Vol. 10, No. 1, April 2025, Page. 13-27
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Deteksi dan Pemulihan Serangan *Adversarial* : Studi tentang *Fast Gradient Sign Method* dan *Autoencoder* untuk Pencitraan Medis

Yefta Christian[*1], Tony Tan[2], Jesen Winardo[3]
[1,2] Program Studi Sistem Informasi, Fakultas Ilmu Komputer Universitas Internasional Batam, Indonesia
Email: [1]Yefta@uib.ac.id, [2]Tony@uib.ac.id, [3]Jsnwnrdo@gmail.com

Abstrak

Penelitian ini membahas tantangan serangan *Fast Gradient Sign Method* (FGSM) dalam pencitraan medis dan keamanan data. Karena layanan kesehatan semakin mengandalkan model *Machine Learning*, jaminan integritas dan kerahasiaan dalam data medis menjadi sangat penting. Studi ini meninjau literatur terkait serangan adversarial pada *Machine Learning*, khususnya dalam konteks aplikasi pencitraan medis. Kerentanan ini mencakup risiko manipulasi data yang dapat menyebabkan kesalahan diagnosis atau pelanggaran privasi. Literatur yang dianalisis mencakup jurnal medis dan konferensi pembelajaran mesin guna memahami metode serangan, dampaknya, serta pendekatan mitigasinya.
Hasil penelitian menunjukkan bahwa serangan adversarial dapat secara signifikan menurunkan performa model *deep learning*. Untuk mengatasinya, Autoencoder terbukti efektif dalam memulihkan data yang diserang dan meningkatkan akurasi prediksi. Selain itu, teknik augmentasi data dapat memperkuat ketahanan model, terutama pada dataset yang tidak seimbang, serta mengurangi risiko overfitting. Dengan menerapkan kerangka kerja SEMMA (*Sample, Explore, Modify, Model, and Assess*), studi ini membuktikan bahwa serangan FGSM yang hanya menambahkan sedikit noise dapat menggagalkan prediksi model, namun dapat dipulihkan secara efektif oleh Autoencoder. Penelitian ini memberikan kontribusi penting dalam memahami serangan adversarial dalam bidang medis dan menawarkan strategi pertahanan yang menjanjikan.

Kata kunci: Adversarial Robustness, Autoencoder, Fast Gradient Sign Method, Serangan Adversarial, Xception

Abstract

This research addresses the challenges of Fast Gradient Sign Method (FGSM) attacks in medical imaging and data security. As healthcare increasingly relies on Machine Learning models, assurance of integrity and confidentiality in medical data is of paramount importance. This study reviews the literature related to adversarial attacks on Machine Learning, specifically in the context of medical imaging applications. These vulnerabilities include the risk of data manipulation that could lead to misdiagnosis or privacy breaches. The literature analysed includes medical journals and machine learning conferences to understand attack methods, their impact, and mitigation approaches.
This research addresses the challenges of Fast Gradient Sign Method (FGSM) attacks in medical imaging and data security. As healthcare increasingly relies on Machine Learning models, assurance of integrity and confidentiality in medical data is of paramount importance. This study reviews the literature related to adversarial attacks on Machine Learning, specifically in the context of medical imaging applications. These vulnerabilities include the risk of data manipulation that could lead to misdiagnosis or privacy breaches. The literature analysed includes medical journals and machine learning conferences to understand attack methods, their impact, and mitigation approaches.

Keywords: *Adversarial Attacks, Adversarial Robustness, Autoencoder, Fast Gradient Sign Method, Xception*

1. PENDAHULUAN

Kemajuan pesat dalam pencitraan medis telah merevolusi perawatan kesehatan, memungkinkan diagnosis yang lebih akurat, deteksi dini penyakit, dan hasil yang lebih baik bagi pasien.[1] Dengan memanfaatkan teknologi canggih seperti pembelajaran mendalam dan kecerdasan buatan, sistem pencitraan medis semakin banyak digunakan untuk diagnosis otomatis, pemantauan perkembangan penyakit, dan perencanaan perawatan.[2][3] Sistem ini menjadi solusi penting untuk mengatasi tantangan perawatan kesehatan global, terutama di wilayah dengan keterbatasan keahlian dan sumber daya medis.[4] Namun, seiring dengan adopsi luas teknologi ini, muncul kerentanan baru yang dapat

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

dieksploitasi oleh serangan jahat, seperti serangan *adversarial*, yang mengancam integritas data medis dan keakuratan sistem diagnosis berbasis AI. Salah satu teknik serangan yang terkenal adalah Fast Gradient Sign Method (FGSM), yang memperkenalkan gangguan kecil pada data masukan untuk memanipulasi prediksi model *Machine learning*. *Adversarial attack*, dalam konteks *Machine Learning* terjadi karena model sering kali bergantung pada pola-pola tertentu dalam data pelatihan. Hal ini menciptakan celah di mana input yang sedikit dimodifikasi dapat dengan mudah mengeksploitasi bias model tersebut. Dalam sistem *Medical Imaging*, ini berarti bahwa gangguan kecil yang tidak terlihat oleh manusia dapat menyebabkan model membuat prediksi yang salah.[5] Musuh yang memanfaatkan serangan ini dapat berupa peretas individu, kelompok kriminal siber, pihak internal yang tidak puas, atau bahkan entitas kompetitif. Misalnya, serangan ini dapat digunakan untuk memanipulasi hasil diagnosis demi keuntungan finansial, sabotase layanan kesehatan, atau melemahkan kepercayaan publik terhadap teknologi berbasis AI.[6] *FGSM* secara luas dikenal karena efisiensinya dalam menghasilkan contoh-contoh yang berlawanan karena kesederhanaan dan kecepatan komputasinya.[7][8] Dengan menghitung gradien kerugian sehubungan dengan input dan menerapkan gangguan minimal ke arah gradien, FGSM dapat secara efektif mengelabui model klasifikasi.[9]

Namun, seiring dengan adopsi luas teknologi ini, muncul kerentanan baru yang dapat dieksploitasi oleh serangan jahat, seperti serangan adversarial, yang mengancam integritas data medis dan keakuratan sistem diagnosis berbasis AI. Salah satu teknik serangan yang terkenal adalah Fast Gradient Sign Method (FGSM), yang memperkenalkan gangguan kecil pada data masukan untuk memanipulasi prediksi model Machine Learning[10]. Dalam penelitian ini, nilai epsilon (ϵ) yang digunakan adalah 0,01, karena dianggap cukup kecil untuk tidak mengganggu persepsi visual manusia, namun cukup besar untuk memengaruhi keputusan model secara signifikan.

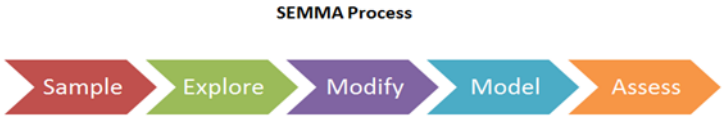
Meski demikian, belum banyak studi yang secara eksplisit menerapkan Autoencoder sebagai strategi pertahanan terhadap serangan FGSM dalam domain citra medis (khususnya X-ray). Mayoritas studi hanya fokus pada metode klasifikasi atau pelatihan adversarial, tetapi belum mengeksplorasi peran Autoencoder dalam pemulihan citra yang diserang[11]. Selain itu, sebagian literatur yang ada lebih banyak menyoroti domain lain seperti intrusi jaringan atau pemrosesan kode program, yang kurang relevan dalam konteks imaging medis[12]. Oleh karena itu, diperlukan pendekatan yang terfokus untuk mengevaluasi efektivitas Autoencoder dalam meningkatkan ketahanan model terhadap serangan FGSM di lingkungan medis yang sensitif.

