BAB 2
LANDASAN TEORI

2.1 Keamanan Siber (Cyber Security)

Keamanan siber adalah disiplin multidimensi yang mempelajari prinsip, teknologi, dan praktik untuk melindungi sistem informasi, infrastruktur jaringan, dan data dari ancaman yang bersifat tidak sah, merusak, atau mengganggu operasi normal. Dalam kajian ini, keamanan siber didefinisikan sebagai rangkaian kebijakan, kontrol teknis, dan prosedur manajerial yang bersama-sama bertujuan menjamin kerahasiaan, integritas, dan ketersediaan aset digital [14, 15]. Perkembangan teknologi *Artificial Intelligence* dan pemodelan generatif telah memperluas permukaan serangan (attack surface) sehingga metode tradisional perlu diperkaya dengan pendekatan yang mempertimbangkan ancaman yang bersumber dari model generatif itu sendiri. Karena itu, paradigma pertahanan tidak lagi hanya fokus pada pencegahan akses tidak sah, melainkan juga deteksi manipulasi canggih pada tingkat konten multimedia, mitigasi penyalahgunaan model generatif, serta ketahanan (robustness) model deteksi terhadap serangan adversarial.

2.1.1 Ancaman dan Serangan dalam Keamanan Siber

Dalam konteks deteksi deephoax dan perlindungan konten, terdapat beberapa kategori ancaman yang relevan dan berisiko tinggi bagi integritas sistem serta kepercayaan publik. Berikut dijelaskan empat ancaman/serangan utama yang menjadi fokus penelitian ini:

1. Impersonation dan Social Engineering Berbasis Deephoax: Teknologi pembuatan wajah dan suara sintetis memungkinkan pelaku melakukan impersonation (penyamaran identitas) yang meyakinkan, sehingga dapat digunakan untuk penipuan finansial, pengambilalihan akun, atau penyebaran disinformasi. Kasus-kasus dimana figur publik atau eksekutif diimitasi secara audio/visual menunjukkan potensi kerugian finansial dan reputasi yang nyata [16, 17].

2. Adversarial Attack terhadap Sistem Deteksi Deephoax: Model pembelajaran mendalam yang digunakan untuk mendeteksi konten sintetis memiliki kerentanan terhadap *adversarial examples*, yaitu gangguan terencana

yang secara visual hampir tidak tampak bagi manusia tetapi cukup untuk menyebabkan model melakukan kesalahan dalam proses klasifikasi. Penelitian akademik dan laporan teknis menunjukkan bahwa serangan white-box maupun black-box dapat mengelabui detector deephoax, menimbulkan kebutuhan untuk metode pelatihan dan regularisasi yang meningkatkan robustnes model deteksi [18, 19].

3. Eksploitasi Rantai Pasokan dan Zero-day pada Infrastruktur AI: Penyalahgunaan atau kompromi model, pipeline data, atau perangkat lunak pihak ketiga (API generatif, modul preprocessing) dapat memberi akses terhadap kemampuan pembuatan deephoax dengan skala besar. Selain itu, vektor serangan tradisional (zero-day di perangkat jaringan) bila digabungkan dengan kemampuan generatif akan memperburuk dampak insiden [14, 15].

4. Penyalahgunaan Aplikasi dan Layanan *Consumer-grade*: Meningkatnya aksesibilitas aplikasi penghasil konten sintetik (layanan nudification, voice cloning sederhana) menurunkan ambang teknis bagi pelaku untuk membuat dan menyebarkan deephoax. Hal ini meningkatkan frekuensi insiden penyalahgunaan serta memperbesar skala korban, mulai dari individu hingga kelompok rentan [17].

2.1.2 Peran AI dan Machine Learning dalam Keamanan Siber

Kecerdasan buatan berperan ganda dalam ekosistem keamanan siber: di satu sisi AI memperkuat kemampuan deteksi ancaman melalui klasifikasi anomali, analisis perilaku jaringan, dan ekstraksi pola pada data besar; di sisi lain, AI juga menjadi sumber ancaman baru ketika model generatif dapat dipakai untuk membuat serangan yang lebih persuasif. Dalam ranah deteksi deephoax, teknik supervised learning (CNN, transformer-based architectures) dan metode hybrid (fitur forensik statis digabungkan dengan representasi mendalam) telah menunjukkan kinerja baik pada kondisi uji standard. Namun, ketika model detektor dihadapkan pada serangan adversarial atau contoh sintetis yang dihasilkan oleh model yang lebih baru, kinerja tersebut sering menurun drastis kecuali strategi pembelajaran robust (adversarial training, data augmentation berbasis penyamaran, dan ensemble) diaplikasikan [18, 20]. Selain itu, metode explainability (XAI) dan teknik verifikasi sumber (provenance, watermarking) menjadi pelengkap penting untuk memberikan bukti

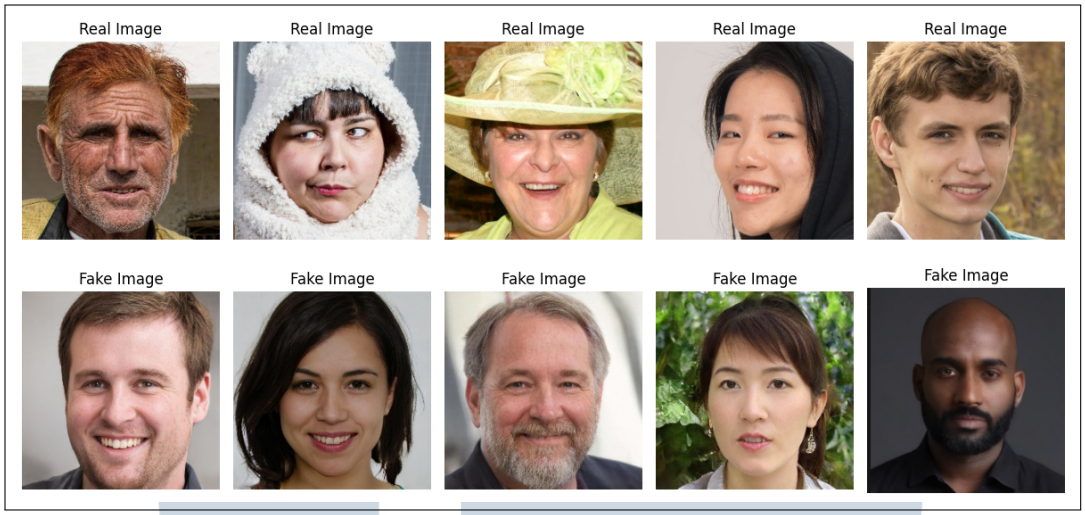
penelusuran asal-usul konten dan meningkatkan kepercayaan terhadap hasil deteksi.

