

# BAB I

## PENDAHULUAN

### A. Latar Belakang

Perkembangan *Artificial Intelligence* (AI), khususnya dalam bidang *deep learning*, telah mendorong kemajuan signifikan pada teknologi generatif, yaitu model yang mampu menghasilkan data baru yang menyerupai data asli berdasarkan pola yang dipelajari dari data pelatihan. Salah satu model generatif awal yang berperan penting adalah *Variational Autoencoders* (VAE) yang diperkenalkan pada tahun 2013, dengan pendekatan ruang laten berbasis distribusi probabilistik seperti *Gaussian* sehingga memungkinkan generasi data melalui proses interpolasi. Selanjutnya, *Generative Adversarial Networks* (GANs) menghadirkan mekanisme pelatihan *adversarial* antara generator dan *discriminator* yang mampu menghasilkan data dengan kualitas tinggi. Perkembangan terbaru ditandai dengan hadirnya *Diffusion Models*, yang memodelkan proses generasi data sebagai proses difusi bertahap untuk menghasilkan konten visual yang semakin realistis [1]. Evolusi model-model generatif ini menunjukkan bahwa kemampuan AI dalam menghasilkan citra dan video sintetis semakin canggih dan sulit dibedakan dari data asli.

Kemampuan model generatif dalam menghasilkan konten visual yang semakin realistis tersebut turut memunculkan fenomena *deepfake*, yaitu teknik yang memanfaatkan AI untuk memanipulasi identitas, ekspresi, atau perilaku seseorang sehingga dapat menghasilkan gambar maupun video palsu yang tampak sangat realistis [2]. Fenomena ini menimbulkan kekhawatiran terhadap keaslian informasi digital karena konten yang telah direkayasa dapat memicu disinformasi dan menurunkan kepercayaan publik terhadap media digital [3]. Kekhawatiran ini semakin relevan dengan munculnya berbagai kasus di Indonesia, seperti penyebaran video *deepfake* yang menggunakan wajah Presiden dalam modus penipuan bantuan dana [4], serta kasus mahasiswa yang dikeluarkan dari Universitas Udayana akibat membuat konten *deepfake* bermuatan pelecehan [5]. Berbagai insiden tersebut menegaskan pentingnya pengembangan sistem deteksi *deepfake* yang akurat dan andal untuk mencegah penyalahgunaan teknologi ini.

Dalam mendeteksi *deepfake*, berbagai penelitian menunjukkan bahwa citra dan video hasil manipulasi umumnya memiliki anomali yang muncul akibat ketidaksempurnaan proses generatif dari model pembuatnya [6]. Meskipun secara visual tampak realistis, konten *deepfake* sering kali mengandung ketidakonsistenan yang dapat dideteksi melalui analisis pola tertentu. Anomali tersebut dapat dikategorikan menjadi tiga jenis. Pertama, *visual artifacts*, seperti tepi wajah yang kabur, pencahayaan yang tidak konsisten, atau tekstur kulit yang tampak tidak wajar akibat proses sintesis yang kurang sempurna [7]. Kedua, *physiological irregularities*, yaitu ketidakwajaran pada aspek biologis manusia, seperti pola kedipan mata, ekspresi wajah, atau gerakan kepala yang tidak alami. Ketiga, *temporal inconsistencies*, yang umumnya ditemukan pada konten berbasis video, berupa ketidaksinkronan antara ucapan dan gerakan bibir atau perubahan ekspresi yang tidak selaras antar frame [8].

Berdasarkan keberadaan berbagai anomali tersebut, pendekatan berbasis *deep learning* menjadi salah satu metode yang paling banyak digunakan dalam mendeteksi *deepfake*. *Convolutional Neural Network* (CNN) merupakan salah satu model *deep learning* paling umum yang digunakan untuk mendeteksi *deepfake*. Model CNN seperti MesoNet menunjukkan tingkat ketahanan yang baik karena mampu mempelajari pola manipulasi secara otomatis langsung dari data [7]. Selain itu, arsitektur ResNet-34 juga banyak digunakan karena kemampuannya mempelajari fitur hierarkis yang kompleks sehingga pelatihan menjadi lebih stabil dan efisien [8]. Kelebihan CNN terletak pada efisiensi komputasi serta kemampuannya mendeteksi detail lokal pada gambar. Namun, CNN sering kali mengabaikan informasi global yang kompleks serta fitur pada domain frekuensi yang menjadi karakteristik penting citra *deepfake*, sehingga kurang efektif dalam menangani bentuk manipulasi yang lebih canggih [9].