Untuk menangkal risiko yang ditimbulkan oleh serangan FGSM, peneliti telah mengeksplorasi berbagai strategi pertahanan, di antaranya *Autoencoder* telah menunjukkan harapan besar. *Autoencoder* adalah jaringan saraf yang dirancang untuk mempelajari representasi yang efisien dari data input dengan mengompresi dan merekonstruksi data tersebut. Sebagai mekanisme pertahanan, *Autoencoders* digunakan untuk mendeteksi dan menolak input yang tidak sesuai, mengembalikan bentuk aslinya yang tidak diubah sebelum meneruskannya ke model klasifikasi[13]. Pendekatan ini memastikan bahwa pengklasifikasi beroperasi pada data yang bersih, dengan demikian menjaga akurasi dan keandalan bahkan dengan adanya manipulasi yang tidak sesuai[14]. *Autoencoder* sangat efektif karena dapat menggeneralisasi dengan baik dan mengidentifikasi input yang menyimpang dari distribusi data normal.[15]

Tujuan penelitian ini adalah untuk mengatasi tantangan kritis yang ditimbulkan oleh serangan musuh pada sistem pencitraan medis, dengan fokus khusus pada serangan FGSM dan implementasi pertahanan berbasis *Autoencoder*. Tujuan utamanya adalah untuk mengevaluasi dampak serangan FGSM pada model klasifikasi Medical Imaging, mengembangkan dan mengimplementasikan mekanisme pertahanan *Autoencoder* untuk mengurangi serangan ini, dan menganalisis pertukaran yang terkait dengan pendekatan ini. Dengan mempelajari efektivitas serangan FGSM secara sistematis dan mengeksplorasi ketahanan *Autoencoder*, penelitian ini berusaha untuk berkontribusi pada pengetahuan yang berkembang dalam *Machine Learning* yang bersifat permusuhan dalam domain perawatan kesehatan. Pada akhirnya, penelitian ini bertujuan untuk meningkatkan keamanan dan keandalan sistem pencitraan medis yang digerakkan oleh AI, memastikan diagnosis yang akurat, menjaga keselamatan pasien, dan menumbuhkan kepercayaan pada solusi perawatan kesehatan yang diaktifkan oleh teknologi.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

2. METODE PENELITIAN



Gambar 1. Proses SEMMA

SEMMA merupakan kerangka kerja yang dirancang oleh SAS Institute untuk mendukung proses data mining secara terstruktur. Proses ini bertujuan untuk memastikan pendekatan sistematis dalam memahami data, membangun model prediksi, dan mengevaluasi hasil. Metodologi penelitian ini dirancang untuk menjawab tantangan serangan *adversarial* pada sistem pencitraan medis berbasis AI, khususnya dalam mendeteksi pneumonia dari gambar *X-Ray* dada. Penelitian ini menggunakan pendekatan kualitatif melalui tinjauan literatur untuk membangun fondasi konseptual dan mengumpulkan wawasan dari penelitian yang sudah ada mengenai klasifikasi Medical Imaging, serangan *adversarial*, dan mekanisme pemulihan. Selain itu, proyek ini secara sistematis mengikuti alur kerja terstruktur berdasarkan kerangka kerja *SEMMA* untuk memastikan pendekatan yang komprehensif dan metodis dalam pelaksanaannya. Pendekatan ini mencakup pengumpulan data, eksplorasi, preprocessing, pelatihan model, simulasi serangan musuh, implementasi strategi pemulihan, dan evaluasi kinerja, yang mengintegrasikan wawasan dari temuan kualitatif dengan langkah-langkah praktis untuk mengembangkan sistem pencitraan medis berbasis AI yang kuat.

2.1 *Sample*

Dataset X-ray dada diperoleh dari sumber terbuka seperti Kaggle, RSNA, dan Huggingface yang menyediakan data secara publik untuk keperluan penelitian non-komersial. Karena data telah dianonimkan oleh penyedia, penggunaan dataset ini tidak memerlukan persetujuan etik tambahan. Namun, jika penelitian dikembangkan lebih lanjut menggunakan data internal rumah sakit atau data pasien baru, maka persetujuan dari komite etik penelitian akan diperlukan. Dataset awal berisi sekitar 4.827 gambar pneumonia dan 2.394 gambar normal. Setelah proses pembersihan data, seperti menghapus gambar duplikat, rusak, atau berkualitas rendah, dataset final terdiri dari 4.273 gambar pneumonia dan 1.583 gambar normal. Dataset ini kemudian dibagi menjadi dua subset: data pelatihan (3.883 gambar pneumonia dan 1.349 gambar normal) serta data pengujian (390 gambar pneumonia dan 234 gambar normal).

2.2 *Explore*

Tahap eksplorasi bertujuan untuk memahami struktur dataset dan memastikan kualitas data. Analisis distribusi dilakukan untuk memeriksa ketimpangan antara kategori Normal dan Pneumonia. Proses pembersihan data dilakukan untuk menghapus gambar yang rusak, duplikat, atau berkualitas rendah, menghasilkan dataset bersih dan konsisten. Selain itu, contoh gambar dari setiap kelas dianalisis untuk mengidentifikasi pola visual yang mungkin membedakan kedua kategori. Ketidakseimbangan kelas juga diperiksa, dan jika ditemukan, teknik *augmentasi* atau class weighting direncanakan untuk diterapkan pada tahap selanjutnya.

2.3 *Modify*

Pada tahap ini, gambar diubah ukurannya menjadi 224x224 piksel sesuai kebutuhan model deep learning, dan nilai piksel dinormalisasi ke rentang (0,1) untuk meningkatkan stabilitas numerik. Teknik *augmentasi* data, seperti rotasi acak, pembalikan, pembesaran, dan penyesuaian kecerahan, diterapkan untuk mengatasi ketidakseimbangan kelas serta meningkatkan variasi data agar model dapat menggeneralisasi lebih baik. Augmentasi mencakup rotasi acak hingga 15°, zoom antara 0.9–1.1, dan penyesuaian brightness ±20%. Teknik ini dipilih untuk meniru variasi posisi pasien, pencahayaan, dan skala radiografi dalam praktik klinis. Dengan augmentasi ini, model dilatih untuk lebih tangguh terhadap perbedaan kondisi gambar, sekaligus mengurangi risiko overfitting pada dataset terbatas.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

2.4 *Model*

Model klasifikasi berbasis Xception dilatih menggunakan dataset yang telah diproses. Pengoptimal Adamax digunakan untuk melatih model, dengan callback seperti EarlyStopping dan ModelCheckpoint untuk mencegah *overfitting* dan menyimpan model terbaik. Kinerja model dievaluasi pada subset pelatihan, validasi, dan pengujian untuk menentukan akurasi dasar sebelum simulasi serangan. Model Xception terdiri dari 71 lapisan dengan struktur depthwise separable convolution dan residual connection. Model dimodifikasi untuk klasifikasi biner. Pelatihan dilakukan menggunakan optimizer Adamax dengan learning rate 0.0001, batch size 32, dan maksimal 30 epoch, dengan EarlyStopping (patience = 5) dan ModelCheckpoint untuk menyimpan model terbaik.

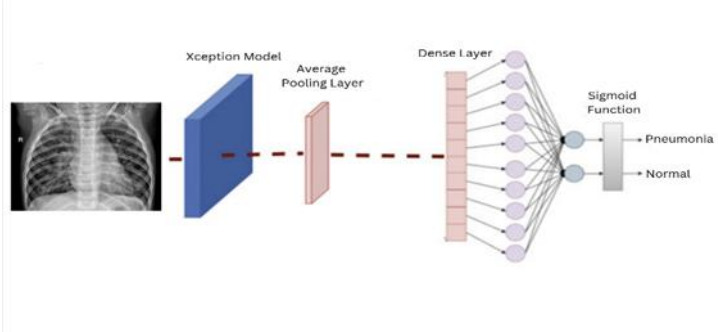
Simulasi serangan dilakukan menggunakan metode FGSM untuk menghasilkan contoh gambar *adversarial* yang bertujuan untuk menurunkan akurasi model. Gambar yang terganggu oleh serangan ini dievaluasi untuk menentukan tingkat keberhasilan serangan. Setelah itu, *Autoencoder* diterapkan untuk memulihkan gambar yang diserang. *Autoencoder* dilatih menggunakan data *adversarial* sebagai input dan gambar asli sebagai target. Efektivitasnya diukur menggunakan kehilangan rekonstruksi, dan kinerja gambar yang dipulihkan diuji kembali pada model Xception untuk menilai akurasi.

