

IMPLEMENTASI TRANSFER LEARNING DAN AUGMENTASI PADA MODEL CNN UNTUK MENINGKATKAN ROBUSTNESS KLASIFIKASI AI-GENERATED IMAGE

Ni Kadek Ratih Ranupadma, Eddy Muntina Dharma, Anak Agung Istri Ita Paramitha

Sistem Informasi, Universitas Primakara

JL. Tukad Badung No. 135 Renon, Denpasar-Bali, 80226, Indonesia

ratihpadma04@gmail.com

ABSTRAK

Perkembangan teknologi *AI-generated image* menghadirkan tantangan baru dalam membedakan gambar asli dan gambar sintetis, terutama dalam konteks keamanan dan keaslian konten digital. Berbagai pendekatan deteksi telah dikembangkan menggunakan *Convolutional Neural Network* (CNN), namun sebagian besar masih diuji pada gambar dalam kondisi bersih. Padahal, gambar yang beredar di dunia nyata sering mengalami gangguan visual seperti *noise*, yang dapat merusak ciri halus pembeda antara gambar asli dan sintetis. Penelitian ini bertujuan menganalisis sejauh mana penerapan *transfer learning* dan augmentasi data dapat meningkatkan performa serta ketahanan (*robustness*) model CNN dalam mendeteksi gambar *AI-generated* pada kondisi terdegradasi. Dataset yang digunakan berasal dari *Kaggle* dengan dua kelas, yaitu gambar asli dan *AI-generated*. Dua skema eksperimen diterapkan, yaitu pelatihan CNN dari awal (*baseline*) dan model *pre-trained* menggunakan arsitektur *NYUAD-ComNets/AI-Generated Images Detector* yang dikombinasikan dengan augmentasi data. Evaluasi *robustness* dilakukan dengan menambahkan *Gaussian Noise* dan *Salt and Pepper Noise* pada berbagai tingkat keparahan. Hasil menunjukkan model *pre-trained* mencapai akurasi 0,83 pada data bersih, menurun menjadi 0,75 pada *Gaussian Noise* dan 0,63 pada *Salt and Pepper Noise*. Nilai AUC turun dari 0,83 menjadi 0,78 dan 0,74. Temuan ini menunjukkan bahwa karakteristik dan tingkat keparahan *noise* berpengaruh signifikan terhadap ketahanan model.

Kata Kunci: *Transfer Learning, Convolutional Neural Network, Augmentasi, Noise, Robustness*

1. PENDAHULUAN

Perkembangan teknologi *AI-generated image* semakin pesat, dengan model generatif seperti GAN dan *diffusion model* yang mampu menghasilkan gambar sangat realistis hingga sulit dibedakan dari gambar asli oleh manusia. Kondisi ini menimbulkan tantangan dalam menjaga keaslian konten visual, khususnya pada bidang keamanan digital dan forensik media. Untuk mengatasi hal tersebut, berbagai pendekatan telah dikembangkan, salah satunya menggunakan *Convolutional Neural Network* (CNN) sebagai model deteksi. Sejumlah penelitian menunjukkan bahwa CNN mampu mengenali artefak tersembunyi dan pola visual tertentu yang menjadi ciri khas gambar sintetis, sehingga memberikan performa deteksi yang tinggi pada berbagai jenis model generatif [1].

Meskipun demikian, sebagian besar pengembangan model deteksi masih dilakukan pada gambar dalam kondisi bersih dan ideal. Pada kenyataannya, gambar yang beredar di dunia nyata umumnya telah mengalami berbagai proses seperti kompresi, perubahan resolusi, penurunan kualitas, blur, maupun penambahan *noise* akibat proses unggah dan unduh di platform digital [2]. Padahal, deteksi gambar *AI-generated* sangat bergantung pada ciri visual yang halus, seperti pola tekstur mikro dan karakteristik frekuensi tertentu, yang justru mudah rusak akibat degradasi tersebut. Hal ini membuat performa model berpotensi menurun ketika diterapkan

pada gambar yang telah mengalami gangguan kualitas [3].

Berbeda dengan klasifikasi objek pada umumnya yang masih dapat mengandalkan bentuk utama objek, deteksi gambar sintetis berfokus pada perbedaan visual yang jauh lebih halus antara gambar asli dan gambar hasil *generated AI* [4]. Perbedaan tersebut dapat berupa ketidakwajaran pada detail kecil, inkonsistensi tekstur, atau pola statistik tertentu yang sulit dikenali oleh mata manusia. Karena sifatnya yang sensitif, ciri-ciri ini mudah berubah atau hilang ketika gambar mengalami *noise*, blur, maupun kompresi, sehingga ketahanan model terhadap gangguan visual menjadi aspek penting yang perlu diperhatikan [5].

Penelitian ini berfokus pada analisis ketahanan model CNN dalam mendeteksi gambar *AI-generated* pada kondisi yang tidak ideal. Evaluasi dilakukan dengan mensimulasikan berbagai bentuk degradasi visual yang merepresentasikan kondisi distribusi gambar digital di dunia nyata. Untuk mendukung kemampuan model dalam menghadapi variasi tersebut, digunakan pendekatan *transfer learning* agar model dapat memanfaatkan fitur visual umum yang telah dipelajari sebelumnya, serta augmentasi data untuk memperkaya variasi kondisi gambar selama pelatihan [6]. Melalui pendekatan ini, diharapkan diperoleh pemahaman yang lebih mendalam mengenai perilaku model deteksi pada kondisi terdegradasi sekaligus menghasilkan model yang lebih robust untuk penerapan di dunia nyata.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian terdahulu menjadi salah satu acuan penting dalam penelitian ini karena memberikan gambaran mengenai perkembangan metode deteksi gambar *AI-generated* menggunakan *deep learning*. Meskipun tidak ditemukan penelitian dengan judul yang sama, beberapa studi memiliki keterkaitan kuat, terutama dalam pemanfaatan model *pre-trained* untuk meningkatkan kemampuan generalisasi dan ketahanan model. Salah satu yang relevan adalah penelitian oleh Laura Gaintseva berjudul “*Improving Interpretability and Robustness for the Detection of AI-Generated Images*” [7] yang mengusulkan pendekatan berbasis CLIP dengan analisis embedding, penghapusan fitur yang tidak relevan, serta pemilihan *attention heads* untuk meningkatkan interpretabilitas dan *robustness* model. Hasilnya menunjukkan peningkatan akurasi klasifikasi *out-of-distribution* hingga 6%, menandakan model menjadi lebih tangguh terhadap variasi data dari generator yang berbeda.

