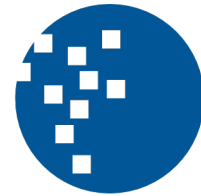


**DETEKSI DEEPHOAX TERHADAP SERANGAN
ADVERSARIAL DENGAN MODEL EFFICIENTNET-B0
UNTUK KEAMANAN SIBER**



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

SKRIPSI

ANDREW THOMAS AGUSTINUS
00000059999

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2026

**DETEKSI DEEPHOAX TERHADAP SERANGAN
ADVERSARIAL DENGAN MODEL EFFICIENTNET-B0
UNTUK KEAMANAN SIBER**



SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

**ANDREW THOMAS AGUSTINUS
00000059999**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2026**

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Andrew Thomas Agustinus

Nomor Induk Mahasiswa : 00000059999

Program Studi : Informatika

Skripsi dengan judul:

Deteksi Deepfoax terhadap Serangan Adversarial dengan model EfficientNet-B0 untuk Keamanan Siber

merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 8 Januari 2026



(Andrew Thomas Agustinus)

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PENGESAHAN

Skripsi dengan judul

DETEKSI DEEPHOAX TERHADAP SERANGAN ADVERSARIAL DENGAN MODEL EFFICIENTNET-B0 UNTUK KEAMANAN SIBER

Oleh

Nama : Andrew Thomas Agustinus
NIM : 00000059999
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

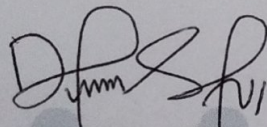
Telah diujikan pada hari Kamis, 15 Januari 2026

Pukul 10.00 s.d 12.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

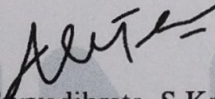
Ketua Sidang



(David Agustriawan, S.Kom., M.Sc.,
Ph.D.)

NIDN: 0525088601

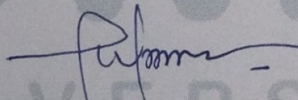
Penguji



(Alethea Suryadibrata, S.Kom., M.Eng)

NIDN: 322099201

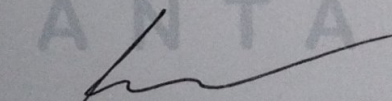
Pembimbing



(SY. Yuliani, S.Kom., MT., Ph.D., CEH)

NIDN: 0411037904

Ketua Program Studi Informatika,



(Arya Wicaksana, S.Kom., M.Eng.Sc., OCA)

NIDN: 0315109103

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini:

Nama : Andrew Thomas Agustinus
NIM : 00000059999
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : Deteksi Deephoax terhadap Serangan Adversarial dengan model EfficientNet-B0 untuk Keamanan Siber

Menyatakan dengan sesungguhnya bahwa saya bersedia (**pilih salah satu**):

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- ☐ Lainnya, pilih salah satu:
 - ☐ Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - ☐ Embargo publikasi karya ilmiah dalam kurun waktu tiga tahun.