Disisi lain, model berbasis *Transformer* seperti *Vision Transformer* (ViT) dirancang untuk memahami hubungan global pada citra melalui mekanisme *self-attention*. Pendekatan ini memungkinkan model untuk mengenali pola konteks yang lebih luas antar bagian citra, sehingga lebih adaptif terhadap variasi manipulasi dari berbagai generator [10]. Namun, kekurangan utama arsitektur ini adalah kebutuhan komputasi yang tinggi dan ketergantungan pada jumlah data

pelatihan yang besar agar dapat mencapai performa optimal [11]. Selain itu, model Transformer murni cenderung kurang sensitif terhadap detail lokal yang sangat halus pada permukaan wajah yang sering kali menjadi indikator penting dalam mendeteksi *deepfake* [12].

Berbagai penelitian sebelumnya menyebutkan bahwa pendekatan *hybrid CNN-Transformer* dinilai mampu mengatasi keterbatasan dari kedua arsitektur tunggal dengan menggabungkan keunggulan keduanya dalam satu sistem deteksi. Penelitian [13] yang mengintegrasikan model spasial dan temporal, yaitu CNN seperti ResNet50, Xception, dan ResNeXt, dikombinasikan dengan *Long Short-Term Memory* (LSTM) serta *Vision Transformer* (ViT). Model tersebut dilatih menggunakan dataset publik seperti FaceForensics++, Celeb-DF, dan Deepfake Detection Challenge, dan mampu mencapai akurasi hingga 98,7%. Penelitian lain [14] mengusulkan model *Generative Convolutional Vision Transformer* (GenConViT) untuk deteksi video *deepfake* dengan menggabungkan arsitektur ConvNeXt dan *Swin Transformer* sebagai ekstraktor fitur, serta memanfaatkan *Autoencoder* dan *Variational Autoencoder* (VAE) untuk mempelajari distribusi data laten. Diuji pada beberapa dataset seperti DFDC, FaceForensics++, TM, DeepfakeTIMIT, dan Celeb-DF (v2), GenConViT mencapai akurasi rata-rata 95,8% dan nilai AUC 99,3%, menandakan efektivitasnya dalam mendeteksi konten manipulatif secara luas.

Pada penelitian ini, arsitektur yang digunakan mengombinasikan ResNet-34 sebagai *backbone* CNN dan *Data-efficient Image Transformer* (DeiT-Small) sebagai komponen *Transformer* dalam skema *hybrid* yang berjalan secara paralel (*dual-branch architecture*). Pada tahap awal, citra wajah yang sama diproses secara bersamaan oleh kedua model. ResNet-34 berperan sebagai ekstraktor fitur lokal yang efektif untuk menangkap detail tekstur, pencahayaan, dan kontur wajah [8], sementara DeiT-Small digunakan untuk memproses hubungan global antar fitur dengan efisiensi data dan kebutuhan komputasi yang lebih rendah dibandingkan *Vision Transformer* murni [11].

*Output* dari masing-masing cabang berupa representasi fitur tingkat tinggi kemudian digabungkan pada tahap *feature-level fusion* menggunakan metode *concatenation*[15]. Pada proses ini, vektor fitur dari ResNet-34 dan DeiT-Small

disusun membentuk representasi *hybrid* yang lebih komprehensif. Representasi *hybrid* tersebut selanjutnya diteruskan ke lapisan *fully connected* untuk proses klasifikasi. Pendekatan paralel dengan *feature concatenation* ini memungkinkan sistem mempertahankan kelebihan masing-masing arsitektur sekaligus mengintegrasikan informasi lokal dan global secara seimbang, sehingga meningkatkan kemampuan deteksi manipulasi kompleks pada citra *deepfake*.

## **B. Perumusan Masalah**

Berdasarkan latar belakang dan judul penelitian, rumusan masalah dalam penelitian ini adalah bagaimana mengevaluasi kinerja model *hybrid* CNN–*Transformer* dalam mendeteksi *deepfake* dengan menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*.

## **C. Batasan Penelitian**

Agar penelitian ini lebih terarah dan fokus, maka ruang lingkup penelitian dibatasi pada beberapa hal sebagai berikut:

1. Penelitian ini hanya berfokus pada deteksi *deepfake* berbasis citra wajah, tidak mencakup deteksi pada video maupun audio.
2. Dataset pelatihan yang digunakan terdiri dari dua versi:
  - a. Dataset single-source: 20.000 citra yang seluruhnya bersumber dari GRAVEX-200K Real vs AI-Generated Faces (Kaggle), dengan komposisi seimbang 10.000 citra *real* dan 10.000 citra *fake*.
  - b. Dataset multi-source: 20.776 citra yang merupakan gabungan dari empat sumber, yaitu GRAVEX-200K (10.470 citra), FaceForensics++ C23 (6.000 citra), Deepfake and Real Images (3.530 citra), dan citra dengan *background removal* (776 citra), dengan komposisi seimbang 10.388 citra *real* dan 10.388 citra *fake*.
3. Pengujian generalisasi lintas dataset (*cross-dataset evaluation*) hanya dilakukan menggunakan satu dataset eksternal, yaitu StyleGAN–StyleGAN2 Deepfake Face Images yang terdiri dari 11.780 citra dengan distribusi seimbang (5.890 citra *real* dan 5.890 citra *fake*).