Penelitian tersebut memiliki persamaan dengan penelitian ini dalam hal fokus pada peningkatan generalisasi dan ketahanan model dalam mendeteksi gambar *AI-generated*. Namun, penelitian ini lebih menekankan pada evaluasi *robustness* model CNN terhadap gangguan visual berupa *noise* dengan tingkat keparahan bertingkat, serta mengombinasikan *transfer learning* dan augmentasi data sebagai strategi peningkatan performa.

2.2. Deep Learning

Deep Learning adalah cabang dari machine learning yang memanfaatkan artificial neural network yang berlapis-lapis layaknya otak manusia dimana neuron terkoneksi satu sama lain untuk membuat sebuah jaringan [8]. Model Pembelajaran deep learning ini dapat mempelajari sebuah pola kompleks pada data untuk mendeteksi atau memproses sebuah gambar, teks, audio dan data lainnya untuk menghasilkan wawasan atau prediksi yang akurat [9]

2.3. CNN (Convolutional Neural Network)

Convolutional Neural Network (CNN) adalah arsitektur deep learning yang sering digunakan mempelajari fitur-fitur kompleks dari suatu objek, khususnya dalam bentuk spasial seperti gambar, sehingga lebih efisien dalam mengenali gambar [10]. Arsitektur CNN terdiri dari beberapa lapisan pemrosesan yang dapat mempelajari berbagai fitur masukan data, misalnya gambar dengan tingkat abstraksi yang berbeda [11].

- Lapisan awal mempelajari fitur dasar (abstraksi rendah), seperti garis dan warna
- Lapisan lebih dalam mempelajari fitur yang lebih kompleks (abstraksi tinggi), seperti bentuk dan pola tertentu

CNN terdiri dari beberapa lapisan (layers) yang bekerja untuk mengenali fitur dalam sebuah gambar, yakni di antaranya:

- Convolutional Layer*, Lapisan ini menggunakan fitur (kernel) untuk mengekstraksi fitur dari gambar input
- Pooling layer*, satu metode untuk mengurangi ukuran matriks data dimana lapisan *pooling* ini akan mengekstrak ukuran spasial dari fitur konvolusi, sehingga dapat mengurangi kebutuhan sumber daya komputasi dengan mereduksi dimensi dari fitur map [12]
- Activation layer*, fungsi aktivasi memutuskan apakah neuron akan mengembalikan input tertentu dengan menghasilkan output yang sesuai [11]
- Dropout*, Teknik ini mendorong model untuk menemukan cara yang lebih robust dalam membuat prediksi tanpa bergantung pada fitur atau neuron dalam data pelatihan [13]

2.4. Transfer Learning

Transfer learning adalah proses memanfaatkan pengetahuan yang telah diperoleh oleh suatu model dari tugas sebelumnya yang memiliki jumlah data besar, lalu menerapkannya pada tugas baru dengan datasets yang lebih terbatas [6]. sebuah model telah dilatih pada dataset yang cukup besar dan bersifat umum, maka model tersebut dapat berfungsi sebagai representasi visual yang efektif untuk berbagai tugas lainnya. Ada tiga pendekatan pada transfer learning yaitu sebagai classifier, *feature extraction* dan *fine tuning*. Menggunakan metode *transfer learning* sebagai *classifier* berarti memanfaatkan jaringan yang telah dilatih untuk masalah serupa dan langsung diterapkan pada model baru. Metode *transfer learning* sebagai *feature extraction* diterapkan ketika tugas baru memiliki kemiripan dengan dataset asli yang digunakan untuk melatih jaringan sebelumnya. Sementara itu, metode *transfer learning* sebagai *fine tuning* dilakukan dengan mengekstrak *feature map* dari model sebelumnya, lalu menyempurnakannya agar lebih sesuai dengan domain target [14].

2.5. Augmentasi

Augmentasi adalah teknik yang digunakan untuk menambah jumlah dan variasi dalam dataset pelatihan dengan cara menciptakan versi modifikasi dari data yang sudah ada [15]. Modifikasi ini bisa berupa perubahan kecil seperti rotasi, *flipping*, perubahan warna, penambahan *noise*, penskalaan atau transformasi lainnya yang tidak mengubah makna dasar dari data tersebut [16].

2.6. Robustness

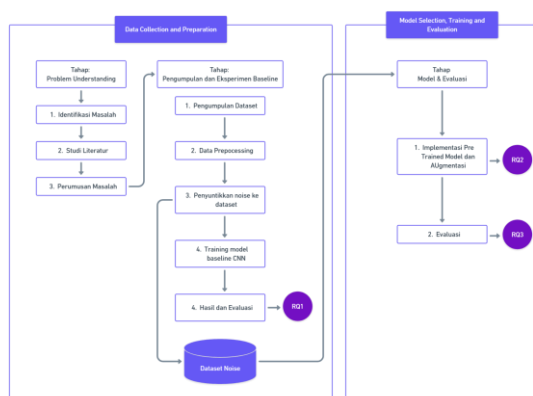
Secara umum, *Robustness* dapat didefinisikan sebagai kemampuan sistem atau model yang dapat mempertahankan kinerja yang stabil walaupun terdapat gangguan dan perubahan yang terjadi dalam kondisi lingkungan [17]. Dalam konteks CNN, *robustness* merujuk pada kemampuan untuk memberikan prediksi yang akurat walau terdapat

gangguan yang dipengaruhi oleh beberapa faktor, misalnya [18]:

- Variasi dalam data input yang memiliki pendistribusian data yang berbeda dengan data pelatihan, misalnya gambar yang buram atau terbalik
- Kehilangan data atau beberapa fitur yang tidak tersedia selama penggunaan model
- Gangguan eksternal seperti gangguan noise yang dapat mempengaruhi kualitas data dan mempengaruhi tingkat akurasi prediksi.