Tangerang, 8 Januari 2026

Yang menyatakan

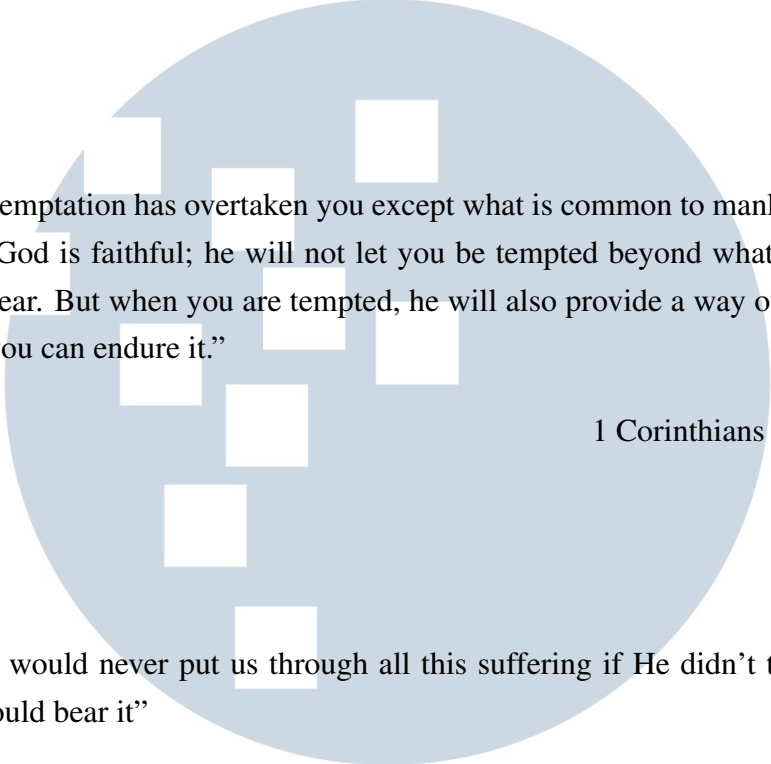


(Andrew Thomas Agustinus)

*Centang salah satu tanpa menghapus opsi yang tidak dipilih

**Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

HALAMAN PERSEMBAHAN / MOTTO



"No temptation has overtaken you except what is common to mankind. And God is faithful; he will not let you be tempted beyond what you can bear. But when you are tempted, he will also provide a way out so that you can endure it."

1 Corinthians 10:13 NIV

"God would never put us through all this suffering if He didn't think we could bear it"

Konno Yuuki (Zekken)

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

KATA PENGANTAR

Puji Syukur atas selesainya penulisan Laporan Skripsi ini dengan judul: "Deteksi Deepfake terhadap Serangan Adversarial dengan model EfficientNet-B0 untuk Keamanan Siber" dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan tugas akhir ini, sangatlah sulit bagi saya untuk menyelesaikan tugas akhir ini. Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Bapak Dr. Ir. Andrey Andoko, M.Sc., selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Arya Wicaksana, S.Kom., M.Eng.Sc., OCA, selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Ibu SY.Yuliani., S.Kom., MT., Ph.D., CEH, sebagai Pembimbing pertama yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Orang Tua dan keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan laporan penelitian ini.

Semoga karya ilmiah ini dapat memberikan manfaat yang baik sebagai sumber informasi maupun inspirasi seluruh kalangan.

Tangerang, 8 Januari 2026



Andrew Thomas Agustinus

DETEKSI DEEPHOAX TERHADAP SERANGAN ADVERSARIAL DENGAN MODEL EFFICIENTNET-B0 UNTUK KEAMANAN SIBER

Andrew Thomas Agustinus

ABSTRAK

Perkembangan pesat teknologi *deephoax* memungkinkan manipulasi wajah manusia secara realistis, namun penggunaannya semakin meluas untuk disinformasi dan ancaman keamanan siber. Sistem deteksi *deephoax* berbasis *deep learning* masih rentan terhadap serangan *adversarial*, khususnya metode *Projected Gradient Descent* (PGD), yang dapat menurunkan akurasi model secara signifikan. Penelitian ini mengkaji penerapan *adversarial training* (AT) berbasis PGD pada arsitektur EfficientNet-B0 untuk meningkatkan ketahanan detektor terhadap serangan *adversarial* sekaligus mempertahankan efisiensi komputasi. Metodologi meliputi pembangunan model deteksi, pra-pemrosesan citra wajah, serta pelatihan menggunakan tiga skenario: Baseline, Adversarial Training, dan Mixed AT. Evaluasi dilakukan pada data bersih serta data yang dimodifikasi menggunakan PGD dengan iterasi 5, 10, dan 20. Hasil menunjukkan model Baseline memiliki *clean accuracy* 98,7% namun akurasinya 0% terhadap serangan PGD. Model AT mempertahankan akurasi *adversarial* sekitar 49,7%. Sementara itu, Mixed AT menunjukkan performa seimbang dengan *clean accuracy* 95,2% dan akurasi PGD 50,4 49,5% (PGD-5/10) serta 32,2% (PGD-20). Penelitian menegaskan bahwa *adversarial training* efektif meningkatkan robustness, dan Mixed AT menyediakan keseimbangan terbaik antara akurasi dan ketahanan terhadap serangan.