2.5 *Assess*

Pada tahap akhir, evaluasi dilakukan terhadap kinerja model Xception pada tiga skenario: data asli, data *adversarial*, dan data yang dipulihkan. Laporan evaluasi akhir mencakup analisis kinerja model, efektivitas FGSM sebagai metode serangan, dan efisiensi *Autoencoder* sebagai mekanisme pertahanan. Proses ini bertujuan untuk memberikan wawasan tentang pengembangan sistem berbasis AI yang tangguh untuk pencitraan medis.

3. HASIL DAN PEMBAHASAN

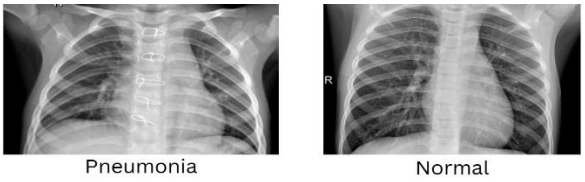
3.1 Xception



Gambar 2. Skema Xception

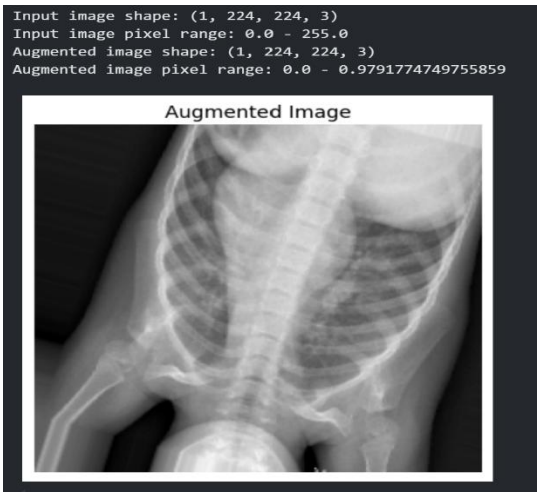
Skema Xception mengilustrasikan alur klasifikasi gambar X-Ray untuk mendeteksi NORMAL atau PNEUMONIA. Gambar input diproses melalui model Xception yang telah dilatih sebelumnya (tanpa lapisan klasifikasi) untuk mengekstrak fitur-fitur penting. Fitur-fitur ini dirangkum menggunakan Average Pooling Layer sebelum diteruskan ke Dense Layer, yang memprosesnya untuk tugas klasifikasi. Lapisan terakhir menggunakan fungsi aktivasi Sigmoid untuk menghasilkan dua probabilitas kelas: NORMAL atau PNEUMONIA. Hasil akhirnya adalah prediksi kelas berdasarkan probabilitas tertinggi. Skema ini menggunakan Xception untuk ekstraksi fitur dan lapisan tambahan untuk klasifikasi.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI



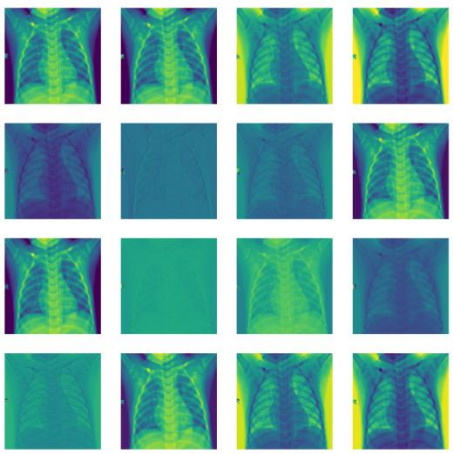
Gambar 3. X-Ray Pneumonia dan Normal

Penelitian ini menggunakan data Medical Image X-Ray dada yang diperoleh dari sumber terbuka, termasuk situs web medis yang menyediakan set data secara publik. Dataset terdiri dari dua kategori utama: Normal dan Pneumonia, yang diatur dalam folder terpisah untuk memudahkan proses pelabelan otomatis. Setelah proses pembersihan data, termasuk penghapusan gambar duplikat, gambar yang rusak, dan gambar dengan kualitas yang tidak sesuai, set data akhir yang digunakan untuk analisis mencakup total 4.827 gambar pneumonia dan 1.583 gambar normal. Data ini kemudian dibagi menjadi dua subset utama, yaitu data pelatihan (Train) yang terdiri dari 3.883 citra pneumonia dan 1.583 citra normal, dan data pengujian (Test) yang terdiri dari 390 citra pneumonia dan 234 citra normal. Model Xception digunakan sebagai arsitektur inti untuk mendeteksi pneumonia dari gambar *X-Ray* dada.



Gambar 4. Gambar Setelah Augmentasi

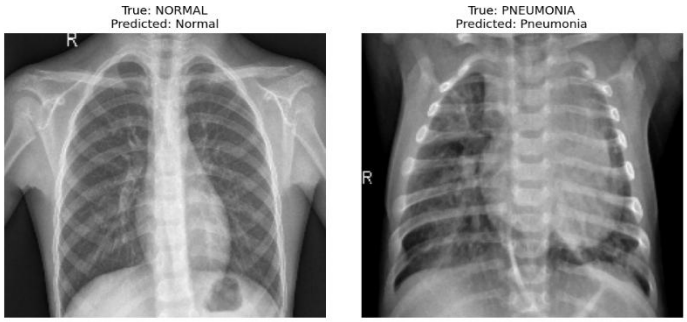
Prosesnya dimulai dengan gambar X-Ray masukan yang telah melalui proses preprocessing, termasuk mengubah ukuran menjadi 224x224 piksel dan menormalkan nilai piksel ke rentang (0,1). Model Xception, yang telah dilatih sebelumnya pada dataset ImageNet, dimodifikasi untuk memenuhi kebutuhan klasifikasi biner (Normal dan Pneumonia).



Gambar 5. *Xception Model Layer*

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Gambar X-Ray diproses melalui lapisan awal model Xception, yang mencakup serangkaian konvolusi dan operasi konvolusi yang dapat dipisahkan berdasarkan kedalaman untuk mengekstrak fitur penting dari gambar. Setelah melalui tahap ekstraksi fitur, lapisan Average Pooling digunakan untuk mengurangi dimensi output tanpa kehilangan informasi penting. Hasil dari Average Pooling kemudian diteruskan ke lapisan Dense, yang bertindak sebagai jembatan untuk memetakan fitur ke dalam dua kategori. Dense layer ini dilengkapi dengan fungsi aktivasi sigmoid, yang mengubah output menjadi nilai probabilitas antara 0 dan 1, yang memungkinkan model untuk memutuskan apakah gambar tersebut termasuk dalam kategori Pneumonia atau Normal. Model ini dirancang untuk mengoptimalkan akurasi prediksi melalui fungsi kehilangan crossentropy biner, yang sangat cocok untuk klasifikasi biner. Proses ini didukung oleh penggunaan callback, seperti EarlyStopping dan ModelCheckpoint, untuk memastikan bahwa pelatihan model berlangsung secara optimal dan model terbaik disimpan untuk dievaluasi. Pada akhirnya, sistem ini memanfaatkan kekuatan pembelajaran transfer dari Xception untuk membuat model yang andal dalam mendeteksi pneumonia dari gambar X-Ray, meskipun dihadapkan pada tantangan seperti variasi gambar dan kemungkinan serangan musuh.