2.1.3 Konsep Dasar Deephoax dan Perbedaannya dengan Deepfake

Istilah *deephoax* diperkenalkan dalam konteks keamanan siber untuk menggambarkan konten manipulatif yang dihasilkan menggunakan teknik pembelajaran mendalam dan dimanfaatkan sebagai sarana penyesatan, disinformasi, atau serangan digital terstruktur. Berbeda dengan istilah *deepfake* yang umumnya merujuk pada teknik atau hasil manipulasi media digital—terutama citra, audio, dan video—dengan fokus pada realisme sintesis, *deephoax* menekankan tujuan, konteks, dan dampak penggunaannya sebagai ancaman terhadap sistem informasi dan kepercayaan publik [21]. Dalam penelitian Yuliani *et al.*, *deephoax* diposisikan sebagai evolusi hoaks konvensional yang diperkuat oleh kemampuan model pembelajaran mendalam, seperti *transformer* dan *generative models*, sehingga mampu menghasilkan konten palsu dalam skala besar dan dengan tingkat kredibilitas yang tinggi. Sementara itu, *deepfake* dipandang sebagai teknologi generatif yang menjadi salah satu sarana pembentukan *deephoax*, namun tidak selalu digunakan untuk tujuan berbahaya [1, 22]. Dengan demikian, *deephoax* merepresentasikan spektrum ancaman keamanan siber yang lebih luas, sedangkan *deepfake* berperan sebagai metode atau teknik pendukung dalam pembentukan konten tersebut.

A Deephoax Berbasis Gambar Wajah

Deephoax berbasis gambar wajah merupakan salah satu bentuk ancaman visual yang memanfaatkan teknik manipulasi citra wajah berbasis pembelajaran mendalam untuk tujuan penyesatan atau penyalahgunaan identitas digital. Bentuk manipulasi ini umumnya dilakukan melalui teknik *face swapping*, *face reenactment*, dan *identity synthesis*, yang dikembangkan menggunakan arsitektur *autoencoder* dan *Generative Adversarial Networks* (GAN) [23]. Proses sintesis wajah melibatkan ekstraksi fitur wajah, pemetaan identitas antara citra sumber dan target, serta tahap *post-processing* untuk menyempurnakan tekstur, pencahayaan, dan kontur wajah agar menyerupai citra fakta [24, 25].



Gambar 2.1. Perbandingan citra wajah fakta dan citra wajah hoax hasil manipulasi berbasis *deep learning*.

Gambar 2.1 menunjukkan contoh perbandingan antara citra wajah fakta dan citra wajah sintetis hasil manipulasi berbasis *deep learning*. Secara visual, citra wajah sintetis tampak menyerupai citra fakta sehingga perbedaannya sulit dikenali secara kasat mata. Hal ini menunjukkan bahwa manipulasi wajah berbasis pembelajaran mendalam dapat menghasilkan citra dengan tingkat realisme yang tinggi.

2.1.4 Dampak Deephoax dalam Keamanan Siber

Deephoax berbasis gambar dan video telah menjadi ancaman serius dalam keamanan siber. Konten sintetis yang sangat realistis menyulitkan proses verifikasi keaslian informasi, sehingga meningkatkan potensi penyalahgunaan, mulai dari manipulasi individu, penipuan identitas, hingga disinformasi berskala politik. Fenomena ini menunjukkan bahwa deephoax tidak hanya berdampak pada aspek teknis, tetapi juga pada stabilitas sosial dan kepercayaan publik terhadap media visual.

A Pemerasan dan Manipulasi Visual terhadap Individu

Pemanfaatan deephoax dalam bentuk manipulasi gambar atau video kompromitatif palsu telah digunakan sebagai alat pemerasan terhadap individu. Pelaku dapat meniru wajah korban dan menempatkannya dalam adegan sensitif atau tidak pantas, sehingga menimbulkan tekanan psikologis yang besar serta potensi

kerugian finansial bagi korban. Praktik ini dilaporkan meningkat seiring dengan kemudahan akses terhadap teknologi generatif yang canggih [26].

Gambar 2.2. Ilustrasi manipulasi wajah menggunakan deephoax untuk tujuan pemerasan.

Ancaman pemerasan berbasis deephoax menjadi semakin berbahaya karena korban sering kali kesulitan membuktikan bahwa konten tersebut bersifat palsu, terutama ketika kualitas visual yang dihasilkan sangat menyerupai citra fakta.