3. METODE PENELITIAN

Penelitian ini menggunakan metodologi CRISP-DM (Cross Industry Standard Process for Data Mining) sebagai kerangka kerja utama. CRISP-DM dipilih karena menyediakan alur penelitian yang sistematis dan terstruktur, serta telah banyak digunakan dalam penelitian di bidang data science, machine learning, dan kecerdasan buatan[19]. Metodologi ini memungkinkan proses penelitian berjalan secara terkontrol mulai dari perumusan masalah hingga evaluasi model. Tahapan CRISP-DM yang diterapkan dalam penelitian ini meliputi *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, dan *Evaluation*.



Gambar 1. Alur Penelitian

3.1. Business Understanding

Tahap ini berfokus pada perumusan permasalahan penelitian, yaitu meningkatnya peredaran gambar *AI-generated* yang secara visual semakin sulit dibedakan dari gambar asli. Model deteksi yang ada masih menunjukkan keterbatasan ketika dihadapkan pada kondisi tidak ideal, khususnya saat gambar mengalami gangguan visual (*noise*). Oleh karena itu, penelitian ini bertujuan untuk meningkatkan performa dan *robustness* model CNN melalui penerapan *transfer learning* dan augmentasi data.

3.2. Data Understanding

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari platform Kaggle, yaitu dataset “*AI Generated Images vs Real*

Images” yang disusun oleh Cash Bowman. Dataset ini terdiri dari 1.069 gambar berformat JPG yang terbagi ke dalam dua kelas, yaitu gambar asli dan gambar *AI-generated*. Data kemudian dibagi menjadi data latih, validasi, dan uji dengan rasio 70:20:10.

3.3. Data Preparation

Pada tahap ini dilakukan proses prapemrosesan data untuk memastikan keseragaman dan kesiapan data sebelum pemodelan. Seluruh gambar diubah ke ukuran 256×256 piksel, dinormalisasi, dan dikonversi ke dalam bentuk tensor agar sesuai dengan masukan model CNN. Selain data bersih, disiapkan pula data uji bergangguhan melalui proses injeksi *noise* dengan tingkat keparahan (*severity*) 1 hingga 10.

3.4. Injeksi Noise dan Severity

Injeksi noise dilakukan untuk merepresentasikan kondisi degradasi visual yang umum terjadi di dunia nyata. Penelitian ini menggunakan pendekatan *severity* bertingkat (1–10) yang dibagi menjadi tiga kategori, yaitu *severity* rendah (1–3), menengah (4–7), dan tinggi (8–10). Pada *Gaussian noise*, *severity* rendah (1–3) menggambarkan fluktuasi pencahayaan ringan atau keterbatasan sensor kamera. *Severity* menengah (4–7) merepresentasikan kondisi pencahayaan tinggi atau *noise* sensor yang mulai signifikan, sedangkan *severity* tinggi (8–10) menggambarkan kondisi pencahayaan ekstrem atau lingkungan dengan tingkat *noise* sensor yang sangat tinggi. Pada *Salt and Pepper noise*, *severity* rendah (1–3) menggambarkan gangguan kecil pada proses transmisi atau penyimpanan data. *Severity* menengah (4–7) merepresentasikan gangguan digital yang lebih sering terjadi, sedangkan *severity* tinggi (8–10) menggambarkan kerusakan piksel yang signifikan seperti hasil pemindaian ulang, tangkapan layar, atau proses digitalisasi berulang. Pada *Gaussian blur*, *severity* rendah (1–3) menggambarkan ketidaktepatan fokus atau pergerakan kamera yang ringan. *Severity* menengah (4–7) menunjukkan penurunan ketajaman gambar yang mulai mengaburkan detail visual, sedangkan *severity* tinggi (8–10) merepresentasikan kondisi gambar yang sangat buram sehingga detail penting sulit dikenali. Pada *JPEG compression*, *severity* rendah (1–3) umumnya terjadi akibat kompresi ringan seperti unggahan satu kali ke platform digital. *Severity* menengah (4–7) menggambarkan kompresi berulang yang menyebabkan degradasi kualitas visual secara bertahap, sedangkan *severity* tinggi (8–10) merepresentasikan kompresi agresif yang menghilangkan detail visual penting. Pada *color shift*, *severity* rendah (1–3) menggambarkan perbedaan pencahayaan atau pengaturan warna yang masih ringan. *Severity* menengah (4–7) menunjukkan pergeseran warna yang mulai memengaruhi tampilan visual secara keseluruhan, sedangkan *severity* tinggi (8–10) merepresentasikan perubahan warna yang signifikan akibat manipulasi warna atau perbedaan ekstrem antar perangkat tampilan.

