KOMPARASI SERANGAN *ADVERSARIAL* PADA MODEL *RESNET34* UNTUK TUGAS KLASIFIKASI GAMBAR *IMAGENET DATASET*

TUGAS AKHIR

Diajukan Oleh:
AIDIL NAJAR
190705023
Mahasiswa Fakultas Sains Dan Teknologi
Program Studi Teknologi Informasi

FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI AR-RANIRY
BANDA ACEH
2025 M/1446 H

LEMBAR PERSETUJUAN

KOMPARASI SERANGAN *ADVERSARIAL* PADA MODEL *RESNET34* UNTUK TUGAS KLASIFIKASI GAMBAR *IMAGENET DATASET*

TUGAS AKHIR

Diajukan kepada Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Ar-Raniry Banda Aceh
Sebagai Salah Satu Persyaratan Penulisan Tugas Akhir
Dalam Prodi Teknologi Informasi

Oleh :

AIDIL NAJAR
190705023

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi

Disetujui Untuk Dimunaqasyahkan Oleh :

Pembimbing I,	Pembimbing II
Dr. Hendri Ahmadian, S.Si., M.I.M	**Mulkan Fadhli, S.T., M.T**
NIP. 198301042014031002	**NIP. 198811282020121006**

Mengetahui,
Ketua Program Studi Teknologi Informasi

Malahayati, M.T.
NIP. 198301272015032003

LEMBAR PENGESAHAN

KOMPARASI SERANGAN *ADVERSARIAL* PADA MODEL *RESNET34* UNTUK TUGAS KLASIFIKASI GAMBAR *IMAGENET DATASET*

TUGAS AKHIR

Telah Diuji Oleh Panitia Ujian Munaqasyah Tugas Akhir
Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh dan Dinyatakan Lulus
Serta Diterima Sebagai Salah Satu Beban Studi Program Sarjana (S1)
Dalam Program Studi Teknologi Informasi

Pada Hari/Tanggal: Rabu, 02 Juli 2025
7 Muharram 1447 H
Di Darussalam, Banda Aceh

Panitia Ujian Munaqasyah Tugas Akhir:

Ketua,	Sekretaris,
Dr. Hendri Ahmadian, S.Si., M.I.M	**Mulkan Fadhli, S.T., M.T**
NIP. 198301042014031002	**NIP. 198811282020121006**
Penguji I,	Penguji II,
Malahayati, M.T.	**Khairan AR, M.Kom.**
NIP. 198301272015032003	**NIP. 198607042014031001**

Mengetahui:
Dekan Fakultas Sains dan Teknologi
UIN Ar-Raniry Banda Aceh

Prof. Dr. Ir. Muhammad Dirhamsyah, M.T., IPU
NIP. 19620021988111001

LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini:

Nama	: Aidil Najar
NIM	: 190705023
Program Studi	: Teknologi Informasi
Fakultas	: Sains dan Teknologi
Judul	: Komparasi Serangan Adversarial Pada Model Resnet34 Untuk Tugas Klasifikasi Gambar Imagenet Dataset

Dengan ini menyatakan bahwa dalam penulisan tugas akhir ini, Saya:

1. Tidak menggunakan ide orang lain tanpa mampu mengembangkan dan mempertanggungjawabkan;
2. Tidak melakukan plagiasi terhadap naskah karya orang lain;
3. Tidak menggunakan karya orang lain tanpa menyebutkan sumber asli atau tanpa izin pemilik karya;
4. Tidak memanipulasi dan memalsukan data;
5. Mengerjakan sendiri karya ini dan mampu bertanggungjawab atas karya ini.

Bila dikemudian hari ada tuntutan dari pihak lain atas karya Saya, dan telah melalui pembuktian yang dapat dipertanggungjawabkan dan ternyata memang ditemukan bukti bahwa Saya telah melanggar pernyataan ini, maka Saya siap dikenai sanksi berdasarkan aturan yang berlaku di Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh.

Demikian pernyataan ini Saya buat dengan sesungguhnya dan tanpa paksaan dari pihak manapun.