B Penipuan Identitas Visual dan Serangan terhadap Sistem Autentikasi

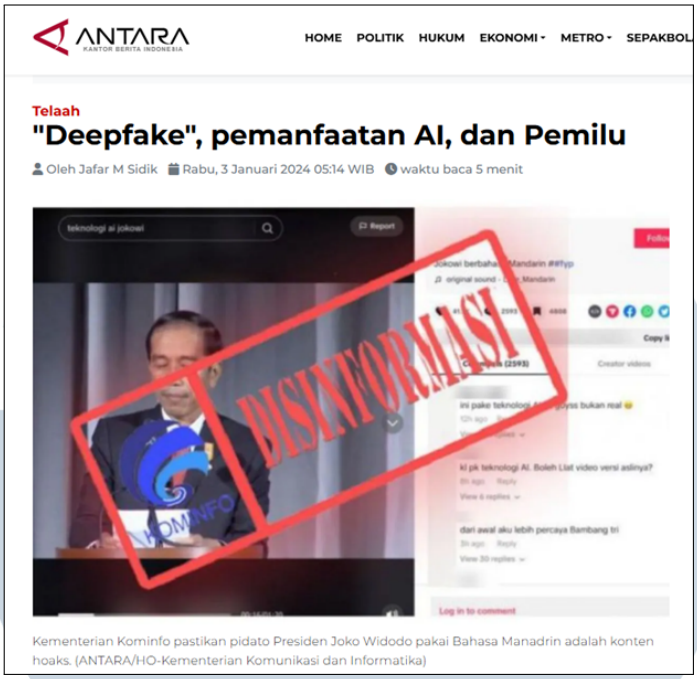
Deephoax juga dimanfaatkan untuk melakukan penipuan identitas visual dengan mengecoh sistem autentikasi berbasis biometrik, seperti verifikasi wajah. Konten visual palsu dapat digunakan untuk memperoleh akses ilegal ke layanan keuangan, akun pribadi, maupun sistem informasi sensitif lainnya. Ancaman ini menunjukkan celah keamanan pada sistem autentikasi yang terlalu bergantung pada satu modalitas biometrik [27].

Gambar 2.3. Contoh deephoax wajah atau video yang berpotensi menipu sistem autentikasi.

Kondisi tersebut menegaskan perlunya mekanisme keamanan berlapis *(multi-factor authentication)* untuk mengurangi risiko penyalahgunaan identitas berbasis konten visual sintetis.

C Disinformasi Politik melalui Manipulasi Video Tokoh Publik

Dalam konteks politik, deephoax video dapat digunakan untuk menciptakan rekaman palsu tokoh publik, seperti pidato atau pernyataan yang sebenarnya tidak pernah disampaikan. Konten manipulatif semacam ini berpotensi membentuk opini publik yang keliru dan memicu instabilitas sosial, terutama menjelang pemilihan umum [28].

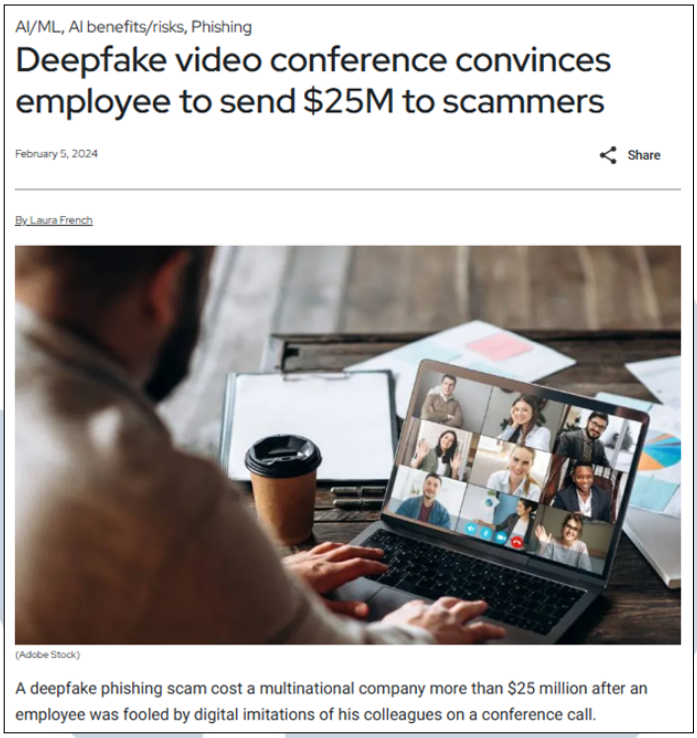


Gambar 2.4. Ilustrasi manipulasi video tokoh publik menggunakan deephoax yang berpotensi menimbulkan disinformasi politik.

Beberapa laporan media menunjukkan adanya video manipulatif yang menampilkan pidato palsu kepala negara, yang diedit menggunakan deephoax dan disebarkan sebagai bagian dari kampanye disinformasi selama periode pemilu [28].

D Penipuan Korporasi melalui Panggilan Video Deephoax

Pada lingkungan korporasi, deephoax dimanfaatkan dalam skema penipuan berbasis konferensi video, di mana pelaku menyamar sebagai eksekutif atau pejabat perusahaan. Melalui penyamaran visual dan suara yang meyakinkan, pelaku dapat memberikan instruksi palsu kepada staf, seperti permintaan transfer dana atau pengungkapan informasi rahasia perusahaan.



Gambar 2.5. Ilustrasi penipuan korporasi menggunakan konferensi video berbasis deephoax.

Kasus nyata menunjukkan bahwa korban yang mempercayai keaslian video tersebut dapat mengalami kerugian finansial dalam jumlah besar akibat instruksi palsu yang disampaikan melalui media deephoax [29].

2.2 Artificial Intelligence dan Deep Learning

Artificial Intelligence (AI) merupakan bidang ilmu komputer yang berfokus pada pengembangan sistem yang mampu meniru kemampuan kognitif manusia, termasuk dalam mengidentifikasi pola kompleks pada data visual untuk tugas pengenalan dan klasifikasi objek [30]. Deep Learning sebagai sub-bidang AI memanfaatkan jaringan saraf tiruan berlapis untuk mempelajari representasi data secara hierarkis, sehingga sangat efektif dalam pemrosesan citra dan analisis pola yang mendasari citra wajah asli maupun hasil manipulasi [31]. Dalam konteks deteksi deephoax, pendekatan deep learning, khususnya arsitektur *Convolutional Neural Networks* (CNN), mampu mengekstraksi fitur visual yang mendeskripsikan artefak manipulatif yang tidak mudah dikenali secara kasat mata oleh manusia [1, 32]. Deep learning juga memungkinkan sistem deteksi untuk belajar secara otomatis tanpa memerlukan ekstraksi fitur manual, sehingga menjadi dasar bagi

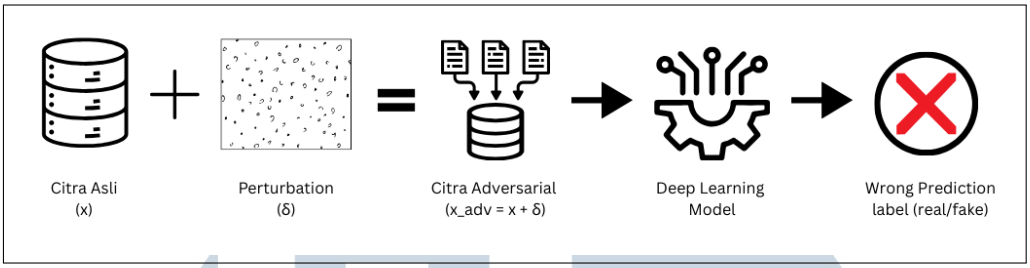
penelitian deteksi deephoax berbasis citra wajah.