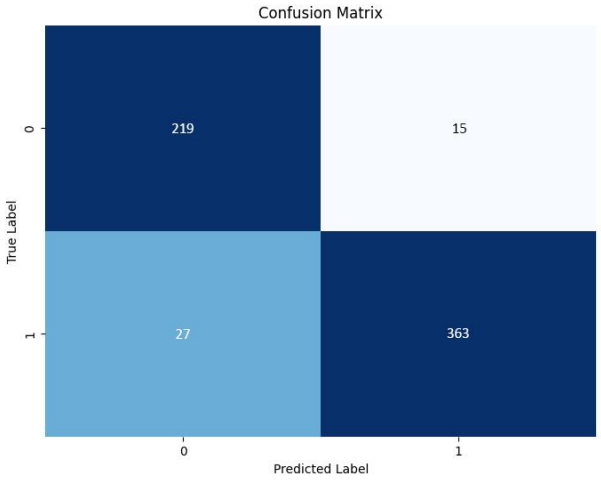


Gambar 6. *XCeption Model Prediction*

Gambar hasil prediksi menunjukkan dua contoh data X-Ray yang digunakan untuk menguji model Xception yang telah dilatih. Gambar pertama menunjukkan hasil prediksi untuk kategori Normal, di mana model dengan tepat mengidentifikasi gambar sebagai Normal, sesuai dengan label aslinya. Gambar kedua menunjukkan hasil prediksi untuk kategori Pneumonia, di mana model juga dengan benar memprediksi label aslinya. Pada gambar ini, terdapat tanda-tanda opasitas (penggelapan) di area paru-paru, yang merupakan indikasi klinis pneumonia. Model ini menggunakan arsitektur Xception dengan pembelajaran transfer, yang memungkinkan ekstraksi fitur mendalam dari data X-Ray, seperti pola jaringan paru-paru, untuk membuat klasifikasi yang akurat.

Untuk mengevaluasi lebih lanjut ketahanan model Xception dalam aplikasi dunia nyata di mana sistem pencitraan medis mungkin menghadapi tantangan seperti serangan musuh, penelitian ini mengeksplorasi implementasi simulasi musuh. Meskipun model Xception menunjukkan akurasi yang tinggi dalam mendeteksi pneumonia dalam kondisi normal, kerentanannya terhadap gangguan yang disengaja, seperti gangguan yang disengaja, perlu diselidiki untuk memastikan keandalan dan ketahanannya dalam skenario perawatan kesehatan yang kritis.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI



Gambar 7. *Confusion Matrix*

Matriks confusion ini dihasilkan menggunakan data testing/validasi yang telah dipersiapkan sebelumnya sejak awal. Data testing/validasi ini digunakan untuk mengevaluasi performa model klasifikasi terhadap dua kelas, yaitu Kelas 0 ('Normal') dan Kelas 1 ('Pneumonia'). Matriks ini mencerminkan distribusi hasil prediksi model, termasuk prediksi yang benar dan salah, berdasarkan data testing yang telah dipisahkan dari data pelatihan. Berikut ini adalah interpretasi dari setiap elemen matriks:

- *True Negative* (TN): Sebanyak 219 sampel dari kelas 'Normal' diprediksi dengan benar sebagai 'Normal'.
- *False Positive* (FP): Sebanyak 15 sampel dari kelas 'Normal' salah diprediksi sebagai 'Pneumonia'.
- *False Negative* (FN): Seba Sebanyak 27 sampel dari kelas 'Pneumonia' salah diprediksi sebagai 'Normal'.
- *True Positive* (TP): Sebanyak 363 sampel dari kelas 'Pneumonia' diprediksi dengan benar sebagai 'Pneumonia'.

Dari confusion matrix ini, dapat dilihat bahwa model memiliki akurasi yang cukup baik dalam mengenali kedua kelas, meskipun terdapat 1 false negative, dimana satu kasus pneumonia gagal terdeteksi dan diprediksi sebagai kasus normal. Model cukup handal, dengan 2 prediksi yang benar untuk setiap kelas, namun masih ada ruang untuk perbaikan terutama dalam mengurangi false negative pada Kelas 1. Negatif palsu dapat memiliki dampak yang lebih serius pada aplikasi medis, karena salah mendiagnosis penyakit dapat mengakibatkan keterlambatan pengobatan. Oleh karena itu, langkah-langkah perbaikan dapat difokuskan pada peningkatan recall untuk Kelas 1 untuk memastikan lebih banyak kasus pneumonia yang dapat dideteksi dengan benar.

Tabel 1. Precision , Recall , F1-Score dan Support pada Xception

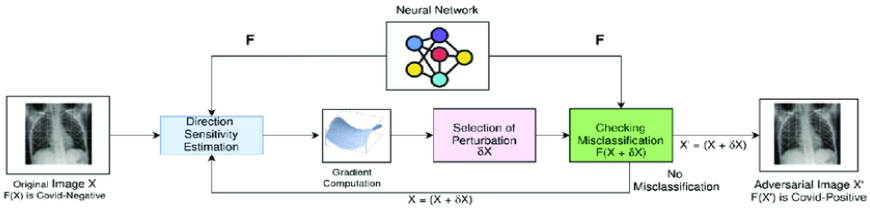
	Precision	**Recall**	**F1-Score**	**Support**
Class 0	0.93	0.89	0.913	246
Class 1	0.93	0.96	0.94	378
Accuracy			0.93	
Macro Avg	0.93	0.92	0.92	
Weighted Avg	0.93	0.93	0.93	

Meskipun model menunjukkan akurasi keseluruhan yang tinggi, masih terdapat sejumlah *false negative* — yaitu kasus pneumonia yang salah diklasifikasikan sebagai normal. Dalam praktik klinis, kesalahan ini sangat krusial karena dapat menyebabkan keterlambatan pengobatan dan memperburuk kondisi pasien. Oleh karena itu, peningkatan *recall* untuk kelas pneumonia sangat penting.

Vol. 10, No. 1, April 2025, Page. 13-27
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Hasil evaluasi kinerja model klasifikasi ditunjukkan pada tabel di atas. Untuk kelas 0 (Normal), presisi tercatat sebesar 0.94, yang berarti bahwa 94% dari prediksi untuk kelas ini benar, sedangkan recall mencapai nilai 0.89, yang berarti bahwa 89% dari total contoh kelas 0 berhasil dikenali oleh model. Nilai F1 untuk kelas 0 adalah 0.91, yang mencerminkan keseimbangan antara presisi dan recall, dengan data pendukung sebanyak 246 kasus. Sementara itu, untuk kelas 1 (Pneumonia), presisi tercatat sebesar 0.93, yang berarti bahwa 93% dari prediksi untuk kelas ini benar. Recall untuk kelas ini mencapai nilai 0.96, yang menunjukkan bahwa 96% dari total kasus pneumonia berhasil dideteksi oleh model. Nilai F1 untuk kelas 1 adalah 0.95, dengan data pendukung sebanyak 378 kasus. Secara keseluruhan, model memiliki akurasi 0.93, yang berarti bahwa model memprediksi 93% dari data uji dengan benar. Rata-rata makro dari metrik menunjukkan nilai presisi 0.93, recall 0.93, dan skor F1 0.93, yang menggambarkan kinerja yang konsisten antara kedua kelas. Selain itu, rata-rata tertimbang menunjukkan presisi 0.93, recall 0.93, dan skor F1 0.93, yang mempertimbangkan jumlah contoh di setiap kelas.