Kata kunci: *adversarial training, deephoax, EfficientNet-B0, Projected Gradient Descent* (PGD)

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

DEEPHOAX DETECTION AGAINST ADVERSARIAL ATTACKS USING EFFICIENTNET-B0 MODEL FOR CYBERSECURITY

Andrew Thomas Agustinus

ABSTRACT

The rapid development of deepphoax technology enables realistic manipulation of human faces, yet its usage has increasingly been exploited for disinformation and cyber security threats. Deep learning-based deepphoax detection systems remain vulnerable to adversarial attacks, particularly the Projected Gradient Descent (PGD) method, which can significantly reduce model accuracy. This study investigates the application of PGD-based adversarial training (AT) on the EfficientNet-B0 architecture to enhance detector robustness against adversarial attacks while maintaining computational efficiency. The methodology includes model development, facial image preprocessing, and training using three scenarios: Baseline, Adversarial Training, and Mixed AT. Evaluation was conducted on clean data as well as data modified with PGD at 5, 10, and 20 iterations. Results show that the Baseline model achieved 98.7% clean accuracy but 0% accuracy under PGD attacks. The AT model maintained adversarial accuracy around 49.7%. Meanwhile, the Mixed AT model demonstrated a balanced performance, with 95.2% clean accuracy and PGD accuracy of 50.4%, 49.5% (PGD-5/10) and 32.2% (PGD-20). The study confirms that adversarial training effectively enhances robustness, and Mixed AT provides the best trade-off between accuracy and resilience against attacks.

Keywords: *adversarial training, deepphoax, EfficientNet-B0, Projected Gradient Descent (PGD)*



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR	vi
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xiii
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Permasalahan	4
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	5
1.6 Sistematika Penulisan	5
BAB 2 LANDASAN TEORI	7
2.1 Keamanan Siber (Cyber Security)	7
2.1.1 Ancaman dan Serangan dalam Keamanan Siber	7
2.1.2 Peran AI dan Machine Learning dalam Keamanan Siber	8
2.1.3 Konsep Dasar Deepfoax dan Perbedaannya dengan Deepfake	9
2.1.4 Dampak Deepfoax dalam Keamanan Siber	10
2.2 Artificial Intelligence dan Deep Learning	14
2.3 Serangan Adversarial dalam Deepfoax Detection	15
2.3.1 Perbedaan Serangan Adversarial Berdasarkan Akses Penyerang	16
2.3.2 Metode Serangan Adversarial	17
2.4 Pertahanan terhadap Serangan Adversarial	19
2.4.1 Adversarial Training (AT)	20
2.5 Convolutional Neural Network (CNN)	21
2.5.1 Arsitektur EfficientNet-B0	21
2.5.2 Metrik Evaluasi	23
2.6 Studi Literatur Terkait	25
BAB 3 METODOLOGI PENELITIAN	27
3.1 Diagram Metodologi Penelitian untuk Deteksi Deepfoax	27
3.1.1 Pengumpulan Data	28
3.1.2 Pra-Pemrosesan Data	29
3.1.3 Konfigurasi Pelatihan dan Serangan Adversarial	30
3.1.4 Skenario Pelatihan	31
3.1.5 Pembangunan Arsitektur Model	33
3.2 Perhitungan Model Architecture	35
3.2.1 Perhitungan Lapisan Konvolusi (Conv Stem dan MBConv)	35
3.2.2 Contoh Proses Konvolusi dan Pembentukan Feature Map	36
3.2.3 Perhitungan Global Average Pooling	40
3.2.4 Perhitungan Lapisan Fully Connected untuk Dua Kelas	41

