikiran

Gambar 1.1 merupakan kerangka pemikiran menggambarkan alur berfikir penelitian ini. Adapun penjelasan dari kerangka pemikiran tersebut disajikan sebagai berikut.

1. Fenomena

   Perkembangan teknologi *Artificial Intelligence* (AI) bisa digunakan untuk membuat *deepfake*. Fenomena *deepfake* di Indonesia semakin meningkat dan menjadi ancaman di ruang digital karena sulit dibedakan dari media asli. Media sosial ramai dengan penyalahgunaan AI untuk membuat foto manipulatif bersama artis, atlet, atau figur publik dengan pose yang tidak pantas.

2. Masalah

   Gambar hasil AI yang semakin realistis menyebabkan masyarakat sulit membedakan gambar asli dan palsu secara manual. Hal itu disebabkan hasil AI semakin meningkat dan realistis. Sehingga dibutuhkan sistem klasifikasi otomatis yang mampu membedakan gambar wajah asli dan palsu secara akurat dan efisien, terutama pada dataset terbatas.

3. Studi Literatur

   Penelitian sebelumnya menggunakan metode CNN, DNN, RNN, dan Vision Transformer (ViT) untuk klasifikasi gambar hasil AI.

4. Peluang

   Teknologi AI dapat dimanfaatkan untuk membantu mengklasifikasi gambar wajah asli maupun palsu hasil AI generatif. Misalnya menggunakan model DeiT yang berpotensi menjadi solusi karena mampu bekerja efektif pada data berukuran kecil hingga menengah.

5. Pendekatan

   Pendekatan penelitian menggunakan Data-Efficient Image Transformer (DeiT) varian Tiny untuk klasifikasi gambar wajah asli atau palsu hasil AI.

6. Metode

   Pengembangan sistem klasifikasi dilakukan menggunakan metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*).

7. Implementasi

   Implementasi penelitian menggunakan Python dan Google Colab, sedangkan dataset yang digunakan berupa citra wajah asli dan palsu hasil AI generatif. Model yang digunakan adalah *Data-Efficient Image Transformer* (DeiT-Tiny).

8. Hasil

   Sistem mampu mengklasifikasikan gambar wajah asli dan palsu hasil AI generatif sebagai langkah awal dalam membantu deteksi konten manipulatif di media digital.