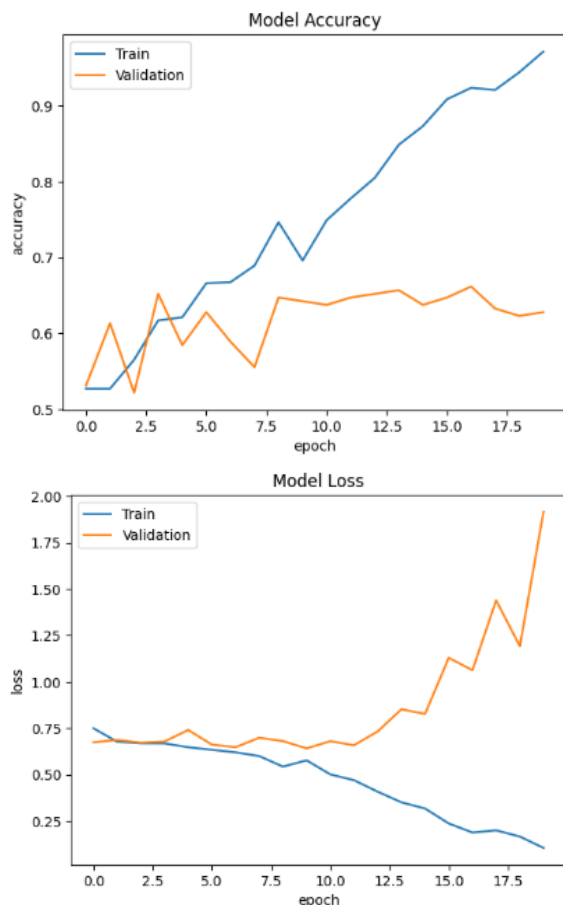
3.5. Modeling

Tahap pemodelan melibatkan pengujian dua pendekatan model CNN. Model pertama adalah CNN *baseline* yang dilatih dari awal menggunakan arsitektur sederhana. Model kedua adalah model *pre-trained NYUAD-ComNets/AI-Generated Images Detector* yang diterapkan dengan pendekatan transfer learning. Pada model *pre-trained*, layer *backbone* dibekukan dan hanya layer klasifikasi akhir yang dilatih ulang. Eksperimen dilakukan dalam dua skema, yaitu pelatihan menggunakan data bersih dan pengujian pada data bergangguan, serta pelatihan menggunakan data beraugmentasi dan pengujian pada data bersih dan bergangguan.

3.6. Evaluation

Evaluasi model dilakukan menggunakan metrik akurasi dan ROC-AUC untuk mengukur performa klasifikasi. *Robustness* model dianalisis berdasarkan tren perubahan performa terhadap peningkatan *severity noise*. Model dikatakan *robust* apabila mampu mempertahankan performa klasifikasi secara relatif stabil meskipun tingkat gangguan visual meningkat.

4. HASIL DAN PEMBAHASAN



Gambar 2. Grafik Learning Curve Baseline CNN

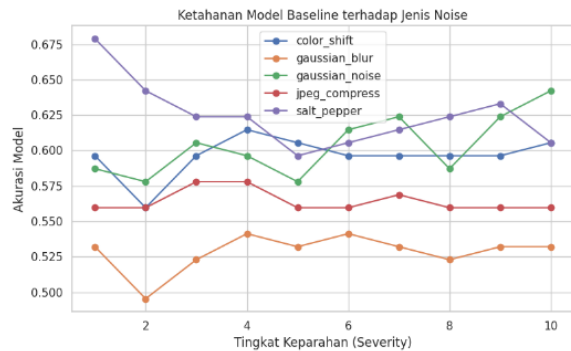
Penelitian ini membandingkan dua pendekatan model, yaitu *baseline CNN* yang dilatih dari awal dan model *pre-trained NYUAD-ComNets/AI-Generated Images Detector* yang di-fine-tune menggunakan dataset penelitian. Evaluasi dilakukan pada dua kondisi, yaitu data bersih dan data dengan gangguan visual berupa *Gaussian Noise* dan *Salt and Pepper Noise* pada tingkat keparahan 1–10. Analisis ini dilakukan dengan cara membangun model *baseline CNN from scratch* menggunakan dataset *AI Generated vs Real Images* yang diperoleh dari Kaggle dimana dataset ini terdiri atas 1.069 gambar dengan dengan rasio pembagian 70:20:10 yang menghasilkan 734 gambar untuk pelatihan, 207 gambar untuk validasi dan 109 gambar untuk pengujian.

Hasil pelatihan menunjukkan bahwa model *baseline* mencapai akurasi sebesar 0.98 (98%) dan akurasi validasi cenderung stagnan di kisaran 0.6618 (66%) yang menunjukkan adanya *overfitting* pada model tersebut. Namun pada *epoch* ke-5, nilai akurasi validasi berada pada titik paling optimal sebelum mulai menurun, sehingga *epoch* ke-4 dipilih sebagai model terbaik (*best checkpoint*) untuk digunakan pada tahap pengujian (*testing*). Tahap berikutnya adalah penyuntikan *noise* pada dataset uji untuk melihat dampak setiap gangguan terhadap performa model. Penyuntikan dilakukan dengan parameter yang telah disesuaikan agar menyerupai kondisi gangguan nyata, seperti blur optik, artefak kompresi dan perubahan warna. Setiap diterapkan pada seluruh data uji berjumlah 109 gambar, kemudian model diuji ulang untuk menghitung perubahan akurasi. Hasil pengujian menunjukkan bahwa setiap jenis *noise* memiliki tingkat dampak berbeda terhadap performa model *baseline*, sebagaimana ditunjukkan pada tabel berikut:

Tabel 1. Hasil Pengujian Noise pada Dataset Uji

Jenis Noise	Severity	Akurasi baseline	Akurasi akhir	Drop akurasi
Gaussian Noise	10	0.6618	0.50	-0.1618
Salt & Pepper	10	0.6618	0.50	-0.1618
Gaussian Blur	10	0.6618	0.57	-0.0918
JPEG Compression	10	0.6618	0.61	-0.0518
Color Shift	10	0.6618	0.6	-0.0618