3.2 *Fast Gradient Sign Method*

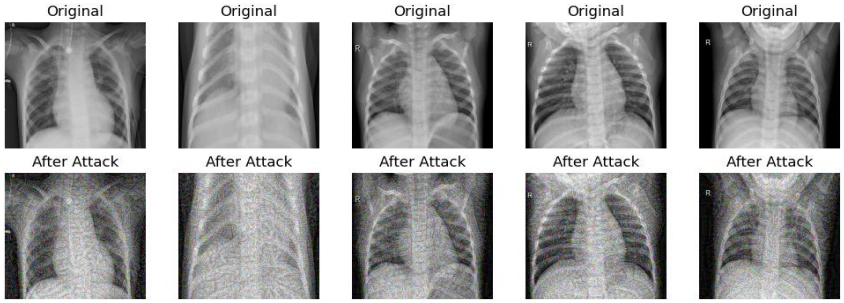


Gambar 8. Skema FGSM

Skema serangan lawan menguraikan proses pembuatan gambar lawan untuk menguji kerentanan jaringan saraf. Dimulai dengan Gambar Asli, model menjalani Estimasi Sensitivitas Arah melalui Komputasi Gradien untuk mengidentifikasi gangguan optimal δX. Gangguan ini diterapkan untuk menghasilkan Citra Lawan X′, yang dievaluasi oleh jaringan saraf. Langkah Memeriksa Misklasifikasi menentukan apakah X′ berhasil menghasilkan prediksi yang salah. Proses ini menunjukkan kerentanan model terhadap perubahan kecil yang dibuat dengan hati-hati, menekankan perlunya langkah-langkah pertahanan yang kuat terhadap serangan musuh. Pada fase simulasi serangan musuh, proyek ini menggunakan Fast Gradient Sign Method (FGSM), salah satu teknik yang paling populer untuk menghasilkan contoh serangan musuh. FGSM menambahkan gangguan kecil, yang hampir tidak terlihat oleh manusia, pada gambar asli dengan tujuan menipu model untuk membuat prediksi yang salah. Metode ini bekerja dengan mengeksploitasi gradien dari fungsi kerugian relatif terhadap gambar input. Derau ditambahkan ke arah gradien dengan skala epsilon (ϵ) yang mengontrol intensitas derau.

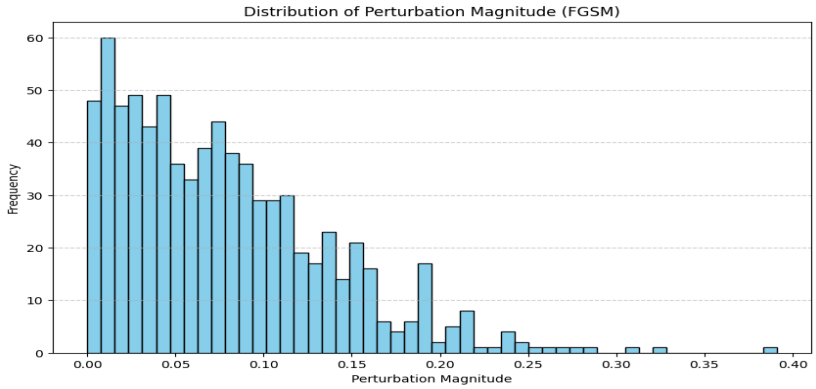
Langkah pertama dalam simulasi ini adalah membuat contoh-contoh yang berlawanan dari kumpulan data yang ada. Untuk setiap gambar dalam dataset, noise dengan nilai epsilon 0,01 ditambahkan ke gambar asli. Nilai epsilon dipilih untuk menjaga agar noise cukup kecil sehingga tidak terlihat secara visual, tetapi cukup kuat untuk mempengaruhi prediksi model. Proses ini diterapkan pada data pelatihan dan validasi, menghasilkan dua set gambar yang berlawanan: *train_x_adversarial* dan *val_x_adversarial*.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI



Gambar 9. Gambar Original dan Gambar Setelah Diserang FGSM

Setelah menghasilkan gambar yang berlawanan, model Xception yang telah dilatih sebelumnya diuji pada dataset ini. Hasil pengujian menunjukkan penurunan akurasi yang signifikan. Untuk mengevaluasi sensitivitas model terhadap gangguan, digunakan tiga nilai epsilon: 0.01, 0.03, dan 0.05. Masing-masing menghasilkan dataset adversarial berbeda. Hasil menunjukkan bahwa semakin besar epsilon, semakin rendah akurasi model. Penurunan ini divisualisasikan melalui grafik dan tabel perbandingan akurasi antar-epsilon. Sebagai contoh, jika akurasi model pada data asli mendekati 93%, akurasinya bisa turun di bawah 80% setelah serangan *adversarial*. Penurunan ini mengindikasikan bahwa model sangat rentan terhadap serangan ini, meskipun noise yang ditambahkan kecil.

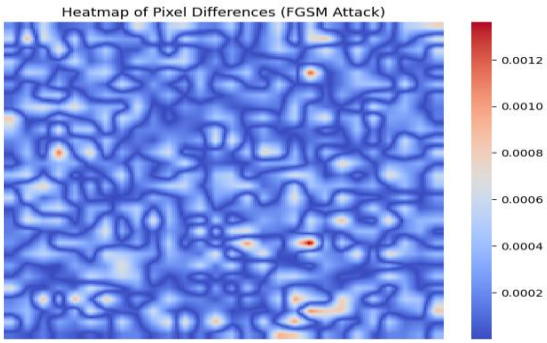


Gambar 10. Distribution of Perturbation Magnitude pada FGSM

Gambar 10 menunjukkan distribusi besaran gangguan yang dihasilkan oleh serangan FGSM. Sebagian besar nilai gangguan terkonsentrasi pada kisaran rendah antara 0,00 dan 0,10, dengan frekuensi tertinggi di sekitar nilai 0,00 hingga 0,05, yang mengindikasikan bahwa hanya sedikit gangguan yang diperlukan untuk mengganggu prediksi model. Frekuensi mulai menurun secara bertahap seiring dengan meningkatnya magnitudo, dengan gangguan yang lebih besar (di atas 0,20) menjadi lebih jarang terjadi. Magnitudo tertinggi dalam distribusi ini mendekati 0,40, tetapi jumlahnya sangat kecil, menunjukkan bahwa serangan dengan gangguan yang besar tidak sering terjadi.

Distribusi ini menunjukkan bahwa model ini cukup sensitif terhadap gangguan kecil, dengan sebagian besar noise tingkat rendah yang cukup untuk menghasilkan contoh lawan yang berhasil. Hal ini menunjukkan bahwa meskipun FGSM efektif dalam menyerang model, tindakan defensif seperti pelatihan *adversarial* atau mekanisme *Autoencoder* dapat difokuskan untuk menekan dampak gangguan kecil pada gambar. Secara keseluruhan, distribusi ini memberikan wawasan tentang pola derau musuh yang dihasilkan oleh FGSM dan sensitivitas model terhadap serangan tersebut.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

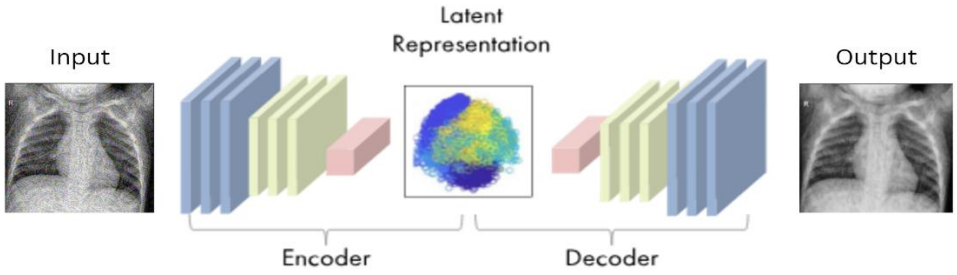


Gambar 11. Heatmap pada FGSM

Gambar 11 menunjukkan peta panas perbedaan piksel antara gambar asli dan gambar yang diserang oleh FGSM. Warna biru pada peta panas menunjukkan perbedaan piksel yang kecil, sedangkan warna merah menunjukkan perbedaan yang lebih besar. Dari visualisasi ini, dapat dilihat bahwa sebagian besar perbedaan piksel berada dalam kisaran kecil (diwakili oleh warna biru), yang mengindikasikan bahwa FGSM menambahkan gangguan yang halus dan hampir tidak terlihat pada penglihatan manusia. Area merah, yang tersebar di beberapa lokasi, menunjukkan piksel dengan tingkat gangguan yang lebih tinggi, yang mungkin terdapat pada fitur penting yang digunakan oleh model klasifikasi.