Banda Aceh,
Yang Menyatakan,

Aidil Najar

ABSTRAK

Nama	: Aidil Najar
NIM	: 190705023
Program Studi	: Teknologi Informasi
Fakultas	: Sains dan Teknologi
Judul	: Komparasi Serangan Adversarial Pada Model Resnet34 Untuk Tugas Klasifikasi Gambar Imagenet Dataset
Tanggal Sidang	: 02 Juli 2025
Pembimbing I	: Dr. Hendri Ahmadian, S.Si., M.I.M
Pembimbing II	: Mulkan Fadhli, S.T., M.T

Keamanan model *deep learning* dalam tugas klasifikasi gambar semakin menjadi sorotan, terutama karena kemunculan serangan *adversarial* yang dapat mengelabui prediksi model hanya dengan gangguan kecil pada input. Penelitian ini bertujuan untuk menganalisis dan membandingkan efektivitas tiga jenis serangan *adversarial*, yaitu *Fast Gradient Sign Method* (FGSM), *Projected Gradient Descent* (PGD), dan *Patch Attack* terhadap model *ResNet34* pada *dataset ImageNet*. Model *ResNet34* dipilih karena arsitekturnya yang dalam dan telah terbukti unggul dalam berbagai kompetisi visi komputer. *Dataset* yang digunakan adalah subset dari *ImageNet* yang telah melalui proses pra-pemrosesan untuk kompatibilitas dengan arsitektur *ResNet34*.

Hasil penelitian menunjukkan bahwa ketiga serangan mampu menurunkan akurasi klasifikasi secara signifikan. Serangan PGD terbukti paling destruktif, menghasilkan misklasifikasi dengan tingkat kepercayaan tinggi, diikuti oleh FGSM dan *Patch Attack*. Visualisasi *adversarial examples* juga mengonfirmasi bahwa perubahan kecil pada input dapat menyebabkan perubahan besar dalam prediksi model. Penelitian ini memberikan wawasan penting mengenai kerentanan model *deep learning* dan mendorong pengembangan strategi pertahanan yang lebih efektif di masa mendatang.

Kata Kunci: *Adversarial Attack*, *ResNet34*, FGSM, PGD, *Patch Attack*, *ImageNet*.

ABSTRACT

Name	: Aidil Najar
NIM	: 190705023
Department	: Teknologi Informasi
Faculty	: Sains dan Teknologi
Title	: Comparison Of Adversarial Attacks On Resnet34 Model For Image Classification Task Of Imagenet Dataset
Date	: 02 July 2025
Supervisor I	: Dr. Hendri Ahmadian, S.Si., M.I.M
Supervisor II	: Mulkan Fadhli, S.T., M.T

The security of deep learning models in image classification tasks has become increasingly critical due to the emergence of adversarial attacks that can deceive model predictions through subtle input perturbations. This study aims to analyze and compare the effectiveness of three adversarial attack methods, specifically Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Patch Attack against the ResNet34 model using the ImageNet dataset. ResNet34 was chosen for its deep architecture and proven performance in various computer vision benchmarks. A preprocessed subset of the ImageNet dataset was used to ensure compatibility with the ResNet34 input requirements.

The experimental results reveal that all three attacks significantly degrade classification accuracy, with PGD being the most disruptive, followed by FGSM and Patch Attack. Visualizations of the adversarial examples confirm that even minor input modifications can lead to drastic changes in model predictions. This research provides crucial insights into the vulnerability of deep learning models and underscores the need for more robust defense strategies in future developments.

Keywords: Adversarial Attack, ResNet34, FGSM, PGD, Patch Attack, ImageNet.

KATA PENGANTAR

Puji syukur Penulis panjatkan atas kehadirat Allah SWT yang telah memberikan nikmat, rahmat, karunia serta hidayah-Nya, sehingga Penulis dapat menyelesaikan tugas akhir ini yang berjudul "**Komparasi Serangan *Adversarial* Pada Model *Resnet34* Untuk Tugas Klasifikasi Gambar *Imagenet Dataset***". Shalawat beserta salam yang tercurahkan kepada Rasulullah SAW beserta keluarga dan sahabat Beliau sekalian yang telah membawa umat Islam kejalan kebenaran yang penuh dengan ilmu pengetahuan yang bermanfaat bagi dunia dan akhirat.

Penyusunan tugas akhir ini diajukan sebagai salah satu syarat untuk menyelesaikan tugas akhir pada Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, UIN Ar-Raniry. Dalam penulisan tugas akhir ini, penulis dengan segala kerendahan hati ingin mengucapkan terimakasih yang sebesar-besarnya kepada :