4. Arsitektur model dibatasi pada ResNet-34 sebagai *backbone* CNN dan DeiT-Small sebagai *backbone* Transformer, dengan mekanisme *feature fusion concatenation*. Eksplorasi *backbone* lain maupun metode *fusion* yang lebih adaptif tidak termasuk dalam lingkup penelitian ini.
5. Proses deteksi hanya dilakukan pada citra wajah yang sudah ter-*crop* atau berfokus pada area wajah. Sistem tidak mencakup tahap *face detection* atau *face cropping* otomatis sebelum proses klasifikasi.
6. Dataset yang digunakan dalam penelitian ini mencakup citra *deepfake* yang dihasilkan oleh berbagai metode generatif, termasuk GAN (*Generative Adversarial Network*) dan beberapa *diffusion model* awal. Namun, kemampuan model terhadap citra yang dihasilkan oleh generator generatif terbaru belum dievaluasi secara khusus.
7. Evaluasi model hanya mencakup *accuracy*, *precision*, *recall*, dan *F1-score*.

#### **D. Tujuan Penelitian**

Berdasarkan rumusan masalah tersebut, maka tujuan penelitian ini adalah mengevaluasi kinerja model yang dihasilkan menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengetahui tingkat efektivitas model dalam mendeteksi *deepfake*.