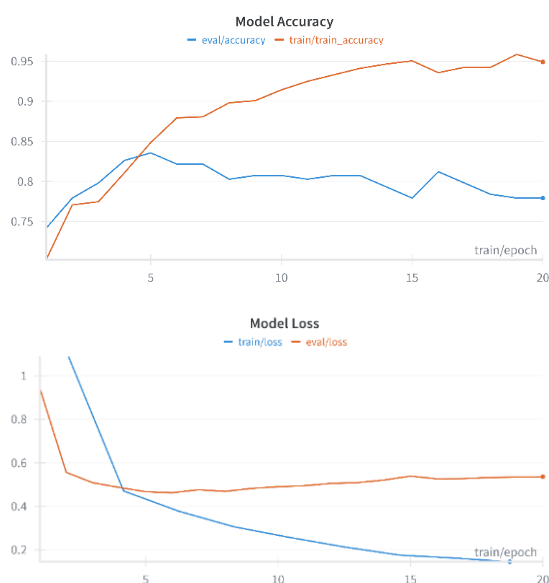
Hasil pengujian menunjukkan bahwa jenis *noise* dengan pola acak pada piksel, seperti *Gaussian Noise* dan *Salt & Pepper* memberikan dampak paling besar terhadap deteksi model CNN. Kedua jenis gangguan ini menyebabkan penurunan akurasi sekitar 0-16 poin persentase dari kondisi awal dimana artinya akurasi turun 16.18 angka persen secara absolut, bukan turun 16% dari performa awal.



Gambar 3. Ketahanan Model Baseline terhadap Noise

Grafik pada gambar 3 memperlihatkan bahwa *Gaussian Noise* dan *Salt & Pepper* memiliki nilai tertinggi, masing-masing dengan penurunan akurasi sekitar 10 poin persentase (pp) dari *baseline*. Hal ini menunjukkan bahwa gangguan acak pada intensitas piksel merupakan bentuk *noise* yang paling merusak performa CNN karena secara langsung mengacaukan informasi tekstur dan pola tepi citra. Setelah model *baseline* CNN menunjukkan penurunan akurasi cukup besar ketika diuji pada data *noise*, tahap ini berfokus pada penerapan metode peningkatan performa agar model menjadi lebih stabil terhadap variasi citra dan *noise*.

Model yang digunakan pada eksperimen ini adalah NYUAD AI-Generated Images Detector yaitu model pre-trained yang dikembangkan oleh NYUAD-ComNets dan tersedia melalui *Hugging Face Transformers*. Dalam penelitian ini, augmentasi data menjadi salah satu fokus utama karena berperang langsung dalam meningkatkan *robustness* model CNN. Berbeda dengan augmentasi konvensional, pada penelitian ini digunakan kombinasi transformasi yang lebih beragam dan adaptif untuk mensimulasikan gangguan nyata.



Gambar 4. Grafik Model Accuracy dan Loss Robustness.

Berdasarkan hasil pelatihan yang divisualisasikan pada Gambar 4, nilai *validation accuracy* mengalami peningkatan stabil hingga mencapai puncak pada *epoch* ke-5 sebesar 0.83. Setelahnya, nilai akurasi validasi mulai menurun secara perlahan meskipun akurasi pelatihan terus meningkat hingga mendekati 0.95 pada *epoch* ke-20. Fenomena ini menunjukkan bahwa model telah mencapai titik konvergensi optimal pada *epoch* ke-5 dan mulai mengalami *overfitting* pada *epoch* selanjutnya. Oleh karena itu, *epoch* ke-5 digunakan sebagai acuan untuk penyimpanan bobot model terbaik (*best model checkpoint*).

Tabel 2. Hasil pengujian robustness model pre-trained pada noise Gaussian Noise

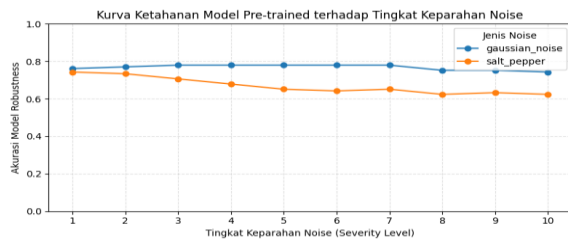
Severity	Akurasi Model	Penurunan Akurasi	Akurasi Robustness
1	0.8357	0.7890	-0.0467
2	0.8357	0.7798	-0.0559
3	0.8357	0.7615	-0.0742
4	0.8357	0.7706	-0.0651
5	0.8357	0.7615	-0.0742
6	0.8357	0.7615	-0.0742
7	0.8357	0.7615	-0.0742
8	0.8357	0.7523	-0.0834
9	0.8357	0.7523	-0.0834
10	0.8357	0.7523	-0.0834

Berdasarkan hasil uji *robustness* terhadap *Gaussian Noise* pada model terbaik yakni *epoch* ke-5, akurasi 0.83568, terlihat bahwa gangguan visual jenis ini menyebabkan penurunan performa deteksi yang konsisten seiring meningkatnya tingkat keparahan (*severity*). Pola ini menunjukkan bahwa semakin tinggi tingkat *Gaussian Noise*, semakin besar gangguan acak pada nilai piksel yang muncul di seluruh gambar.

