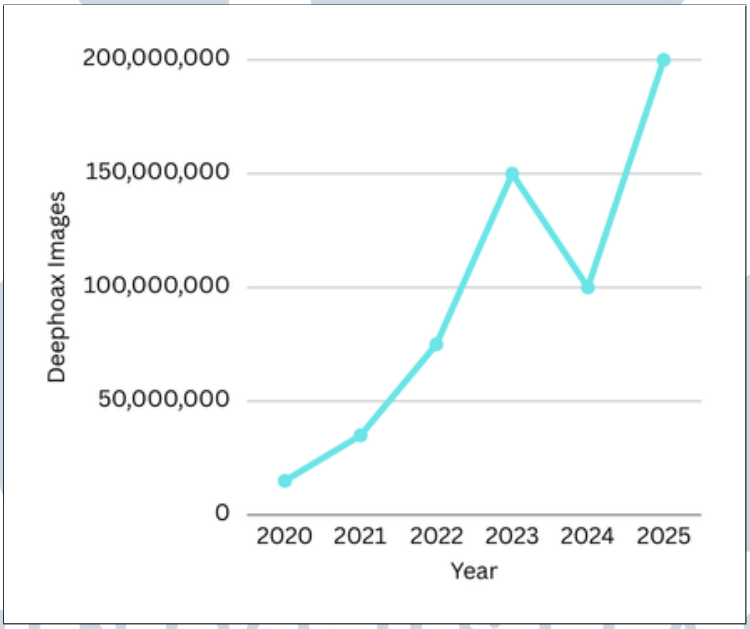
BAB 1
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi kecerdasan buatan (AI) telah membawa perubahan besar dalam bidang pengolahan citra digital. Salah satu inovasi yang paling signifikan adalah teknologi *deephoax*, yaitu teknik yang memanfaatkan model *deep learning* untuk menghasilkan atau memanipulasi wajah manusia secara realistis. Awalnya dikembangkan untuk kepentingan hiburan seperti film dan media sosial, kini teknologi ini sering disalahgunakan untuk menyebarkan disinformasi visual, manipulasi politik, hingga penipuan digital yang mengancam privasi dan keamanan siber [1]. Fenomena ini menjadikan deteksi *deephoax* sebagai topik penting dalam upaya menjaga keaslian dan kepercayaan terhadap konten digital.



Gambar 1.1. Tren peningkatan jumlah konten *deephoax* secara global pada tahun 2020–2025.

Dengan meningkatnya kemampuan model generatif modern seperti StyleGAN3, DALL·E 3, dan Stable Diffusion 3, kualitas citra hoax yang dihasilkan kini semakin sulit dibedakan dari citra fakta, baik oleh manusia maupun sistem deteksi otomatis. Fenomena ini terlihat jelas pada tren peningkatan jumlah konten *deephoax* yang beredar di internet selama lima tahun terakhir. Sebagaimana

ditunjukkan pada Gambar 1.1, jumlah *deephoax* global meningkat secara drastis dari sekitar 15 juta pada tahun 2020 menjadi lebih dari 200 juta kasus pada tahun 2025. Tren eskalatif ini sejalan dengan temuan Chauhan et al. (2025), yang melaporkan bahwa lonjakan kualitas dan jumlah konten sintetik dipicu oleh kemajuan model generatif berbasis CNN, GAN, dan Transformer [2]. Berdasarkan laporan resmi Kementerian Komunikasi dan Digital (Kemkomdigi) pada September 2025, jumlah konten *deephoax* di Indonesia meningkat hingga 550% dibandingkan tahun sebelumnya [3]. Peningkatan ini menunjukkan bahwa penyalahgunaan teknologi sintesis berbasis AI semakin meluas, bahkan lebih cepat daripada kemampuan sistem deteksi konvensional dalam mengikuti perkembangan teknologi tersebut.