1. Ibunda Cut Khairaniah dan Ayahanda Alwi beserta keluarga yang senantiasa memberikan do'a dan dukungan setulus hati yang tak pernah harap kembali.
2. Bapak Dr. Hendri Ahmadian, S.Si., M.I.M, selaku Pembimbing I tugas akhir tugas akhir yang selalu memberikan arahan dan bimbingannya dalam penyusunan tugas akhir ini.
3. Bapak Mulkan Fadhli, S.T., M.T, selaku Pembimbing II tugas akhir tugas akhir yang selalu memberikan arahan dan bimbingannya dalam penyusunan tugas akhir ini.
4. Ibu Malahayati, M.T. dan Bapak Khairan AR, M.Kom, selaku Ketua dan Sekretaris Program Studi Teknologi Informasi, yang telah memberikan bimbingan dan arahannya dalam menyelesaikan tugas akhir ini.
5. Ibu Cut Ida Rahmadiana, S.Si, selaku Staf Program Studi Teknologi Informasi, yang selalu membantu dalam pemberkasan administrasi.
6. Bapak Dr. Ir. M. Dirhamsyah, M.T., IPU selaku Dekan Fakultas Sains dan Teknologi UIN Ar-Raniry.
7. Bapak Dr. Khairizzaman., M.Ag, selaku Ketua Program Studi Magister Ilmu Al-Qur'an dan Tafsir, yang telah memberikan arahan dan bimbingan dalam menyelesaikan tugas akhir ini.

8. Bapak dan Ibu dosen Prodi Teknologi Informasi yang telah membekali penulis dengan ilmu pengetahuan di bidang Teknologi Informasi.
9. Sahabat dan teman-teman yang selalu memberikan dukungan moral untuk menyelesaikan penyusunan tugas akhir ini.
10. Pihak-pihak terkait lainnya yang membantu penulis dalam penyusunan tugas akhir ini.

Penulis menyadari sepenuhnya bahwa dalam peyusunan tugas akhir ini tidak sempurna, untuk itu dengan segala kerendahan hati Penulis menerima saran dan kritikan guna menyempurnakan penyusunan tugas akhir ini. Akhir kata, semoga tugas akhir ini dapat bermanfaat bagi Penulis dan Pembaca dan semoga dicatat sebagai sebuah amal kebaikan oleh Allah SWT. Amin Ya Rabbal A'lamin.

Banda Aceh, 17 Juli 2025
Yang Menyatakan

Aidil Najar

DAFTAR ISI

جامعة الرانري
AR-RANIRY

DAFTAR GAMBAR

DAFTAR TABEL

جامعة الرانري
AR-RANIRY

BAB I
PENDAHULUAN

1.1. Latar Belakang

Machine Learning merupakan bagian dari kecerdasan buatan yang menginstruksikan komputer untuk mempelajari informasi dari data dan mengeluarkan prediksi atau tindakan berdasarkan interpretasi terhadap data tersebut. Dengan *Machine Learning*, komputer dapat mengambil keputusan atau melakukan tindakan berdasarkan analisis data tanpa memerlukan pemrograman eksplisit. Secara sederhana, mesin "mengambil pelajaran" dari pengalaman sebelumnya dan menggunakan pengetahuan yang diperoleh untuk menghadapi situasi baru (Firdaus et al., 2023).

Perkembangan teknologi *Deep Learning* telah melahirkan model-model *Convolutional Neural Network* (CNN) yang canggih, salah satunya adalah *ResNet34*. *ResNet34*, dengan arsitektur residual learning-nya, telah menunjukkan performa unggul dalam berbagai tugas visi komputer, termasuk klasifikasi gambar pada *Dataset* ImageNet. Kemampuannya dalam mengekstraksi fitur-fitur penting dari gambar dan mengatasi masalah vanishing *gradient* menjadikannya berharga dalam berbagai aplikasi, seperti pengenalan objek, deteksi objek, dan segmentasi gambar (Alzubaidi et al., 2021).

Meskipun demikian, model *Deep Learning* seperti *ResNet34* rentan terhadap *Adversarial Attack*, yaitu input yang dimodifikasi secara halus namun dapat menipu model dan menyebabkan kesalahan prediksi. Pada kasus klasifikasi gambar, *Adversarial Attack* dapat berupa gambar dengan sedikit perubahan piksel yang menyebabkan *ResNet34* salah mengklasifikasikan objek dalam gambar (M. Gao et al., 2021).

Pada penelitian Sen and Dasgupta (2023) yang berjudul *"Adversarial Attacks on Image Classification Models - FGSM and Patch Attacks and their Impact"* berfokus pada analisis dampak serangan *Adversarial*, khususnya *Fast Gradient Sign Method* (FGSM) dan *Adversarial Patch Attack*, terhadap akurasi

model klasifikasi gambar seperti *ResNet34*, *GoogleNet*, dan *DenseNet-161* pada *Dataset ImageNet*. Penelitian ini menunjukkan bahwa model-model tersebut, meskipun memiliki kinerja tinggi dalam kondisi normal, sangat rentan terhadap serangan *Adversarial*. Serangan FGSM dan *Adversarial Patch* menyebabkan penurunan signifikan dalam akurasi klasifikasi, menunjukkan perlunya pengembangan metode pertahanan yang lebih kuat untuk melindungi model *Deep Learning* dari ancaman *Adversarial* (Sen & Dasgupta, 2023).