Tabel 3. Hasil pengujian robustness model pre-trained pada noise Salt and Pepper

Severity	Akurasi Model	Akurasi Robustness	Penurunan akurasi
1	0.8357	0.7339	-0.1017
2	0.8357	0.7523	-0.0834
3	0.8357	0.6972	-0.1384
4	0.8357	0.6789	-0.1568
5	0.8357	0.6514	-0.1843
6	0.8357	0.6422	-0.1935
7	0.8357	0.6330	-0.2027
8	0.8357	0.6239	-0.2118
9	0.8357	0.6330	-0.2027
10	0.8357	0.6330	-0.2027

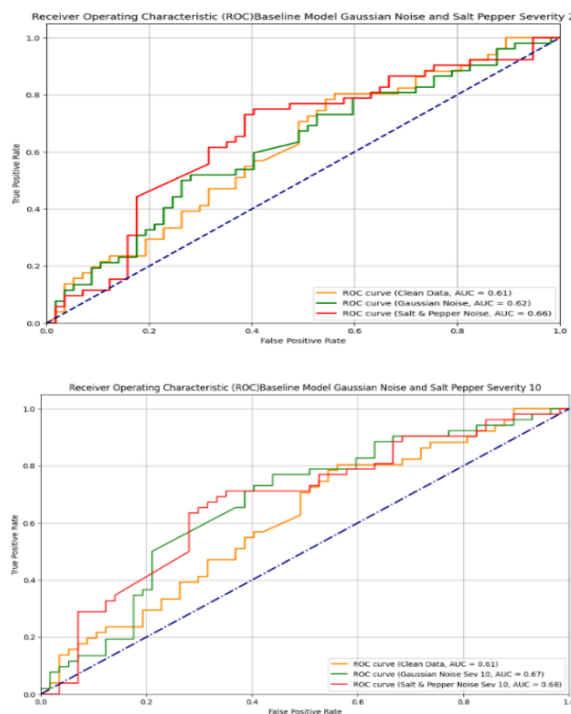
Pada tingkat keparahan ringan (*severity* 1–2), akurasi model masih relatif baik di kisaran 0.73–0.75, dengan penurunan sekitar 8–10 poin persentase dari kondisi bersih. Namun, ketika tingkat keparahan meningkat di atas *severity* 3, performa model menurun tajam hingga 0.63–0.68, dengan total penurunan mencapai 18–21 poin persentase pada level *severity* tertinggi.



Gambar 4. Kurva Ketahanan Model Pre-Trained terhadap Noise

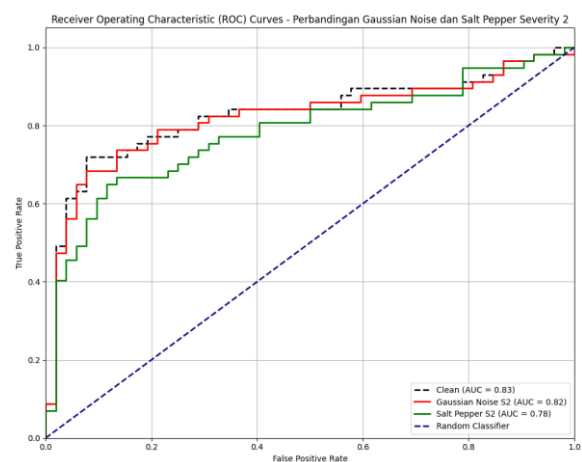
Grafik tersebut membandingkan ketahanan model pre-trained terhadap dua jenis noise, yaitu *Gaussian Noise* dan *Salt & Pepper* pada *severity* tingkat 10. Seiring meningkatnya tingkat gangguan, akurasi model mengalami penurunan bertahap. Pada *Gaussian Noise*, penurunan akurasi relatif cenderung menurun dengan nilai yang masih bertahan di kisaran 0.75-0.79, sedangkan pada *salt & pepper noise* penurunan terjadi lebih tajam hingga mencapai sekitar 0.63 pada tingkat *severity* tertinggi. Dengan demikian, meskipun model masih dapat mempertahankan performa cukup baik pada *Gaussian Noise*, hasil pengujian ini menunjukkan bahwa robustness model terhadap *salt & pepper* masih belum optimal dan masih perlu peningkatan.

Untuk menilai apakah akurasi yang dihasilkan model benar-benar mencerminkan kemampuan dalam membedakan gambar *AI-generated* dan gambar asli, dilakukan pengujian menggunakan *Receiver Operating Characteristic (ROC) Curve*. ROC menampilkan hubungan antara *True Positive Rate* (Rate) dan *False Positive Rate* (FPR) pada berbagai ambang Keputusan



Gambar 5. ROC Baseline Model pada Gaussian Noise dan Salt Pepper Severity 2 dan 10

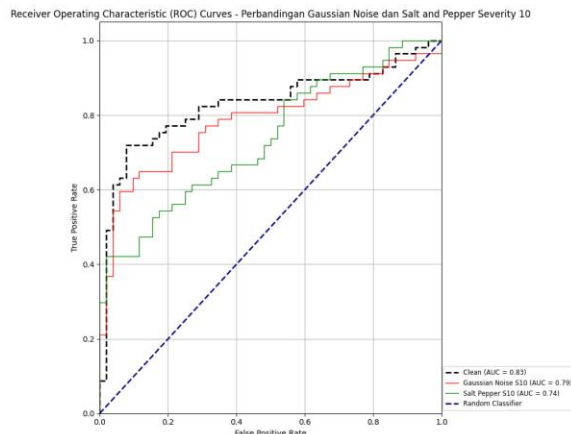
Ketika diuji dengan *noise* tingkat rendah atau *severity* 2. Pada data bersih, model memiliki nilai AUC 0.61 yang sudah menunjukkan bahwa model kurang mampu membedakan gambar *AI-generated* dan gambar asli. Setelah diberi *Gaussian Noise severity* 2, nilai AUC hanya naik sedikit menjadi 0.62 dan *Salt and Pepper* menghasilkan AUC 0.66. Hal itu juga berlaku pada *Severity* 10 pada grafik 4.6 yang memperlihatkan AUC naik menjadi 0.67 pada *gaussian Noise* dan *Salt and Pepper* menghasilkan AUC 0.68. Dengan kata lain model dapat membedakan kedua kelas karena gangguan *noise* secara tidak sengaja memperbesar perbedaan visual antar kelas dan model seharusnya mampu mengenali karakteristik asli gambar bukan mengandalkan pola yang muncul akibat kerusakan gambar.