Meskipun memiliki kemampuan representasi yang kuat, model deep learning diketahui rentan terhadap gangguan kecil yang terstruktur pada data masukan yang tidak memengaruhi persepsi visual manusia, yang dikenal sebagai serangan adversarial [8, 33]. Kerentanan ini memiliki dampak signifikan terhadap sistem deteksi deephoax, karena penyerang dapat menambahkan "noise" minimal yang sulit dideteksi secara visual namun cukup untuk mengecoh model klasifikasi [34]. Oleh karena itu, dalam penelitian ini pemahaman mengenai kerentanan model deep learning terhadap serangan adversarial menjadi aspek penting, terutama ketika model deteksi diuji pada kondisi nyata di mana data masukan mungkin telah dimanipulasi baik dengan teknik generatif maupun dengan teknik serangan yang dirancang khusus.

2.3 Serangan Adversarial dalam Deephoax Detection

Serangan adversarial merupakan ancaman signifikan dalam pengembangan sistem deteksi deephoax berbasis *deep learning*. Ancaman ini muncul karena model deteksi umumnya dilatih dengan asumsi bahwa data uji memiliki distribusi yang serupa dengan data latih. Pada kondisi nyata, asumsi tersebut dapat dilanggar ketika penyerang secara sengaja memanipulasi input untuk mengeksploitasi kerentanan model. Penelitian terdahulu menunjukkan bahwa model berbasis *convolutional neural networks* (CNN) rentan terhadap gangguan kecil yang terstruktur, meskipun gangguan tersebut tidak memengaruhi persepsi visual manusia [33, 34].

Dalam konteks keamanan multimedia, serangan adversarial memperluas ancaman deephoax dari sekadar manipulasi konten menjadi manipulasi terhadap sistem pendeteksi itu sendiri. Penyerang tidak hanya berupaya menghasilkan konten palsu yang realistis, tetapi juga mengoptimalkan perturbasi agar konten tersebut dapat melewati mekanisme deteksi otomatis. Oleh karena itu, evaluasi ketahanan (*robustness*) model terhadap serangan adversarial menjadi aspek penting dalam penelitian deephoax detection, khususnya terhadap serangan berbasis *white-box* seperti *Projected Gradient Descent* (PGD) yang sering digunakan sebagai skenario terburuk (*worst-case*) [8].



Gambar 2.6. Ilustrasi konsep serangan adversarial pada sistem deteksi deephoax.

Gambar 2.6 menggambarkan konsep dasar serangan adversarial pada sistem deteksi deephoax. Citra asli x dimodifikasi dengan menambahkan perturbasi kecil δ sehingga menghasilkan citra adversarial $x_{adv} = x + \delta$. Meskipun perubahan tersebut tidak mengubah persepsi visual manusia, citra adversarial dapat menyebabkan model *deep learning* menghasilkan prediksi yang salah. Konsep ini menunjukkan bagaimana manipulasi kecil pada data masukan dapat dimanfaatkan untuk mengecoh sistem deteksi otomatis [4].

2.3.1 Perbedaan Serangan Adversarial Berdasarkan Akses Penyerang

Serangan adversarial dapat diklasifikasikan berdasarkan tingkat akses yang dimiliki penyerang terhadap model target. Tingkat akses ini menentukan strategi serangan yang digunakan serta efektivitas perturbasi yang dihasilkan. Secara umum, serangan adversarial dibedakan menjadi serangan *white-box* dan *black-box*. Pemahaman mengenai perbedaan ini diperlukan untuk menempatkan posisi penelitian dalam konteks ancaman yang relevan, meskipun tidak seluruh jenis serangan dievaluasi secara eksperimental dalam penelitian ini.

A Serangan White-box

Serangan *white-box* merupakan skenario di mana penyerang memiliki akses penuh terhadap model target, termasuk arsitektur jaringan, parameter pembelajaran, serta fungsi kerugian (*loss function*). Dengan akses ini, penyerang dapat menghitung gradien fungsi loss terhadap input untuk membangun perturbasi yang secara optimal meningkatkan kesalahan prediksi model. Dalam literatur, serangan *white-box* sering digunakan sebagai tolok ukur utama dalam evaluasi robustness karena mencerminkan kondisi serangan terburuk [8].

Dalam penelitian ini, serangan *white-box* dipilih sebagai fokus evaluasi karena memberikan batas atas terhadap kerentanan model deteksi deephoax. Secara

khusus, metode *Projected Gradient Descent* (PGD) digunakan untuk menguji ketahanan model terhadap serangan adversarial yang bersifat iteratif dan kuat.

B Serangan Black-box

Serangan *black-box* merupakan skenario di mana penyerang tidak memiliki akses terhadap struktur internal maupun parameter model. Penyerang hanya dapat mengamati keluaran model berupa label prediksi atau skor kepercayaan. Dalam kondisi ini, serangan biasanya dilakukan melalui estimasi gradien berbasis *query*, pendekatan berbasis keputusan, atau transfer serangan dari model *surrogate* [35–37].

Meskipun serangan *black-box* relevan dalam skenario dunia nyata, pendekatan ini tidak dievaluasi dalam penelitian ini. Fokus penelitian diarahkan pada serangan *white-box* untuk memperoleh analisis ketahanan model yang lebih komprehensif terhadap ancaman adversarial.