Seiring meningkatnya kompleksitas dan realisme konten *deephoax*, model deteksi *deephoax* berbasis *deep learning* menghadapi tantangan serius dari *adversarial attack*, yaitu gangguan kecil pada citra masukan yang secara visual hampir tidak terlihat namun mampu menyesatkan proses klasifikasi [4]. Dalam skenario *white-box*, metode *Projected Gradient Descent* (PGD) merupakan salah satu serangan paling efektif karena secara iteratif memanfaatkan gradien model untuk menghasilkan perturbasi yang berdampak signifikan terhadap keputusan klasifikasi. Chen et al. [5] menunjukkan bahwa model deteksi *deephoax* dengan akurasi tinggi pada data bersih mengalami penurunan performa yang sangat drastis ketika dievaluasi menggunakan serangan PGD iteratif, dengan akurasi turun hingga mendekati 0%–0.13%. Temuan yang konsisten juga dilaporkan oleh Jia et al. [6], yang menunjukkan bahwa serangan PGD memiliki tingkat keberhasilan tinggi terhadap berbagai arsitektur model deteksi *deephoax*, dengan *attack success rate* berkisar antara 77.7% hingga 85.4%. Selain itu, Liu et al. [7] melaporkan bahwa akurasi rata-rata berbagai model deteksi *deephoax* dapat turun hingga berada pada kisaran sekitar 45% di bawah serangan PGD, meskipun sebelumnya menunjukkan performa yang sangat baik pada data bersih, yang mengindikasikan bahwa performa tinggi pada data bersih tidak secara langsung menjamin ketahanan model terhadap eksploitasi gradien.

Sebagai respons terhadap permasalahan tersebut, pendekatan pertahanan berbasis *adversarial training* (AT) banyak diusulkan dalam literatur sebagai strategi yang efektif untuk meningkatkan ketahanan model terhadap serangan *adversarial*. *Adversarial training* dilakukan dengan melibatkan contoh *adversarial* yang dihasilkan selama proses pelatihan, sehingga model secara eksplisit dilatih untuk mengenali dan mengklasifikasikan data yang telah mengalami perturbasi. Madry et

al. [8] memperkenalkan AT berbasis *Projected Gradient Descent* (PGD) sebagai pendekatan standar dalam membangun model yang memiliki ketahanan tinggi terhadap serangan *white-box*. Sejalan dengan temuan tersebut, Abed et al. dan Banait et al. [9, 10] menunjukkan bahwa *adversarial training* secara konsisten lebih efektif dibandingkan mekanisme pertahanan lainnya ketika dievaluasi terhadap serangan *adversarial* berbasis PGD. Temuan-temuan tersebut menegaskan bahwa *adversarial training* merupakan pendekatan pertahanan yang kuat dan relevan dalam menghadapi ancaman *adversarial* pada model pembelajaran mendalam.

Meskipun *adversarial training* berbasis *Projected Gradient Descent* (PGD) terbukti efektif dalam meningkatkan ketahanan model terhadap serangan *adversarial*, sebagian besar penelitian masih berfokus pada dataset klasifikasi citra umum sehingga belum sepenuhnya merepresentasikan karakteristik visual serta pola manipulasi spesifik pada citra wajah *deephoax*. Selain itu, penerapan *adversarial training* berbasis PGD menimbulkan biaya komputasi yang tinggi akibat sifat serangan PGD yang bersifat iteratif. Oleh karena itu, pemilihan arsitektur dengan kompleksitas yang seimbang menjadi faktor krusial dalam pengembangan sistem deteksi *deephoax*. Sejumlah penelitian menunjukkan bahwa arsitektur Efficientnet-B0 mampu mencapai tingkat ketahanan terhadap serangan PGD yang sebanding dengan model berukuran lebih besar, tanpa peningkatan signifikan pada kebutuhan komputasi [11–13]. Berdasarkan temuan tersebut, EfficientNet-B0 dipandang sebagai arsitektur yang relevan untuk dikombinasikan dengan adversarial training berbasis PGD guna mencapai keseimbangan antara *robustness* dan efisiensi komputasi.