Gambar 6. ROC Gaussian Noise and Salt & Pepper Severity 2

Kurva ROC pada tingkat gangguan *severity* 2 menunjukkan bahwa model memperoleh AUC 0.83 pada data bersih dan hanya turun sedikit menjadi 0.82 pada *Gaussian Noise*. Perbedaan yang kecil ini menandakan bahwa gangguan halus yang terdistribusi merata tidak mengganggu kemampuan model dalam membedakan gambar asli dan gambar buatan *AI*. Pada *Salt and Pepper Noise*, kurva ROC bergeser ke kanan bawah dengan AUC 0.78. Pergeseran ini menunjukkan bahwa gangguan acak mulai merusak pola tekstur lokal yang penting bagi *CNN* sehingga meningkatkan kesalahan prediksi. Secara keseluruhan, model masih stabil terhadap *Gaussian Noise* pada tingkat gangguan ringan, tetapi mulai kehilangan performa ketika diberikan *Salt and Pepper Noise*.

Temuan ini menunjukkan bahwa transfer learning efektif menghadapi gangguan halus, namun ketahanan terhadap noise acak masih perlu ditingkatkan.



Gambar 7. ROC Gaussian Noise and Salt & Pepper Severity 10

Perbandingan kurva ROC pada *severity* 10 antara kondisi data bersih memperoleh AUC 0.83, menunjukkan kemampuan diskriminatif yang cukup baik. Ketika diberikan *Gaussian Noise* tingkat tinggi, AUC turun menjadi 0.79 dan sebaliknya, *Salt and Pepper* menurunkan AUC menjadi 0.74 karena *noise* acak ini merusak piksel secara langsung. Hasil ini menunjukkan bahwa model lebih tahan terhadap gangguan distribusi halus seperti *Gaussian*, namun masih rentan terhadap gangguan yang merusak struktur lokal seperti *Salt and Pepper*.

Hasil pengujian menunjukkan bahwa model *baseline* memiliki akurasi validasi 0.6618 dengan selisih besar terhadap akurasi pelatihan, menandakan *overfitting* dan kemampuan generalisasi yang rendah. Ketika diuji dengan *noise* pada tingkat *severity* 1 sampai 10, performanya menurun tajam terutama pada *Gaussian Noise* dan *Salt and Pepper Noise*. Pada *severity* tertinggi, akurasi turun hingga 0.50 atau sekitar 16 poin persentase dari kondisi awal. Penurunan ini terjadi karena *noise* acak merusak struktur lokal seperti tekstur dan tepi yang menjadi dasar ekstraksi fitur CNN, sedangkan gangguan seperti *Gaussian Blur* dan *JPEG Compression* hanya menyebabkan penurunan moderat karena bentuk global citra masih dapat dikenali. Setelah menerapkan *transfer learning* pada model *pre trained NYUAD ComNets*, performa meningkat signifikan dengan akurasi 0.83 pada data bersih. Ketika diuji dengan *Gaussian Noise*, akurasi hanya turun ke kisaran 0.75, namun *Salt and Pepper Noise* menurunkannya lebih drastis hingga 0.63.

Hal ini menunjukkan bahwa *transfer learning* meningkatkan ketahanan model terhadap gangguan halus tetapi tidak sepenuhnya mengatasi kerentanan terhadap *noise* acak. Pada *Gaussian Noise*, AUC turun dari 0.83 menjadi 0.78 pada *severity* 10 sehingga model masih mampu menjaga kemampuan diskriminatifnya. Namun pada *Salt and Pepper Noise*, AUC turun ke 0.74 dan kurva ROC bergeser ke kanan bawah yang menandakan peningkatan kesalahan klasifikasi. Secara keseluruhan, model lebih *robust* terhadap gangguan yang terdistribusi merata,

sedangkan gangguan acak yang merusak pola tekstur tetap menjadi tantangan utama

5. KESIMPULAN DAN SARAN

Hasil penelitian menunjukkan bahwa model CNN *baseline* yang dilatih dari awal memiliki kemampuan generalisasi dan ketahanan yang terbatas terhadap gangguan visual. Performa model cenderung menurun signifikan ketika diuji pada data berganggu, yang mengindikasikan sensitivitas tinggi terhadap perubahan nilai piksel dan ketidakmampuan mempertahankan karakteristik pembeda pada kondisi tidak ideal. Penerapan *transfer learning* menggunakan bobot *pre-trained NYUAD-ComNets/AI-Generated Images Detector* terbukti mampu meningkatkan performa dan robustness model secara signifikan. Model menunjukkan peningkatan akurasi yang jelas pada data bersih serta ketahanan yang lebih baik ketika dihadapkan pada gangguan visual, khususnya *Gaussian Noise*. Namun demikian, penurunan performa yang lebih tajam pada *Salt and Pepper Noise* mengindikasikan bahwa model masih memiliki keterbatasan dalam menghadapi *noise* acak yang merusak struktur spasial lokal. Temuan ini menegaskan bahwa karakteristik dan tingkat keparahan gangguan visual sangat memengaruhi ketahanan model deteksi gambar *AI-generated*, serta bahwa CNN cenderung lebih *robust* terhadap gangguan yang terdistribusi secara halus dibandingkan gangguan acak dengan intensitas tinggi.

Sebagai pengembangan lanjutan, penelitian selanjutnya dapat memperluas evaluasi generalisasi model dengan menggunakan dataset dari berbagai sumber dan melakukan pengujian lintas dataset untuk mengukur ketahanan terhadap domain dan generator AI yang berbeda. Selain itu, eksplorasi pendekatan tambahan seperti *adversarial training*, perbandingan beberapa arsitektur CNN *pre-trained*, serta penerapan *fine-tuning* parsial pada *backbone* model berpotensi meningkatkan robustness dan performa klasifikasi. Evaluasi juga dapat dikembangkan menggunakan skenario dunia nyata, misalnya dengan menghasilkan gambar AI secara mandiri yang kemudian mengalami degradasi alami seperti kompresi, *resize*, *screenshot*, dan proses unggah-unduh ulang, sehingga hasil penelitian menjadi lebih representatif terhadap kondisi penggunaan sebenarnya.