2.3.2 Metode Serangan Adversarial

Berbagai metode serangan adversarial telah dikembangkan untuk mengevaluasi dan mengeksploitasi kerentanan model *deep learning*. Metode-metode ini berbeda dalam hal kompleksitas, efisiensi komputasi, serta tingkat keberhasilan dalam menurunkan performa model. Dalam konteks deteksi deephoax, metode serangan adversarial digunakan untuk mengukur sejauh mana model mampu mempertahankan performanya ketika dihadapkan pada input yang telah dimanipulasi secara adversarial.

Penelitian ini secara khusus memfokuskan analisis pada serangan *Projected Gradient Descent* (PGD), sementara metode lain disajikan sebagai referensi konseptual dalam landasan teori.

A Projected Gradient Descent (PGD)

Serangan *Projected Gradient Descent* (PGD) merupakan metode serangan adversarial bersifat iteratif yang dirancang untuk memaksimalkan nilai fungsi kerugian (*loss function*) model terhadap data masukan dengan tetap membatasi besar perturbasi yang diberikan pada ruang input [8]. PGD termasuk ke dalam kategori serangan *first-order* karena hanya memanfaatkan informasi gradien orde

pertama, dan secara luas digunakan sebagai *worst-case benchmark* dalam evaluasi ketahanan model terhadap serangan adversarial.

Secara konseptual, PGD dapat dipandang sebagai perluasan dari metode *Fast Gradient Sign Method* (FGSM) yang dilakukan secara berulang (*iterative*). Perbedaan utama PGD dibandingkan FGSM terletak pada penerapan mekanisme proyeksi pada setiap langkah pembaruan, yang bertujuan untuk memastikan bahwa perturbasi yang dihasilkan tetap berada dalam batas maksimum yang ditentukan oleh model ancaman (*threat model*).

Dalam formulasi aslinya, PGD umumnya diawali dengan inisialisasi contoh adversarial dari suatu titik awal di sekitar input asli, yang sering dilakukan secara acak di dalam wilayah perturbasi yang diperbolehkan. Pendekatan ini bertujuan untuk meningkatkan kekuatan serangan dengan memungkinkan proses optimisasi menjelajahi permukaan fungsi kerugian secara lebih luas dan menghindari solusi lokal yang lemah [8].

Pada setiap iterasi, PGD melakukan pembaruan contoh adversarial dengan menaikkan nilai fungsi kerugian model terhadap input melalui arah gradien. Proses pembaruan ini dirumuskan sebagai berikut:

$$x_{\text{adv}}^{t+1} = \Pi_{\mathscr{B}_\varepsilon(x)} \left(x_{\text{adv}}^{t} + \alpha \cdot \text{sign} \left(\nabla_x \mathscr{L} \left(f_\theta(x_{\text{adv}}^{t}), y \right) \right) \right) \tag{2.1}$$

di mana x_{adv}^{t} merupakan contoh adversarial pada iterasi ke-t, α adalah parameter *step size* yang mengontrol besar pembaruan pada setiap iterasi, $f_\theta(\cdot)$ menyatakan model dengan parameter θ, $\mathscr{L}(\cdot)$ adalah fungsi kerugian, dan y merupakan label kebenaran.

Gradien $\nabla_x \mathscr{L}(f_\theta(x_{\text{adv}}^{t}), y)$ merepresentasikan arah perubahan input yang paling meningkatkan nilai fungsi kerugian. Oleh karena itu, pembaruan dilakukan searah dengan tanda gradien (*gradient sign ascent*) untuk memaksimalkan *loss*. Operator proyeksi $\Pi_{\mathscr{B}_\varepsilon(x)}(\cdot)$ digunakan untuk memetakan kembali hasil pembaruan ke dalam himpunan $\mathscr{B}_\varepsilon(x)$, yaitu bola dengan jari-jari ε di sekitar input asli x, sehingga perturbasi yang dihasilkan tetap berada dalam batas yang diperbolehkan oleh model ancaman.

Dalam serangan PGD, himpunan $\mathscr{B}_\varepsilon(x)$ umumnya didefinisikan menggunakan norma ℓ_∞, yang membatasi besar perubahan maksimum pada setiap elemen input secara individual, sehingga perturbasi $\delta = x_{\text{adv}} - x$ memenuhi $\|\delta\|_\infty \leq \varepsilon$. Pembatasan ini memastikan bahwa setiap piksel hanya mengalami gangguan kecil yang terkontrol, sehingga struktur visual citra

tetap terjaga meskipun secara matematis gangguan tersebut dirancang untuk memaksimalkan fungsi kerugian model. Norma ℓ_∞ banyak digunakan sebagai model ancaman standar dalam evaluasi ketahanan model pembelajaran mendalam karena merepresentasikan skenario gangguan terburuk pada tingkat piksel (*pixel-wise worst-case perturbation*) dan secara luas diadopsi dalam penelitian serangan dan pertahanan adversarial berbasis gradien seperti FGSM dan PGD [8, 23].

Selama proses serangan PGD, parameter model θ dipertahankan tetap dan tidak mengalami pembaruan. Model hanya digunakan untuk menghitung nilai fungsi kerugian dan gradien terhadap input. Dengan mekanisme pembaruan iteratif dan proyeksi tersebut, PGD mampu menghasilkan contoh adversarial yang secara efektif mengeksploitasi kerentanan model tanpa melanggar batas perturbasi yang telah ditentukan [8].

B Metode Serangan Adversarial Lainnya

Selain PGD, terdapat beberapa metode serangan adversarial lain yang umum dibahas dalam literatur. Metode-metode ini tidak digunakan secara langsung dalam eksperimen penelitian ini, namun disertakan untuk memberikan gambaran umum mengenai spektrum serangan adversarial:

1. Fast Gradient Sign Method (FGSM), yaitu metode serangan satu-langkah yang memanfaatkan arah gradien fungsi loss untuk menghasilkan perturbasi secara cepat [38].
2. Carlini & Wagner (C&W), yaitu serangan berbasis optimasi yang bertujuan meminimalkan norma perturbasi sambil tetap menyebabkan misklasifikasi [34].
3. DeepFool, yaitu metode iteratif yang mendekati batas keputusan terdekat untuk menghasilkan contoh adversarial secara efisien [39].
4. AutoAttack, yaitu pendekatan *ensemble* yang mengombinasikan beberapa metode serangan untuk evaluasi robustness yang lebih andal dan terstandarisasi [40].