Pola noise yang dihasilkan menunjukkan bahwa FGSM secara efektif menargetkan fitur-fitur tertentu dalam gambar untuk memaksimalkan kesalahan prediksi, bahkan dengan noise yang minimal. Distribusi noise yang halus di seluruh gambar menegaskan bahwa FGSM dapat menciptakan contoh-contoh yang sulit dideteksi secara visual, namun tetap efektif menyerang model. Peta panas ini memberikan wawasan penting tentang bagaimana noise dimasukkan ke dalam gambar, yang merupakan dasar untuk mengevaluasi dan mengembangkan pertahanan seperti *Autoencoder*.

3.3 *Autoencoder*



Gambar 12. Skema *Autoencoder*

Arsitektur *Autoencoder* yang digunakan dalam penelitian ini terdiri dari dua komponen utama: encoder dan decoder, yang bekerja bersama-sama untuk merekonstruksi gambar yang bersih dari gambar yang terganggu. Encoder bertanggung jawab untuk mengompresi gambar input yang terganggu menjadi representasi laten, secara efektif mengurangi dimensinya sambil mempertahankan fitur-fitur penting yang diperlukan untuk rekonstruksi. Hal ini dicapai melalui serangkaian lapisan konvolusi yang mengekstrak pola-pola penting dari gambar sambil membuang noise yang tidak relevan. Representasi laten bertindak sebagai ringkasan ringkas dari input, memungkinkan model untuk fokus pada fitur-fitur utama. Di sisi lain, decoder merekonstruksi gambar dari ruang laten ini. Dekoder menggunakan lapisan konvolusi yang dialihkan untuk meningkatkan representasi terkompresi kembali ke dimensi gambar asli, memastikan bahwa detail halus dipertahankan. Selain itu, koneksi lompatan digabungkan antara encoder dan decoder untuk mempertahankan detail spasial dan meningkatkan akurasi rekonstruksi. Selama pelatihan, gambar yang berlawanan disediakan sebagai input ke *Autoencoder*, dengan gambar bersih yang sesuai

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

digunakan sebagai target. Fungsi kerugian Mean Squared Error (MSE) mengkuantifikasi kesalahan rekonstruksi, memandu model untuk menghilangkan gangguan yang berlawanan dengan tetap mempertahankan struktur gambar asli. Pengaturan ini memungkinkan *Autoencoder* untuk memulihkan gambar yang bersih dan berkualitas tinggi yang kuat terhadap serangan musuh.

Untuk mengatasi kerentanan model Xception terhadap serangan lawan, mekanisme pemulihan telah diimplementasikan dengan menggunakan *Autoencoder*. Tujuan utama dari *Autoencoder* adalah untuk mengurangi efek dari gangguan dan mengembalikan gambar yang terganggu ke kualitas aslinya, yang memungkinkan model Xception mendapatkan kembali akurasi prediktifnya. Dengan melatih *Autoencoder* pada contoh asli dan contoh yang berlawanan, *Autoencoder* belajar untuk mengidentifikasi dan menghilangkan noise yang berlawanan sambil mempertahankan fitur-fitur utama yang diperlukan untuk klasifikasi yang akurat.

Tabel 2. Arsitektur *Autoencoder*

Layer (Type)	Output Shape	Param #	Connected To
Input_layer_5 (InputLayer)	(None, 224, 224, 3)	0	-
Conv2d_11 (Conv2D)	(None, 224, 224, 32)	896	Input_layer_5[0]...
Max_pooling2d_7	(None, 112, 112, 32)	-	Conv2d_11[0][0]
Conv2d_12 (Conv2D)	(None, 112, 112, 64)	18,496	Max_pooling2d_7[...
Max_pooling2d_8	(None, 56, 56, 64)	0	Conv2d_12[0][0]
Conv2d_13 (Conv2D)	(None, 56, 56, 128)	73,856	Max_pooling2d_8[...
Flatten_1 (Flatten)	(None, 401408)	0	Conv2d_13[0][0]
Z_mean (Dense)	(None, 128)	51,380,352	Flatten_1[0][0]
Z_log_var (Dense)	(None, 128)	51,380,352	Flatten_1[0][0]
Z (Lambda)	(None, 128)	0	Z_mean[0][0] z-log_var[0][0]
Dense_9 (Dense)	(None, 401408)	51,781,632	Z[0][0]
Reshape (Reshape)	(None, 56, 56, 128)	0	Dense_9[0][0]
Conv2d_transpose (Conv2Dtranspose)	(None, 112, 112, 64)	73,792	Reshape[0][0]
Concatenate (Concatenate)	(None, 112, 112, 128)	0	Conv2d_traspose... Conv2d_12[0][0]
Conv2d_transpose_1 (Conv2Dtranspose)	(None, 224, 224, 32)	36,896	Concatenate[0][0]
Concatenate_1 (Concatenate)	(None, 224, 224, 64)	0	Conv2d_transpose... Conv2d_11[0][0]
Conv2d_14 (Conv2D)	(None, 224, 224, 3)	1,731	Concatenate_1[0]...

Proses pemulihan Autoencoder menggunakan arsitektur *encoder/decoder* ganda untuk menghilangkan *noise* yang mengganggu :

- Encoder: Gambar input (224x224x3) diproses melalui serangkaian lapisan konvolusi (Conv2D) dan lapisan max-pooling, yang secara progresif mengurangi dimensi spasial sekaligus

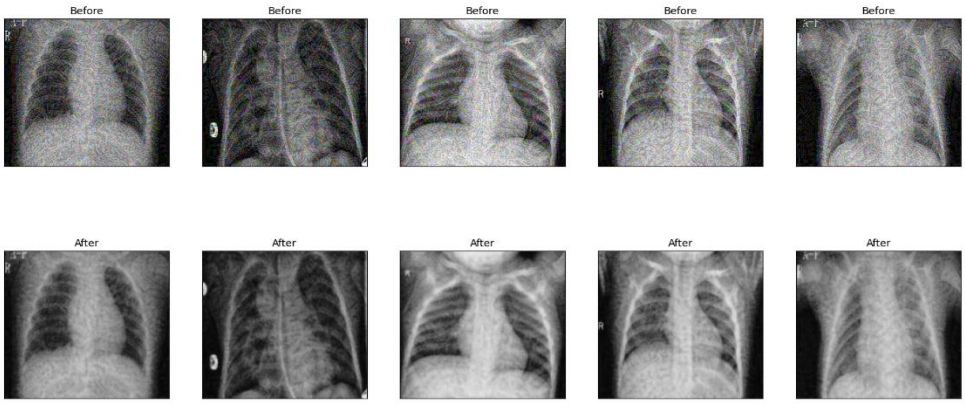
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

meningkatkan kedalaman fitur. Representasi akhir yang diratakan diteruskan ke dua lapisan padat (z_mean dan z_log_var) untuk membuat representasi ruang laten dengan ukuran 128.

- Ruang Laten: Lapisan sampling digunakan untuk mereparameterisasi ruang laten, memperkenalkan variabilitas yang terkendali untuk rekonstruksi.
- Dekoder: Representasi laten dilewatkan melalui serangkaian lapisan konvolusi yang dialihkan untuk meningkatkan sampel dan merekonstruksi gambar kembali ke dimensi aslinya (224x224x3). Koneksi lompatan digunakan untuk menggabungkan fitur dari encoder, meningkatkan akurasi rekonstruksi.