DAFTAR PUSTAKA

- [1] L. Zhang *et al.*, "X-Transfer: A Transfer Learning-Based Framework for GAN-Generated Fake Image Detection," Jan. 2024, [Online]. Available: <http://arxiv.org/abs/2310.04639>
- [2] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting CNN-Generated Facial Images in Real-World Scenarios," May 2020. doi: <https://doi.org/10.48550/arXiv.2005.05632>.
- [3] S. A. Hakim *et al.*, "KLASIFIKASI CITRA GENERASI ARTIFICIAL INTELLIGENCE MENGGUNAKAN METODE FINE TUNING

- PADA RESIDUAL NETWORK AI GENERATED IMAGE CLASSIFICATION USING FINE TUNING ON RESIDUAL NETWORK,” vol. 11, no. 3, pp. 655–666, 2024, doi: 10.25126/jtiik.938118.
- [4] N. Zhong, Y. Xu, S. Li, Z. Qian, and X. Zhang, “Rich and Poor Texture Contrast: A Simple yet Effective Approach for AI-generated Image Detection,” Mar. 2024. [Online]. Available: <http://arxiv.org/abs/2311.12397>
- [5] C. Liu, T. Zhu, S. Shen, and W. Zhou, “Towards Robust Gan-Generated Image Detection: A Multi-View Completion Representation,” 2023.
- [6] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, “Feature-Based Transfer Learning,” in *Transfer Learning*, Cambridge University Press, 2020, pp. 34–44. doi: 10.1017/9781139061773.005.
- [7] A. Raup, W. Ridwan, Y. Khoeriyah, Q. Yuliati Zaqiah, and U. Islam Negeri Sunan Gunung Djati Bandung, “Deep Learning dan Penerapannya dalam Pembelajaran,” *JiIP - Jurnal Ilmiah Ilmu Pendidikan*, Sep. 2022, doi: <https://doi.org/10.54371/jiip.v5i9.805>.
- [8] D. T. Ananto *et al.*, “Edukasi dan Pelatihan Pengenalan Machine Learning dan Computer Vision Untuk Mengeksplorasi Potensi Visual,” Oct. 2023, [Online]. Available: <http://jurnal.umj.ac.id/index.php/semnaskat>
- [9] K. Azmi, S. Defit, and U. Putra Indonesia YPTK Padang Jl Raya Lubuk Begalung-Padang-Sumatera Barat, “Implementasi Convolutional Neural Network (CNN) Untuk Klasifikasi Batik Tanah Liat Sumatera Barat,” *Unitek*, vol. 16, no. 1, Jun. 2023, doi: <https://doi.org/10.52072/unitek.v16i1.504>.
- [10] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, “Fundamental concepts of convolutional neural network,” *Intelligent Systems Reference Library*, vol. 172, pp. 519–567, Jan. 2020, doi: 10.1007/978-3-030-32644-9_36.
- [11] P. Meliawati and E. Kurniati, “Ekstraksi Data Digital Menggunakan Teknik Max Pooling dan Average Pooling,” *Jurnal Riset Matematika*, pp. 137–144, Dec. 2022, doi: 10.29313/jrm.v2i2.1338.
- [12] B. Jabir and N. Falih, “Dropout, a basic and effective regularization method for a deep learning model: A case study,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, pp. 1009–1016, Nov. 2021, doi: 10.11591/ijeecs.v24.i2.pp1009-1016.
- [13] Fauzan Muhammad, Aniati Murni Arimurthy, and Dina Chahyati, “Transfer learning pada Network VGG16 dan ResNet50,” *The Indonesian Journal of Computer Science*, vol. 12, Feb. 2023, doi: 10.33022/ijcs.v12i1.3130.
- [14] M. Saini and S. Susan, “Tackling class imbalance in computer vision: a contemporary review,” *Artif. Intell. Rev.*, vol. 56, pp. 1279–1335, Oct. 2023, doi: 10.1007/s10462-023-10557-6.
- [15] D. Husen, “EVALUASI TEKNIK AUGMENTASI DATA UNTUK KLASIFIKASI TUMOR OTAK MENGGUNAKAN CNN PADA CITRA MRI (PERFORMANCE EVALUATION OF CNN MODELS WITH VARIOUS DATA AUGMENTATION TECHNIQUES ON MRI IMAGES FOR BRAIN TUMOR CLASSIFICATION),” *Teknimedia*, Dec. 2024, doi: <https://doi.org/10.46764/teknimedia.v5i2.220>.
- [16] H. Ben Braiek and F. Khomh, “Machine Learning Robustness: A Primer,” Mar. 2024. [Online]. Available: <http://arxiv.org/abs/2404.00897>
- [17] A. Balendran *et al.*, “A scoping review of robustness concepts for machine learning in healthcare,” *NPJ Digit. Med.*, vol. 8, no. 1, Dec. 2025, doi: 10.1038/s41746-024-01420-1.
- [18] L. Fabris, “Is it possible to identify which AI Generator generated an image?,” Vienna University of Economics and Business, 2024. Accessed: Dec. 21, 2025. [Online]. Available: <https://penni.wu.ac.at/supervision/Laura%20Fabris%20Bachelor%20Thesis%202024.pdf>
- [19] A. Rianti *et al.*, “CRISP-DM: Metodologi Proyek Data Science,” Jul. 2023. Accessed: Dec. 21, 2025. [Online]. Available: <https://ojs.udb.ac.id/index.php/Senatib/article/view/3015>