2.4 Pertahanan terhadap Serangan Adversarial

Pertahanan terhadap serangan adversarial bertujuan untuk meningkatkan ketahanan model pembelajaran mendalam terhadap gangguan kecil yang disengaja

pada data masukan. Berbeda dengan pendekatan pasif yang hanya berfokus pada evaluasi performa model, metode pertahanan bersifat proaktif dengan mengintegrasikan mekanisme ketahanan langsung ke dalam proses pelatihan. Pendekatan ini menjadi sangat penting dalam sistem keamanan siber berbasis kecerdasan buatan, termasuk sistem deteksi *deephoax*, karena serangan adversarial dapat secara signifikan menurunkan performa model meskipun perubahan input yang diberikan tidak mudah terdeteksi secara visual oleh manusia.

2.4.1 Adversarial Training (AT)

Adversarial Training (AT) merupakan salah satu metode pertahanan yang paling banyak diteliti dan dianggap sebagai pendekatan paling efektif secara empiris dalam meningkatkan ketahanan model terhadap serangan adversarial *white-box* [8, 38]. Pendekatan ini memformulasikan proses pelatihan sebagai masalah optimasi robust, di mana model tidak hanya dilatih menggunakan data bersih, tetapi juga menggunakan contoh *adversarial* yang dihasilkan secara eksplisit selama proses pelatihan.

Dalam AT, model secara sistematis dihadapkan pada skenario terburuk melalui proses *inner maximization*, yang bertujuan mencari perturbasi adversarial yang memaksimalkan nilai fungsi kerugian. Dengan demikian, model dipaksa untuk mempelajari representasi fitur yang stabil dan tidak sensitif terhadap gangguan kecil yang terstruktur pada data masukan.

Secara formal, prinsip *adversarial training* dirumuskan sebagai masalah optimasi minimax dua tingkat sebagai berikut:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathscr{D}} \left[\max_{\delta\in\mathscr{B}_{\varepsilon}} \mathscr{L}(f_{\theta}(x+\delta), y) \right] \tag{2.2}$$

Pada persamaan tersebut, proses *inner maximization* bertujuan menghasilkan perturbasi adversarial δ dalam batas ε yang memaksimalkan nilai *loss function*, sedangkan proses *outer minimization* menyesuaikan parameter model θ agar meminimalkan kesalahan prediksi pada kondisi terburuk tersebut. Dalam implementasi praktis, serangan *Projected Gradient Descent* (PGD) umum digunakan untuk menyelesaikan proses *inner maximization* karena kekuatannya dalam mengeksplorasi ruang perturbasi [8].

Pendekatan ini menjadikan *adversarial training* berbasis PGD sebagai standar de facto dalam evaluasi dan pengembangan model pembelajaran mendalam

yang robust terhadap serangan adversarial, termasuk dalam konteks sistem deteksi *deephoax*.

2.5 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) merupakan salah satu arsitektur jaringan saraf dalam (*deep neural network*) yang dirancang untuk memproses data berbentuk grid, seperti citra digital. CNN bekerja dengan mengekstraksi pola spasial melalui operasi konvolusi menggunakan kernel berukuran kecil yang digeser secara sistematis pada citra input [41]. Operasi ini memungkinkan jaringan untuk menangkap fitur lokal seperti tepi, sudut, dan tekstur, serta membangun representasi visual tingkat tinggi secara hierarkis pada lapisan yang lebih dalam.

Secara umum, CNN tersusun atas beberapa komponen utama, yaitu *convolutional layer*, *activation function*, *pooling layer*, dan *fully connected layer*. Lapisan konvolusi berfungsi sebagai ekstraktor fitur utama, sementara fungsi aktivasi non-linear seperti ReLU digunakan untuk meningkatkan kemampuan representasi jaringan. Lapisan pooling berperan dalam mereduksi dimensi spasial dan meningkatkan ketahanan terhadap variasi posisi fitur. Pada tahap akhir, lapisan *fully connected* digunakan untuk memetakan fitur hasil ekstraksi ke dalam ruang kelas keluaran.

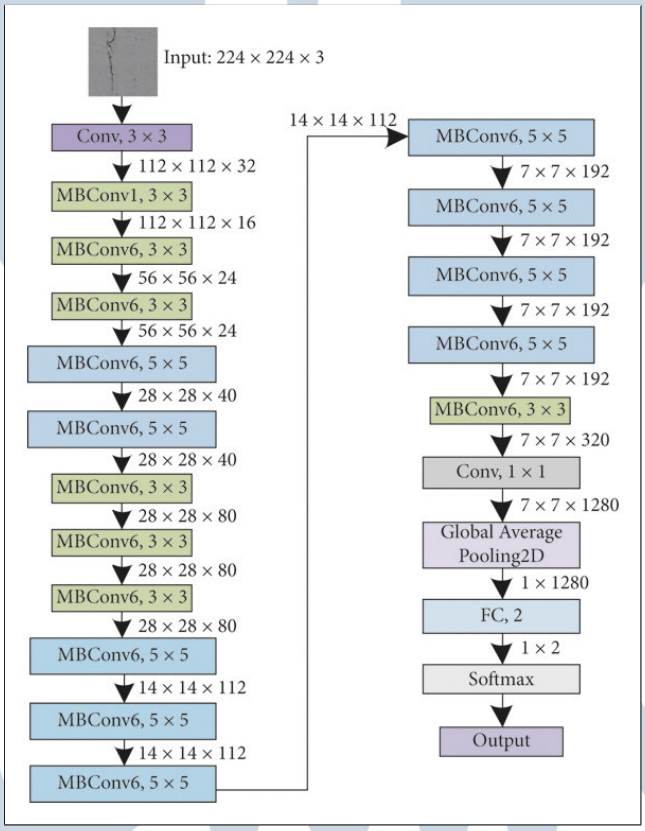
Dalam konteks analisis citra wajah, termasuk klasifikasi citra wajah hasil manipulasi atau *deephoax*, CNN digunakan untuk mengidentifikasi inkonsistensi visual halus yang sulit dikenali secara manual, seperti artefak tekstur, distorsi struktur wajah, serta ketidaksesuaian pola pencahayaan [42]. Namun, peningkatan kompleksitas CNN secara konvensional melalui penambahan kedalaman atau lebar jaringan sering kali berdampak pada meningkatnya jumlah parameter dan biaya komputasi. Kondisi ini mendorong pengembangan arsitektur CNN yang lebih efisien, salah satunya adalah EfficientNet.