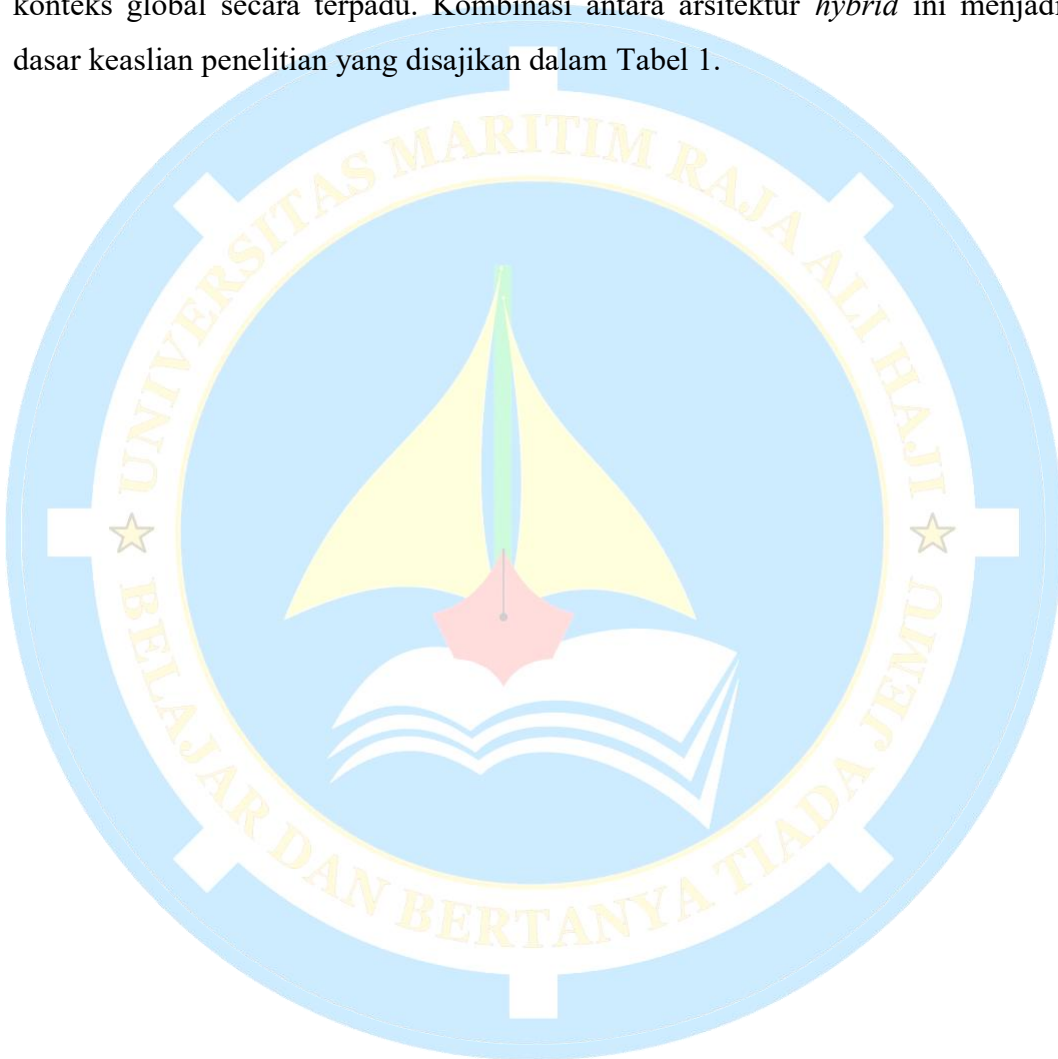
#### **E. Manfaat Penelitian**

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Secara teoretis, penelitian ini dapat memperkaya kajian di bidang kecerdasan buatan, khususnya mengenai penerapan arsitektur *hybrid* CNN–*Transformer* dalam deteksi *deepfake* pada citra wajah.
2. Secara praktis, penelitian ini diharapkan menghasilkan model deteksi *deepfake* berbasis arsitektur *hybrid* yang dapat digunakan sebagai referensi dalam pengembangan sistem verifikasi keaslian citra wajah digital serta mendukung upaya peningkatan keamanan dan kepercayaan terhadap konten digital.

## F. Keaslian Penelitian

Penelitian ini mengusulkan arsitektur *hybrid* yang menggabungkan ResNet-34 sebagai representasi (CNN) untuk mengekstraksi fitur lokal dan DeiT-Small sebagai representasi *Transformer* untuk memodelkan hubungan global pada citra wajah dalam mendeteksi *deepfake*. Integrasi kedua arsitektur ini bertujuan untuk membangun representasi fitur yang menggabungkan informasi spasial lokal dan konteks global secara terpadu. Kombinasi antara arsitektur *hybrid* ini menjadi dasar keaslian penelitian yang disajikan dalam Tabel 1.



Tabel 1. Keaslian Penelitian

| Indikator | Penelitian Sebelumnya                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Penelitian yang Diusulkan                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Judul     | <i>Deepfake detection using convolutional vision transformers and convolutional neural networks</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Deteksi <i>Deepfake</i> pada Citra Wajah Menggunakan Arsitektur <i>Hybrid CNN-Transformer</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Metode    | <i>Convolutional Neural Network (CNN)</i> dan <i>Convolutional Vision Transformer (CViT)</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                               | ResNet-34 dan <i>Data-efficient Image Transformer-Small (DeiT-Small)</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Hasil     | Penelitian ini mengusulkan metode deteksi <i>deepfake</i> berbasis <i>deep learning</i> yang terdiri dari tiga komponen utama: pra-pemrosesan, deteksi, dan prediksi. Pada tahap deteksi, digunakan <i>Convolutional Neural Network (CNN)</i> untuk mendeteksi fitur mata dan hidung, serta kombinasi CNN dan <i>Vision Transformer (ViT)</i> untuk deteksi wajah [16]. Komponen prediksi menggunakan pendekatan <i>majority voting</i> untuk menggabungkan hasil dari tiga model berbeda. Model ini dilatih menggunakan dataset FaceForensics++ dan DFDC, | Penelitian ini mengusulkan arsitektur <i>hybrid</i> untuk deteksi citra <i>deepfake</i> dengan menggabungkan dua pendekatan ekstraksi fitur yang berbeda. ResNet-34 digunakan sebagai <i>Convolutional Neural Network (CNN)</i> untuk mengekstraksi fitur lokal, seperti pola tekstur, tepi, serta artefak manipulasi yang sering muncul pada citra hasil <i>deepfake</i> . Sementara itu, DeiT-Small digunakan sebagai <i>Vision Transformer</i> untuk mempelajari hubungan global antar- <i>patch</i> pada gambar, sehingga model dapat memahami konteks keseluruhan wajah secara lebih menyeluruh. |

---

dengan hasil menunjukkan akurasi 97% pada model CNN dan 85% pada model CViT.

Selanjutnya, representasi fitur yang dihasilkan oleh kedua model tersebut digabungkan menggunakan metode *Feature Fusion Concatenation* pada level representasi (*feature-level fusion*). Vektor fitur gabungan ini kemudian diproses oleh lapisan klasifikasi tunggal untuk menghasilkan prediksi akhir berupa label *real* atau *fake*.

---