3.2.5	Perhitungan Softmax dan Prediksi Kelas	42
3.2.6	Implementasi Adversarial Attack	42
3.2.7	<i>Pipeline</i> Proses <i>Training</i>	44
3.2.8	Evaluasi Model	46
3.2.9	Pustaka yang Digunakan	48
BAB 4	HASIL DAN DISKUSI	50
4.1	Spesifikasi Sistem	50
4.1.1	Perangkat Keras	50
4.1.2	Perangkat Lunak	50
4.2	Preprocessing dan Penyiapan Dataset	51
4.3	Arsitektur Model	52
4.3.1	EfficientNet Detector	52
4.4	Implementasi Adversarial Attack	53
4.4.1	Projected Gradient Descent (PGD)	53
4.4.2	Visualisasi Perturbasi Adversarial PGD	54
4.5	Implementasi Pelatihan dan Validasi Model	55
4.6	Evaluasi dan Testing Model	57
4.7	Hasil Training dan Validasi	58
4.7.1	Performa Training dan Validation Accuracy	58
4.7.2	Performa Training dan Validation Loss	59
4.7.3	Ringkasan Performa Model Terbaik	60
4.8	Hasil Evaluasi Komprehensif	61
4.8.1	Performa Pengujian pada Data Bersih dan Data Adversarial	61
4.8.2	Tabel Hasil <i>Testing Accuracy</i>	62
4.8.3	Hasil <i>Attack Success Rate</i> (ASR)	63
4.9	Hasil Confusion Matrix dan Perhitungan Metrik	65
4.9.1	Baseline Model	66
4.9.2	Adversarial Training (AT) Model	68
4.9.3	Mixed Adversarial Training (Mixed AT) Model	69
4.10	Diskusi	71
4.10.1	Analisis Performa Pelatihan Berdasarkan Akurasi dan Loss	71
4.10.2	Analisis Performa Pengujian Model	73
4.10.3	Analisis Trade-off antara Akurasi Bersih dan Robustness	73
4.10.4	Analisis Confusion Matrix	74
BAB 5	SIMPULAN DAN SARAN	76
5.1	Simpulan	76
5.2	Saran	77
DAFTAR PUSTAKA		78

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR TABEL

Tabel 2.1	Studi Literatur Terkait Adversarial Training terhadap Serangan PGD	26
Tabel 3.1	Komposisi dataset DeepDetect-2025 yang digunakan dalam penelitian	28
Tabel 3.2	Pembagian dataset DeepDetect-2025 pada tahap pra-pemrosesan	29
Tabel 3.3	Konfigurasi Pelatihan dan Serangan Adversarial	31
Tabel 3.4	Skenario Pelatihan dan Penggunaan Sampel Adversarial . .	32
Tabel 3.5	Ringkasan arsitektur model EfficientNet-B0 untuk deteksi deephoax	34
Tabel 4.1	Ringkasan Performa Model Terbaik Berdasarkan Validation Loss	60
Tabel 4.2	Hasil <i>Testing Accuracy</i> pada Data Bersih dan Serangan PGD (%)	62