2.5.1 Arsitektur EfficientNet-B0

EfficientNet merupakan keluarga arsitektur CNN yang dirancang melalui pendekatan *neural architecture search* (NAS) dan strategi *compound scaling* [43]. Strategi ini menskalakan tiga dimensi utama jaringan secara bersamaan, yaitu kedalaman jaringan (*depth*), lebar jaringan (*width*), dan resolusi input, menggunakan koefisien skala yang terkontrol. Pendekatan tersebut bertujuan untuk

mempertahankan keseimbangan antara kapasitas representasi jaringan dan efisiensi komputasi.

EfficientNet-B0 merupakan model dasar dari keluarga EfficientNet yang diperoleh langsung dari proses NAS. Arsitektur ini tersusun atas serangkaian blok *Mobile Inverted Bottleneck Convolution* (MBConv) yang dikombinasikan dengan *depthwise separable convolution* dan mekanisme *Squeeze-and-Excitation* (SE). Kombinasi tersebut dirancang untuk meningkatkan efisiensi parameter sekaligus mempertahankan kemampuan ekstraksi fitur spasial.



Gambar 2.7. Struktur arsitektur EfficientNet-B0

Blok MBConv merupakan pengembangan dari konsep *inverted bottleneck* yang diperkenalkan pada MobileNetV2. Setiap blok MBConv terdiri atas tiga tahap utama, yaitu *pointwise convolution* untuk ekspansi jumlah kanal, *depthwise convolution* untuk ekstraksi fitur spasial, dan *pointwise convolution* untuk proyeksi kembali ke dimensi kanal yang lebih kecil. Selain itu, mekanisme *Squeeze-and-Excitation* digunakan untuk melakukan penyesuaian bobot antar kanal berdasarkan informasi global fitur [44].

Berdasarkan Gambar 2.7 [45], proses ekstraksi fitur pada EfficientNet-B0

dimulai dari citra input berukuran $224 \times 224 \times 3$. Lapisan konvolusi awal berukuran 3×3 menghasilkan peta fitur berdimensi $112 \times 112 \times 32$. Selanjutnya, fitur tersebut diproses melalui beberapa blok MBConv dengan variasi kernel 3×3 dan 5×5, yang secara bertahap menurunkan resolusi spasial menjadi 112×112, 56×56, 28×28, 14×14, hingga 7×7, sembari meningkatkan jumlah kanal hingga 320.

Pada tahap akhir ekstraksi fitur, peta fitur berukuran $7 \times 7 \times 320$ dilewatkan ke lapisan konvolusi 1×1 untuk meningkatkan dimensi kanal menjadi 1280. Operasi *Global Average Pooling* kemudian digunakan untuk mereduksi dimensi spasial dan menghasilkan vektor fitur berdimensi 1×1280. Vektor fitur ini selanjutnya dipetakan ke dalam dua kelas keluaran melalui lapisan *fully connected*, yang merepresentasikan kelas citra *fakta* dan *hoax*, sebelum diaktifkan menggunakan fungsi *softmax*.

2.5.2 Metrik Evaluasi

Evaluasi performa pada penelitian ini bertujuan untuk mengukur dua aspek utama, yaitu kemampuan model dalam melakukan klasifikasi pada data bersih serta tingkat ketahanannya terhadap serangan adversarial berbasis *Projected Gradient Descent* (PGD). Untuk mencapai tujuan tersebut, digunakan beberapa metrik evaluasi yang umum diterapkan dalam penelitian adversarial machine learning, meliputi *clean accuracy*, *PGD accuracy*, *precision*, *recall*, dan *F1-score*. Kombinasi metrik ini memungkinkan analisis yang komprehensif terhadap trade-off antara akurasi alami dan ketahanan model terhadap serangan adversarial [8, 46].

A Clean Accuracy

Clean accuracy digunakan untuk mengukur tingkat keberhasilan model dalam melakukan klasifikasi pada data uji tanpa adanya gangguan adversarial. Metrik ini merepresentasikan performa alami (*natural accuracy*) dari model dalam kondisi ideal, yaitu ketika input yang diberikan identik dengan distribusi data asli [47, 48].

Secara matematis, *clean accuracy* dapat dihitung menggunakan konsep True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) sebagai berikut:

$$\text{Clean Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

di mana:

1. TP = jumlah sampel positif yang diklasifikasikan dengan benar
2. TN = jumlah sampel negatif yang diklasifikasikan dengan benar
3. FP = jumlah sampel negatif yang salah diklasifikasikan sebagai positif
4. FN = jumlah sampel positif yang salah diklasifikasikan sebagai negatif

B PGD Accuracy

PGD accuracy digunakan untuk mengukur tingkat akurasi model ketika diuji menggunakan sampel yang telah dimodifikasi oleh serangan adversarial berbasis *Projected Gradient Descent* (PGD). Metrik ini merepresentasikan kemampuan model dalam mempertahankan performa klasifikasi di bawah skenario serangan *white-box* iteratif [8, 49].