Serangan *adversarial* telah menjadi ancaman serius dalam bidang deep learning, terutama dalam tugas klasifikasi gambar pada *dataset ImageNet* menggunakan model seperti *ResNet34*. Tiga metode serangan yang umum digunakan adalah *Fast Gradient Sign Method* (FGSM), *Projected Gradient Descent* (PGD), dan *Patch Attack*, yang mampu menipu model dengan cara berbeda. FGSM adalah metode serangan cepat berbasis gradien yang menambahkan gangguan kecil pada gambar dalam satu iterasi untuk mengubah prediksi model secara instan. PGD, sebagai versi iteratif dari FGSM, melakukan optimasi bertahap untuk menemukan perturbasi yang lebih efektif, sehingga lebih sulit untuk ditangkal. Sementara itu, *Patch Attack* menggunakan gangguan dalam bentuk *Patch* yang ditempatkan pada bagian tertentu dari gambar, memungkinkan manipulasi sistem dalam melakukan klasifikasi. Penelitian ini akan membandingkan tiga jenis serangan *adversarial*, yaitu *Fast Gradient Sign Method* (FGSM), *Projected Gradient Descent* (PGD), dan *Patch Attack*, terhadap model *ResNet34* dalam tugas klasifikasi gambar pada dataset *ImageNet*, guna memberikan wawasan tentang kerentanan *ResNet34* terhadap serangan tersebut dan dimasa depan dapat membantu dalam pengembangan metode pertahanan yang lebih baik.

1.2. Rumusan Masalah

Dari Latar Belakang di atas maka dapat diambil beberapa rumusan masalah sebagai berikut:

1. Bagaimana visualisasi akurasi model *ResNet34* pada *ImageNet Dataset* setelah diserang menggunakan *Fast Gradient Sign Method* (FGSM), *Projected Gradien Descent* (PGD), dan *Patch Attack*?
2. Bagaimana visualisasi Gambar *Adversarial* examples yang dihasilkan oleh *Fast Gradient Sign Method* (FGSM), *Projected Gradien Descent* (PGD), dan *Patch Attack*?
3. Bagaimana komparasi akurasi serangan *adversarial Fast Gradient Sign Method* (FGSM), *Projected Gradien Descent* (PGD), dan *Patch Attack* pada Model *ResNet34*?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disusun, penelitian ini bertujuan untuk:

1. Memvisualisasikan perubahan akurasi model *ResNet34* pada ImageNet Dataset setelah diserang menggunakan *Fast Gradient Sign Method* (FGSM), *Projected Gradien Descent* (PGD), dan *Patch Attack*.
2. Memvisualisasikan gambar *Adversarial* Examples yang dihasilkan oleh serangan *Fast Gradient Sign Method* (FGSM), *Projected Gradien Descent* (PGD), dan *Patch Attack*.
3. Mengomparasikan akurasi serangan *Adevrsarial Fast Gradient Sign Method* (FGSM), *Projected Gradien Descent* (PGD), dan *Patch Attack* pada Model *ResNet34*.

1.4. Batasan Masalah

Dalam penelitian ini mempunyai beberapa batasa masalah, berikut batasan – batasan masalahnya:

1. Penelitian ini hanya menggunakan model *ResNet34* yang telah dilatih sebelumnya pada *Dataset ImageNet*.

2. Serangan *Adversarial* yang digunakan dalam penelitian ini *Fast Gradient Sign Method* (FGSM), *Projected Gradient Descent* (PGD), dan *Patch Attack*
3. *Dataset* yang digunakan adalah subset *ImageNet Dataset*.

1.5. Manfaat Penelitian

Berikut manfaat dalam penelitian ini:

1. Memberikan pemahaman tentang proses serangan *Fast Gradient Sign Method* (FGSM), *Projected Gradient Descent* (PGD), dan *Patch Attack* pada model *Deep Learning*, khususnya *ResNet34*.
2. Memberikan informasi tentang efektivitas proses *Fast Gradient Sign Method* (FGSM), *Projected Gradient Descent* (PGD), dan *Patch Attack* pada model *ResNet34* dan *Dataset ImageNet*.
3. Memberikan informasi komparasi akurasi serangan *adversarial Fast Gradient Sign Method* (FGSM), *Projected Gradien Descent* (PGD), dan *Patch Attack* pada Model *ResNet34*.