Input dikompresi oleh encoder ke dalam ruang laten berdimensi lebih rendah, mempertahankan fitur yang paling penting sambil membuang noise. Decoder merekonstruksi gambar dari representasi laten ini, dengan tujuan untuk mengembalikan kondisi sebelum gangguan yang bersih. Selama pelatihan, gambar yang berlawanan digunakan sebagai input, sedangkan gambar asli yang sesuai berfungsi sebagai target. Autoencoder terdiri dari tiga lapisan konvolusi (Conv2D: 32, 64, 128 filter), diikuti max pooling, dan Dense layer menuju latent space berukuran 128. Decoder menggunakan transposed convolution hingga citra ukuran 224x224x3. Fungsi aktivasi ReLU digunakan di semua lapisan kecuali output (sigmoid). Proses pelatihan mengandalkan loss MSE, optimizer Adam dengan learning rate 0.001, dan EarlyStopping jika val_loss tidak membaik selama 5 epoch. Autoencoder dinyatakan konvergen jika MSE rata-rata < 0.17. Proses ini dipandu oleh fungsi kerugian Mean Squared Error (MSE) yang mengukur kualitas rekonstruksi dan meminimalkan perbedaan antara gambar input dan target.

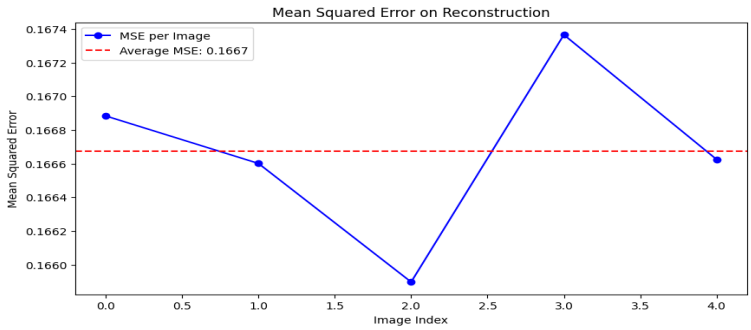
Arsitektur ini mencakup beberapa lapisan konvolusi dalam encoder, yang mengurangi dimensi spasial untuk mengekstrak fitur, dan konvolusi yang ditransposisikan dalam decoder, yang mengambil sampel dan merekonstruksi dimensi asli. Koneksi lompatan menghubungkan lapisan encoder dan decoder untuk mempertahankan detail halus selama rekonstruksi Model ini dioptimalkan menggunakan pengoptimal Adam (tingkat pembelajaran: 0,001) dan menggabungkan EarlyStopping untuk menghentikan pelatihan ketika kehilangan validasi tidak lagi membaik. Hal ini mencegah *overfitting* dan memastikan generalisasi. *Autoencoder* yang terlatih berhasil menghilangkan noise yang tidak diinginkan dan menghasilkan gambar yang, ketika dijalankan melalui model Xception, secara signifikan mengembalikan akurasi klasifikasi.



Gambar 13. Sebelum dan Sesudah Recovery dengan Autoencoder pada X-Ray Dada

Visualisasi di atas mengilustrasikan keefektifan *Autoencoder* dalam memulihkan gambar X-Ray yang terganggu. Baris atas menunjukkan gambar input setelah ditambahkan noise, yang memperkenalkan distorsi yang dapat menyesatkan model klasifikasi. Baris bawah menunjukkan gambar yang direkonstruksi setelah diproses melalui *Autoencoder*. *Autoencoder* berhasil menghilangkan noise dan mengembalikan detail struktural dari gambar X-Ray, menunjukkan kemampuannya untuk memulihkan gambar yang bersih dan mengurangi efek dari serangan musuh.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

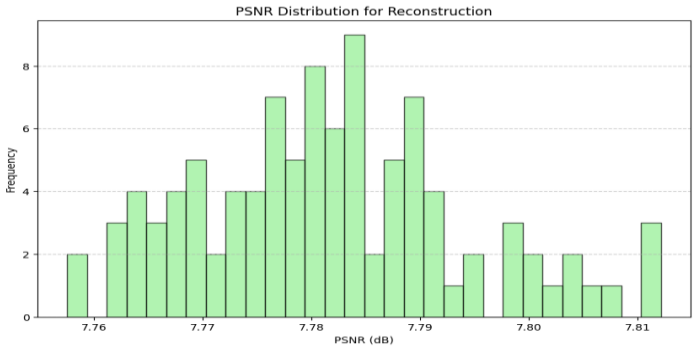


Gambar 14. Mean Squared Error pada Autoencoder

Grafik di atas menunjukkan nilai *Mean Squared Error* (MSE) untuk setiap gambar yang direkonstruksi oleh auto-encoder, dengan garis biru menunjukkan MSE per gambar dan garis merah putus-putus menunjukkan nilai rata-rata MSE sebesar 0,1667. Dapat dilihat dari grafik bahwa nilai MSE untuk setiap gambar bervariasi dalam rentang yang kecil, pada gambar dengan indeks 2, nilai MSE mencapai titik terendah, yang menunjukkan bahwa *Autoencoder* mampu merekonstruksi gambar ini dengan akurasi yang tinggi. Sebaliknya, gambar dengan indeks 3 memiliki MSE tertinggi, mengindikasikan rekonstruksi yang sedikit kurang akurat daripada gambar lainnya. Namun, terlepas dari variasi tersebut, semua nilai MSE tetap mendekati rata-rata, yang mencerminkan kinerja yang konsisten dari *Autoencoder* dalam merekonstruksi gambar dari gangguan lawan. hasil ini menunjukkan bahwa *Autoencoder* mampu menghilangkan sebagian besar noise dari gambar yang diserang, dengan tingkat kesalahan yang rendah dan konsisten di seluruh kumpulan data. Namun, analisis lebih lanjut pada gambar dengan MSE tinggi dapat membantu untuk memahami faktor-faktor yang mempengaruhi kualitas rekonstruksi dalam beberapa kasus tertentu.

Tabel 3. Perbandingan Akurasi Model Xception pada Tiga Kondisi

Nilai ϵ	Akurasi Data Asli	Setelah Serangan FGSM	Setelah Recovery (Autoencoder)
0.01	93%	80.4%	91.2%
0.03	93%	65.7%	88.1%
0.05	93%	54.2%	84.5%

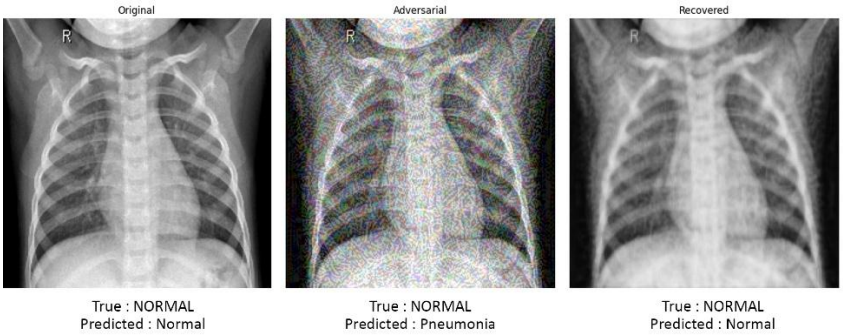


Gambar 15. *Peak Signal-to-Noise Ratio Distribution* pada *Autoencoder*

Gambar 15 menunjukkan distribusi *Peak Signal-to-Noise Ratio* (PSNR) untuk gambar rekonstruksi yang dihasilkan oleh *Autoencoder*. Sebagian besar nilai PSNR terdistribusi antara 7,76 dB dan 7,81 dB, dengan frekuensi puncak pada kisaran 7,78 dB hingga 7,79 dB, yang mengindikasikan bahwa sebagian besar gambar yang direkonstruksi memiliki kualitas yang sama dengan gambar asli dalam kisaran yang konsisten. Variasi frekuensi menunjukkan bahwa ada

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

beberapa gambar yang memiliki kualitas rekonstruksi yang sedikit lebih tinggi atau lebih rendah dari rata-rata, tetapi perbedaannya relatif kecil. Distribusi yang sempit ini mengindikasikan bahwa *Autoencoder* mampu merekonstruksi gambar dengan kualitas yang stabil di seluruh dataset. Namun, nilai PSNR yang berada di kisaran tengah sekitar 7,7 dB hingga 7,8 dB mengindikasikan bahwa masih ada ruang untuk perbaikan, terutama untuk meningkatkan akurasi rekonstruksi agar lebih mendekati citra asli. Secara keseluruhan, distribusi ini menunjukkan bahwa *Autoencoder* memberikan hasil rekonstruksi yang konsisten dengan mempertahankan detail penting pada Medical Image sekaligus mengurangi noise, meskipun kualitas rekonstruksi masih bisa ditingkatkan untuk mencapai nilai PSNR yang lebih tinggi.