Secara matematis, *PGD accuracy* dapat dirumuskan dengan analogi yang sama:

$$\text{PGD Accuracy} = \frac{TP_{adv} + TN_{adv}}{TP_{adv} + TN_{adv} + FP_{adv} + FN_{adv}} \tag{2.4}$$

di mana TP_{adv}, TN_{adv}, FP_{adv}, dan FN_{adv} dihitung berdasarkan prediksi model terhadap sampel adversarial.

C Precision, Recall, dan F1-Score

Selain akurasi, penelitian ini juga menggunakan metrik *precision*, *recall*, dan *F1-score* untuk menilai kualitas prediksi model secara lebih detail, khususnya pada kasus ketidakseimbangan data.

1. **Precision** mengukur seberapa tepat model dalam memprediksi kelas positif:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.5}$$

2. **Recall** mengukur kemampuan model dalam mendeteksi semua sampel positif:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.6}$$

3. **F1-score** merupakan harmonisasi antara *precision* dan *recall*, sehingga memberikan ukuran kinerja yang seimbang:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.7}$$

Penggunaan metrik ini memungkinkan evaluasi performa model tidak hanya dari akurasi secara keseluruhan, tetapi juga dari kemampuan model dalam mendeteksi kelas minoritas dan menghindari kesalahan prediksi yang kritis.

D Attack Success Rate

Attack Success Rate (ASR) digunakan untuk mengukur tingkat keberhasilan serangan adversarial dalam menyebabkan kesalahan klasifikasi pada model. Dalam penelitian ini, ASR didefinisikan secara langsung sebagai komplemen dari *PGD accuracy*, sehingga merepresentasikan proporsi sampel uji yang gagal diklasifikasikan dengan benar setelah dikenai serangan PGD [8, 49].

Pendefinisian ASR sebagai komplemen dari *PGD accuracy* umum digunakan dalam evaluasi ketahanan adversarial karena memberikan interpretasi yang lebih sederhana dan konsisten, terutama ketika serangan bersifat *untargeted*. Secara matematis, ASR dirumuskan sebagai:

$$\text{ASR} = 1 - \text{PGD Accuracy} \tag{2.8}$$

dengan *PGD Accuracy* menyatakan tingkat akurasi model pada data uji yang telah dimodifikasi oleh serangan PGD. Nilai ASR yang tinggi mengindikasikan bahwa serangan adversarial berhasil menurunkan performa model secara signifikan, sedangkan nilai ASR yang rendah menunjukkan bahwa model memiliki tingkat ketahanan yang lebih baik terhadap serangan adversarial.

2.6 Studi Literatur Terkait

Studi literatur ini membahas penelitian-penelitian terkait penerapan *Adversarial Training* (AT) sebagai mekanisme pertahanan utama terhadap serangan *Projected Gradient Descent* (PGD). Fokus kajian diarahkan pada pendekatan AT berbasis PGD yang telah diakui secara luas sebagai standar de facto dalam meningkatkan ketahanan model pembelajaran mendalam terhadap serangan *adversarial* pada skenario *white-box*.

Tabel 2.1. Studi Literatur Terkait Adversarial Training terhadap Serangan PGD

No	Peneliti & Tahun	Model	Attack	ϵ	Dataset	Akurasi	Advantages	Disadvantages
1	Madry et al. (2018) [8]	ResNet-50	PGD20	8/255	CIFAR-10	45.80%	Baseline adversarial training yang kuat dan formal	Biaya komputasi tinggi; robust accuracy terbatas
2	Qin et al. (2019) [50]	ResNet-152	PGD50	4/255	ImageNet	47.00%	Pendekatan local linearization yang stabil	Peningkatan robustness masih terbatas
3	Mao et al. (2019) [51]	Wide ResNet	PGD20	8/255	CIFAR-10	50.03%	Metric learning membantu separasi fitur	Tidak melampaui batas robustness umum
4	Zhang et al. (2020) [52]	Wide ResNet	PGD20	16/255	CIFAR-10	49.86%	Curriculum memperhalus proses training	Robust accuracy stagnan
5	Wang et al. (2019) [53]	8-Layer ConvNet	PGD20	8/255	CIFAR-10	42.40%	Model ringan dan efisien	Kapasitas model rendah
6	Pang et al. (2019) [54]	Wide ResNet	PGD10	0.005	CIFAR-100	32.10%	Diversitas ensemble meningkat	Overhead tinggi; robustness rendah
7	Kariyappa and Qureshi (2019) [55]	ResNet-20	PGD30	0.09/1	CIFAR-10	46.30%	Mengurangi gradient masking	Masih gagal melawan PGD kuat
8	Ding et al. (2020) [56]	Wide ResNet	PGD100	8/255	CIFAR-10	47.18%	Margin-based adversarial training (MMA)	Robust accuracy tetap terbatas
9	Shafahi et al. (2019) [48]	Wide ResNet	PGD100	8/255	CIFAR-10	46.19%	Efisiensi tinggi melalui Free AT	Rentan catastrophic overfitting
10	Lee et al. (2020) [57]	PreActResNet-18	PGD20	8/255	Tiny ImageNet	20.31%	Evaluasi pada dataset kompleks	Robustness sangat rendah

Berdasarkan kajian literatur yang telah diuraikan, *adversarial training* berbasis PGD terbukti sebagai pendekatan pertahanan yang paling konsisten dan efektif dalam meningkatkan ketahanan model terhadap serangan adversarial *white-box*. Namun, sebagian besar penelitian masih berfokus pada dataset benchmark umum seperti CIFAR-10 dan CIFAR-100, serta belum banyak mengkaji penerapan *adversarial training* pada tugas deteksi *deephoax* berbasis citra wajah.

Selain itu, keterbatasan sumber daya komputasi akibat sifat iteratif serangan PGD menjadikan pemilihan arsitektur yang efisien sebagai aspek penting yang belum banyak dieksplorasi. Oleh karena itu, penelitian ini berfokus pada penerapan *adversarial training* berbasis PGD pada arsitektur EfficientNet-B0 sebagai upaya untuk menjembatani kesenjangan penelitian antara ketahanan model, efisiensi komputasi, dan konteks deteksi *deephoax* berbasis citra wajah.