Berdasarkan permasalahan dan celah penelitian tersebut, penelitian ini berfokus pada penerapan *adversarial training* berbasis serangan *Projected Gradient Descent* (PGD) pada sistem deteksi *deephoax* menggunakan arsitektur EfficientNet-B0. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem deteksi *deephoax* yang lebih tangguh terhadap serangan *adversarial*, serta mendukung keamanan siber dengan meningkatkan keandalan deteksi konten hoax dan mencegah penyebaran informasi yang merugikan masyarakat.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah yang diajukan dalam penelitian ini adalah sebagai berikut:

1. Bagaimana pengaruh penerapan *adversarial training* terhadap *robustness*

model EfficientNet-B0 dalam mendeteksi citra *deephoax* terhadap serangan *Projected Gradient Descent* (PGD)?

2. Bagaimana perbandingan performa model EfficientNet-B0 dengan dan tanpa *adversarial training* dalam mendeteksi citra *deephoax* pada data bersih dan data yang telah mengalami serangan *Projected Gradient Descent* (PGD)?

1.3 Batasan Permasalahan

Batasan permasalahan dalam penelitian ini ditetapkan sebagai berikut:

1. Penelitian ini berfokus pada deteksi *deephoax* berbasis citra wajah statis dan tidak mencakup data video maupun audio.
2. Arsitektur model yang digunakan dibatasi pada *EfficientNet-B0* dengan pertimbangan efisiensi komputasi dan jumlah parameter.
3. Pendekatan pertahanan *adversarial* yang diterapkan adalah *adversarial training* berbasis serangan *Projected Gradient Descent* (PGD).
4. Jenis serangan *adversarial* yang digunakan terbatas pada skenario *white-box* berbasis *Projected Gradient Descent* (PGD).
5. Dataset yang digunakan adalah *DeepDetect-2025* yang terdiri atas citra wajah fakta dan citra hasil model generatif modern seperti StyleGAN3, DALL·E 3, dan Stable Diffusion 3.
6. Evaluasi performa model dilakukan menggunakan metrik *clean accuracy*, *PGD accuracy*, *attack success rate* (ASR), *precision*, *recall*, dan *F1-score*.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, tujuan penelitian ini adalah sebagai berikut:

1. Menganalisis pengaruh penerapan *adversarial training* terhadap *robustness* model EfficientNet-B0 dalam mendeteksi citra *deephoax* terhadap serangan *Projected Gradient Descent* (PGD).
2. Membandingkan performa model EfficientNet-B0 dengan dan tanpa *adversarial training* dalam mendeteksi citra *deephoax* pada data bersih dan data yang telah mengalami serangan *Projected Gradient Descent* (PGD).

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. **Manfaat Akademis**: Memberikan kontribusi ilmiah berupa kajian empiris mengenai penerapan *adversarial training* berbasis PGD pada sistem deteksi *deephoax* berbasis citra wajah.
2. **Manfaat Praktis**: Menyediakan pendekatan pertahanan *adversarial* yang efisien dan dapat diterapkan pada sistem deteksi *deephoax* dengan keterbatasan sumber daya komputasi.
3. **Manfaat Sosial**: Mendukung upaya mitigasi penyebaran konten visual manipulatif berbasis AI melalui penguatan ketahanan sistem keamanan digital terhadap ancaman *deephoax*.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini disusun dalam lima bab sebagai berikut:

1. **Bab I Pendahuluan**
   Bab ini memuat latar belakang penelitian, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, serta sistematika penulisan laporan.
2. **Bab II Landasan Teori**
   Bab ini membahas konsep teoretis yang mendasari penelitian, meliputi *deephoax*, model generatif modern, serangan *adversarial*, *adversarial training*, serta arsitektur *EfficientNet*. Selain itu, dibahas pula penelitian terdahulu yang relevan.
3. **Bab III Metodologi Penelitian**
   Bab ini menguraikan tahapan penelitian yang meliputi deskripsi dataset *DeepDetect-2025*, proses *preprocessing*, perancangan model EfficientNet-B0, *attack implementation*, penerapan *adversarial training* berbasis PGD, serta prosedur evaluasi model.

4. **Bab IV Hasil dan Pembahasan**
   Bab ini menyajikan hasil eksperimen serta pembahasan mengenai performa dan ketahanan model terhadap serangan *adversarial* berbasis PGD.

5. **Bab V Kesimpulan dan Saran**
   Bab ini memuat kesimpulan dari hasil penelitian dan saran untuk pengembangan penelitian selanjutnya.