Gambar 16. Gambar X-Ray Dada Setelah Diserang (FGSM) dan Setelah Dipulihkan (*Autoencoder*)

Untuk memastikan bahwa peningkatan akurasi yang diperoleh setelah pemulihan oleh Autoencoder tidak terjadi secara acak, dilakukan uji statistik menggunakan *paired t-test*. Uji ini membandingkan nilai akurasi model Xception pada dataset FGSM dan dataset hasil recovery. Hasil pengujian menunjukkan bahwa **nilai p < 0.05**, yang berarti terdapat perbedaan signifikan secara statistik antara akurasi sebelum dan sesudah pemulihan. Dengan demikian, Autoencoder tidak hanya mengembalikan struktur visual gambar, tetapi juga secara nyata meningkatkan performa klasifikasi.

Visualisasi ini menampilkan perkembangan gambar X-Ray dada melalui tahapan serangan dan pemulihan. Gambar pertama di sebelah kiri mewakili X-Ray asli, yang menunjukkan struktur paru-paru yang tidak berubah dan jelas, yang berfungsi sebagai garis dasar untuk klasifikasi. Gambar tengah menggambarkan X-Ray yang terganggu yang dibuat menggunakan serangan FGSM. Meskipun noise tampak halus, namun cukup mengganggu fitur gambar sehingga berpotensi menyesatkan model ke dalam kesalahan klasifikasi. Gambar akhir di sebelah kanan mengilustrasikan X-Ray yang dipulihkan setelah diproses oleh *Autoencoder*. *Autoencoder* berhasil menghilangkan noise yang mengganggu, merekonstruksi gambar lebih dekat ke kondisi aslinya sambil mempertahankan detail struktural yang penting untuk klasifikasi yang akurat.

4. KESIMPULAN

Penelitian ini membuktikan bahwa Autoencoder efektif dalam mengurangi dampak serangan FGSM pada citra medis, bahkan dapat memulihkan akurasi model klasifikasi hingga kembali di atas 90%. Dengan mengadopsi kerangka kerja SEMMA, studi ini menunjukkan bahwa serangan adversarial dengan gangguan kecil sekalipun dapat menurunkan performa model secara signifikan, namun dapat ditanggulangi melalui pemulihan berbasis Autoencoder. Kontribusi kebaruan dari penelitian ini terletak pada eksplorasi spesifik terhadap kombinasi serangan FGSM dan pertahanan *Autoencoder* dalam domain X-ray medis, serta analisis sensitivitas multi-ε yang jarang dilakukan dalam studi sebelumnya. Penelitian ini juga memberikan evaluasi kuantitatif yang diperkuat dengan uji statistik (paired t-test) untuk memastikan peningkatan akurasi yang signifikan secara matematis.

Batasan studi ini meliputi fokus pada satu jenis serangan (FGSM) dan belum diuji terhadap teknik adversarial yang lebih kompleks seperti PGD atau DeepFool. Selain itu, pemrosesan Autoencoder membutuhkan sumber daya komputasi yang tinggi, sehingga implementasinya di lingkungan klinis real-time masih menjadi tantangan teknis. Sebagai rekomendasi teknis, Autoencoder dapat diintegrasikan ke dalam pipeline klasifikasi sebagai lapisan praproses otomatis yang membersihkan citra input sebelum dianalisis oleh model utama. Pendekatan ini dapat meningkatkan ketahanan sistem terhadap serangan

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

musuh tanpa mengubah arsitektur model inti. Selain itu, studi lanjutan sebaiknya mempertimbangkan strategi pertahanan ensemble dan pengujian lintas domain untuk memperluas penerapannya secara klinis.

DAFTAR PUSTAKA

[1] A. S. Chauhan, R. Singh, N. Priyadarshi, B. Twala, S. Suthar, and S. Swami, "Unleashing the power of advanced technologies for revolutionary medical imaging: pioneering the healthcare frontier with artificial intelligence," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 58, Aug. 2024, doi: 10.1007/s44163-024-00161-0.

[2] L. Pinto-Coelho, "How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications," Dec. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/bioengineering10121435.

[3] A. E. Putra, K. Kartini, and A. P. Sari, "Metode Convolutional Neural Network dan Extreme Gradient Boost untuk Mengklasifikasi Penyakit Pneumonia," *JASIEK (Jurnal Aplikasi Sains, Informasi, Elektronika dan Komputer)*, vol. 6, no. 1, pp. 33–40, Jul. 2024, doi: 10.26905/jasiek.v6i1.11464.

[4] J. Al-Jaroodi, N. Mohamed, and E. Abukhousa, "Health 4.0: On the Way to Realizing the Healthcare of the Future," *IEEE Access*, vol. 8, pp. 211189–211210, 2020, doi: 10.1109/ACCESS.2020.3038858.

[5] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine Learning for Medical Imaging," *RadioGraphics*, vol. 37, no. 2, pp. 505–515, Mar. 2017, doi: 10.1148/rg.2017160130.

[6] M. Adelusola, "Integrated AI Solutions: From Combating Fake News to Revolutionizing Healthcare and VLSI." [Online]. Available: https://www.researchgate.net/publication/387018390

[7] J. Zhang and C. Li, "Adversarial Examples: Opportunities and Challenges," *IEEE Trans Neural Netw Learn Syst*, vol. 31, no. 7, pp. 2578–2593, Jul. 2020, doi: 10.1109/TNNLS.2019.2933524.

[8] L. Wu, Z. Zhu, C. Tai, and W. E, "Understanding and Enhancing the Transferability of Adversarial Examples," Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.09707

[9] S. A. Khowaja, I. H. Lee, K. Dev, M. A. Jarwar, and N. M. F. Qureshi, "Get your Foes Fooled: Proximal Gradient Split Learning for Defense against Model Inversion Attacks on IoMT data," Jan. 2022, doi: 10.1109/TNSE.2022.3188575.

[10] S. M. A. Naqvi, M. Shabaz, M. A. Khan, and S. I. Hassan, "Adversarial Attacks on Visual Objects Using the Fast Gradient Sign Method," *J Grid Comput*, vol. 21, no. 4, p. 52, Dec. 2023, doi: 10.1007/s10723-023-09684-9.

[11] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective," *ACM Comput Surv*, vol. 55, no. 1, pp. 1–38, Jan. 2023, doi: 10.1145/3485133.

[12] S. B. Weber, S. Stein, M. Pilgermann, and T. Schrader, "Attack Detection for Medical Cyber-Physical Systems–A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 41796–41815, 2023, doi: 10.1109/ACCESS.2023.3270225.

[13] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: a survey," *Artif Intell Rev*, vol. 57, no. 2, p. 28, Feb. 2024, doi: 10.1007/s10462-023-10662-6.

[14] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, IEEE, Apr. 2021, pp. 13–24. doi: 10.1109/ICDE51399.2021.00009.

[15] G. Spigler, "Denoising Autoencoders for Overgeneralization in Neural Networks," Sep. 2017, doi: 10.1109/TPAMI.2019.2909876.