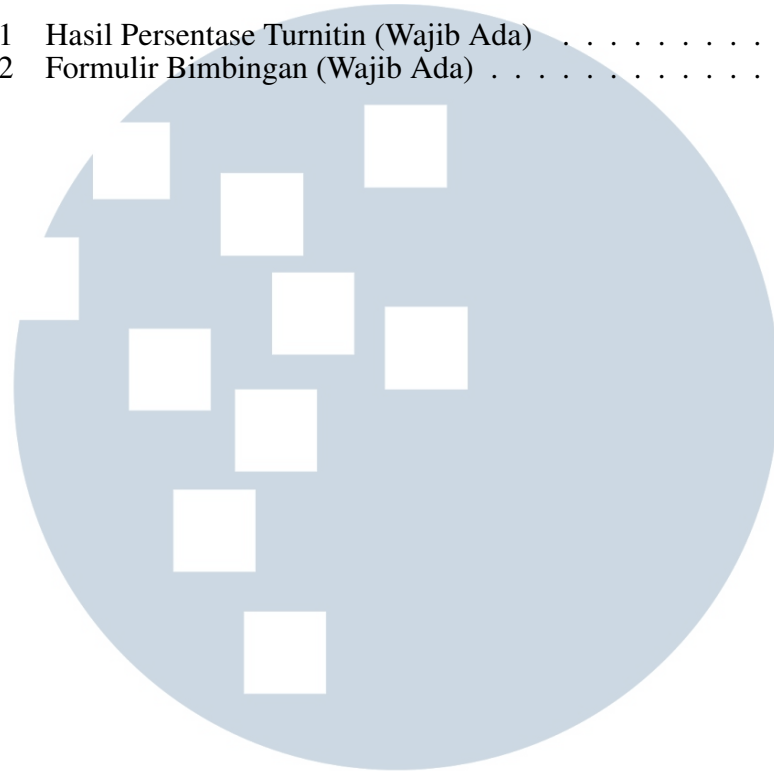


DAFTAR GAMBAR

Gambar 1.1	Tren peningkatan jumlah konten <i>deephoax</i> secara global pada tahun 2020–2025.	1
Gambar 2.1	Perbandingan citra wajah fakta dan citra wajah hoax hasil manipulasi berbasis <i>deep learning</i>	10
Gambar 2.2	Ilustrasi manipulasi wajah menggunakan <i>deephoax</i> untuk tujuan pemerasan.	11
Gambar 2.3	Contoh <i>deephoax</i> wajah atau video yang berpotensi menipu sistem autentikasi.	12
Gambar 2.4	Ilustrasi manipulasi video tokoh publik menggunakan <i>deephoax</i> yang berpotensi menimbulkan disinformasi politik.	13
Gambar 2.5	Ilustrasi penipuan korporasi menggunakan konferensi video berbasis <i>deephoax</i>	14
Gambar 2.6	Ilustrasi konsep serangan adversarial pada sistem deteksi <i>deephoax</i>	16
Gambar 2.7	Struktur arsitektur EfficientNet-B0	22
Gambar 3.1	Diagram metodologi penelitian untuk deteksi <i>deephoax</i>	27
Gambar 3.2	Diagram pipeline pra-pemrosesan data	29
Gambar 3.3	Diagram pipeline pembangunan arsitektur model	33
Gambar 3.4	Diagram Pipeline Implementasi Adversarial Attack menggunakan Projected Gradient Descent (PGD)	43
Gambar 3.5	Diagram pipeline proses <i>Training</i>	45
Gambar 3.6	Diagram pipeline proses evaluasi model	47
Gambar 4.1	Contoh transformasi data.	52
Gambar 4.2	Visualisasi perturbasi adversarial PGD.	55
Gambar 4.3	Perbandingan Training Accuracy dan Validation Accuracy per Epoch	58
Gambar 4.4	Perbandingan Training Loss dan Validation Loss per Epoch	59
Gambar 4.5	Perbandingan <i>Testing Accuracy</i> pada Data Bersih dan Serangan PGD	61
Gambar 4.6	Attack Success Rate (ASR) pada Serangan PGD	63
Gambar 4.7	Baseline Model – Confusion Matrices: <i>Clean</i> , PGD-5, PGD-10, PGD-20	66
Gambar 4.8	AT Model – Confusion Matrices: <i>Clean</i> , PGD-5, PGD-10, PGD-20	68
Gambar 4.9	Mixed AT Model – Confusion Matrices: <i>Clean</i> , PGD-5, PGD-10, PGD-20	69

DAFTAR LAMPIRAN

Lampiran 1	Hasil Persentase Turnitin (Wajib Ada)	86
Lampiran 2	Formulir Bimbingan (Wajib Ada)	89



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA