BAB I
PENDAHULUAN

1.1. Latar Belakang

Kemajuan teknologi berperan sebagai faktor utama dalam perubahan berbagai sektor kehidupan yang berpengaruh terhadap individu, perusahaan, maupun pemerintahan. Salah satu inovasi teknologi yang memperoleh perhatian global adalah *Artificial Intelligence* (AI). AI didefinisikan sebagai sistem yang dirancang untuk meniru kemampuan penalaran manusia, memungkinkan mesin melakukan tugas-tugas yang biasanya membutuhkan kecerdasan manusia serta menghasilkan keluaran yang bermanfaat dan sebanding dengan kemampuan manusia [1]. Perkembangan AI semakin pesat dalam berbagai bidang, termasuk industri musik yang kini tengah mengalami transformasi signifikan. Teknologi AI telah mampu menghasilkan komposisi musik secara otomatis dengan kualitas yang menyerupai karya manusia melalui algoritma *generative models*. Salah satu fenomena ini ditunjukkan oleh laporan platform *streaming* musik Deezer, yang pada September 2025 menerima lebih dari 30.000 lagu yang sepenuhnya dihasilkan oleh AI setiap hari meningkat tajam dari sekitar 20.000 lagu pada April dan 10.000 lagu pada Januari tahun yang sama [2]. Lonjakan tersebut mengindikasikan adanya pertumbuhan pesat dalam produksi konten musik berbasis AI.

Di tengah percepatan inovasi teknologi, berbagai platform *AI song generator* memungkinkan pengguna menciptakan lagu hanya melalui *prompt* berbasis teks, dengan sistem yang secara otomatis menghasilkan komposisi lengkap mencakup aransemen, instrumen, genre, lirik, hingga vokal. Meskipun inovasi ini membuka peluang baru dalam kreativitas musik, kehadiran musik AI menimbulkan kekhawatiran terkait isu hak cipta dan orisinalitas karya [3]. Berdasarkan Pasal 40 ayat (1) sub (d) Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta, karya lagu dan musik memperoleh perlindungan hukum di Indonesia [4]. Namun regulasi tersebut belum secara khusus mengatur penggunaan AI dalam penciptaan karya cipta. Kekosongan pengaturan ini berpotensi menimbulkan pelanggaran hak cipta serta tantangan dalam mengidentifikasi keaslian lagu, sehingga diperlukan sistem

yang mampu membedakan karya manusia dan karya hasil generasi AI untuk menjaga integritas serta nilai ekonomi industri musik.

Penelitian mengenai deteksi musik yang dihasilkan oleh kecerdasan buatan (AI-*generated music*) telah dilakukan menggunakan dan diproses dengan berbagai *encoder* lalu dilakukan dengan model konvolusi sederhana. Hasil penelitian menunjukkan bahwa deteksi musik AI bisa mencapai akurasi sangat tinggi, yakni hingga 99,8%. Namun, angka akurasi tersebut menurun ketika sistem berhadapan dengan audio yang telah dimanipulasi [5]. Temuan ini mengindikasikan bahwa meskipun model CNN mampu mencapai akurasi tinggi dalam kondisi laboratorium, performanya cenderung menurun secara signifikan ketika menghadapi variasi data dan kondisi yang lebih menyerupai dunia nyata, sehingga menunjukkan keterbatasan dalam aspek *robustness*. Selain itu, sebagian besar penelitian sebelumnya belum secara eksplisit mengevaluasi ketahanan model pada distribusi data yang berbeda, seperti pengujian lintas genre sebagai bentuk *out-of-distribution testing*. Dengan demikian, fokus penelitian tidak lagi hanya pada akurasi, tetapi pada kemampuan model mempertahankan performa di luar distribusi pelatihan.

Pada penelitian tersebut membuktikan model *Convolutional Neural Network* (CNN) menghasilkan performa yang baik dalam penerapannya terhadap membedakan lagu AI. Model ini sendiri memiliki lapisan konvolusi yang cocok untuk data dengan *topologi grid* seperti gambar dan video. CNN terbukti sangat efektif dalam pengenalan pola dan klasifikasi dalam berbagai domain, termasuk pengolahan citra, pengenalan suara, dan analisis sinyal [6]. CNN memiliki struktur yang dirancang khusus untuk mempelajari pola, sehingga sangat ideal untuk memproses data visual termasuk data audio yang dapat divisualisasikan menjadi spektogram [7]. Spektogram merupakan representasi spasial dari audio yang menyimpan informasi waktu dan frekuensi secara bersamaan, sehingga memungkinkan CNN untuk mendeteksi pola-pola spasial dengan optimal [8].

Performa CNN sangat bergantung pada kualitas, kuantitas, serta keberagaman data latih. Jumlah data yang terbatas maupun distribusi data yang kurang variatif dapat memicu *overfitting*, sehingga model hanya optimal pada kondisi yang serupa dengan data pelatihan [9]. Oleh karena itu, diperlukan

pendekatan untuk meningkatkan *robustness*, yaitu kemampuan model mempertahankan performa ketika menghadapi variasi kondisi *input*. Salah satu upaya peningkatan kekokohan model (*robustness*) yakni kemampuan model mempertahankan akurasi meski terjadi perubahan pada *input* seperti kecerahan, *blur*, atau *noise* dapat dilakukan tanpa biaya besar [10]. Dalam penelitian ini, *robustness* ditingkatkan melalui *data augmentation*, yaitu teknik menambah variasi data pelatihan tanpa mengumpulkan data baru [11]. Pendekatan ini memungkinkan model belajar dari beragam kondisi representasi audio sehingga tidak hanya sensitif terhadap pola spesifik pada data latih, tetapi juga mampu beradaptasi terhadap variasi karakteristik audio yang lebih luas. Pada kasus audio, augmentasi dilakukan dengan memodifikasi data suara atau spektogram mentah (*raw*) yang merepresentasikan amplitudo spektrum terhadap waktu tanpa transformasi lanjutan seperti *cepstral* [12].

Dalam sistem deteksi, interpretasi dibutuhkan untuk memahami dan menjelaskan hasil yang dihasilkan oleh model. Model ini sering dianggap sebagai "*black box*" karena kurangnya transparansi dalam pengambilan keputusan, sehingga menghambat adopsi AI, terutama ketika pengguna membutuhkan penjelasan atas hasil model [13]. Pada data audio interpretabilitas ini penting untuk meningkatkan transparansi sekaligus mengarahkan pemilihan representasi yang lebih sesuai bagi tugas klasifikasi audio khususnya pada representasi visual [14]. Oleh karena itu, pendekatan *Explainable AI* (XAI) diperlukan untuk menjembatani kompleksitas model dengan kebutuhan pemahaman manusia [15]. Dalam konteks deteksi lagu berbasis AI dengan data *input* berupa representasi visual dari audio, teknik XAI yang paling sesuai adalah metode yang menghasilkan *output* berbentuk visual, karena mampu menyoroti area pada representasi visual audio yang paling berpengaruh terhadap hasil klasifikasi model.

Arsitektur *Residual Network* (ResNet) dikembangkan untuk mengatasi permasalahan degradasi performa dan *vanishing gradient* yang sering terjadi pada jaringan CNN berlapis dalam. Dengan adanya *skip connection* yang menghubungkan antar-lapisan di dalam jaringan berfungsi sebagai pemetaan identitas (*identity mapping*), ResNet mampu mempertahankan aliran informasi

melalui gradien dan mempelajari fitur tingkat menengah (*mid-level features*) yang penting untuk mengenali pola kompleks pada data audio sehingga membantu meningkatkan stabilitas serta kinerja proses pelatihan model [16], [17]. Karakteristik ini menjadikannya sesuai untuk tugas klasifikasi berbasis spektogram, khususnya dalam konteks peningkatan *robustness* melalui *data augmentation*. Meskipun arsitektur yang lebih dalam memiliki kapasitas representasi yang lebih besar, peningkatan kedalaman tidak selalu berbanding lurus dengan kemampuan generalisasi, terutama pada dataset yang terbatas, karena berpotensi meningkatkan risiko *overfitting* dan kompleksitas komputasi. Oleh karena itu, ResNet dipilih dalam penelitian ini karena mampu menyeimbangkan kedalaman jaringan, efisiensi komputasi, stabilitas gradien, serta performa generalisasi, sekaligus mendukung penerapan metode interpretabilitas seperti Grad-CAM untuk menganalisis area spektogram yang paling berkontribusi terhadap hasil klasifikasi.

Salah satu penelitian menerapkan metode *Explainable AI* yaitu teknik Grad-CAM untuk memvisualisasikan area penting pada spektogram dengan membandingkan performa VGG19, ResNet-18, dan CNN kustom dalam pengklasifikasian genre musik menggunakan *Mel-spectrogram*, dengan. ResNet-18 menunjukkan akurasi sebesar 75,5%. Meskipun akurasi ini tergolong rendah, model ini menonjol dalam hal interpretabilitas, dengan Grad-CAM menunjukkan fokus pada pola tekstur yang konsisten, sementara area frekuensi tinggi yang kosong cenderung lebih sering teraktivasi pada model lainnya [8]. Penelitian lain yang mengkaji kombinasi berbagai teknik ekstraksi fitur dan arsitektur jaringan saraf untuk klasifikasi suara pernapasan anak-anak, dengan metode ekstraksi STFT, *Mel Spectrogram*, dan Wav2vec, serta pengklasifikasi LightCNN, ResNet-18, dan *Audio Spectrogram Transformer* (AST). Hasil menunjukkan bahwa kombinasi R-STFT (ResNet-18 dengan STFT) memberikan performa terbaik, menghasilkan nilai *harmonic score* tertinggi pada empat tugas klasifikasi yang diuji, dengan keunggulan pada penggabungan representasi frekuensi STFT dan arsitektur residual ResNet-18. Lalu disusul dengan kombinasi R-Mel (ResNet-18 dengan Mel Spektogram) dan kombinasi lain menunjukkan hasil yang cukup kompetitif pada Task 2 [18].

Berdasarkan temuan tersebut, ResNet-18, sebagai salah satu arsitektur *Convolutional Neural Network* (CNN), menunjukkan performa yang baik ketika diterapkan pada data audio dengan representasi visual berbasis spektogram meskipun struktur komputasinya relatif ringan. Keunggulan ini menjadikan ResNet-18 sesuai untuk penelitian yang memanfaatkan spektogram sebagai *input*, karena mampu menyeimbangkan kompleksitas model dengan kebutuhan pemrosesan yang efisien. Selain itu, arsitektur ResNet-18 secara inheren lebih *XAI-friendly* pada teknik Grad-CAM, karena Grad-CAM sangat bergantung pada stabilitas gradien di lapisan konvolusi terakhir. Namun demikian, varian ResNet lain seperti ResNet-34, ResNet-50, atau arsitektur yang lebih dalam memiliki kapasitas representasi yang lebih tinggi akibat jumlah layer dan parameter yang lebih besar. Peningkatan kedalaman ini berpotensi menangkap fitur yang lebih kompleks, tetapi juga meningkatkan risiko *overfitting* serta kebutuhan komputasi yang lebih besar. Oleh karena itu, pemilihan varian ResNet perlu disesuaikan dengan karakteristik dataset, kompleksitas tugas klasifikasi, serta kebutuhan interpretabilitas model.

Permasalahan utama yang diangkat dalam penelitian ini adalah keterbatasan model *Convolutional Neural Network* (CNN) dalam mendeteksi lagu *AI-generated*. Meskipun CNN telah banyak digunakan untuk tugas klasifikasi musik seperti pengenalan genre, emosi, dan instrumen, sebagian besar penelitian belum secara spesifik menyoroti deteksi keaslian lagu yang dihasilkan AI. Tantangan ini menjadi semakin kompleks karena perbedaan antara musik manusia dan musik AI sering kali sangat halus dan sulit dibedakan secara akustik, sementara performa model sangat bergantung pada kualitas serta keragaman data latih. Model cenderung rentan ketika menghadapi variasi audio di luar distribusi pelatihan, baik berupa manipulasi sinyal seperti *noise* dan distorsi maupun perbedaan karakteristik musikal seperti tempo, tonalitas, dan genre. Selain itu, kajian yang membandingkan pengaruh kedalaman arsitektur CNN terhadap kemampuan *robustness* masih terbatas, padahal kompleksitas jaringan yang lebih dalam tidak selalu menjamin peningkatan generalisasi terutama pada konteks deteksi lagu AI yang melibatkan variasi manipulasi sinyal dan distribusi genre berbeda secara simultan. Kondisi ini

menunjukkan perlunya pengembangan dan evaluasi model yang tidak hanya akurat, tetapi juga *robust* terhadap variasi data serta memiliki tingkat interpretabilitas yang memadai untuk mendukung transparansi dalam proses deteksi keaslian lagu.

Penelitian ini bertujuan untuk meningkatkan *robustness* model klasifikasi berbasis *Convolutional Neural Network* (CNN) dalam mendeteksi dan membedakan antara lagu hasil buatan AI dan lagu ciptaan manusia melalui representasi spektogram audio dengan pendekatan data *augmentation*. Fokus utama penelitian ini adalah menganalisis bagaimana arsitektur CNN merespons perubahan dan manipulasi sinyal audio yang berpotensi terjadi dalam kondisi dunia nyata. Selain itu, penelitian ini juga mengintegrasikan metode *Explainable Artificial Intelligence* (XAI), khususnya Grad-CAM, untuk meningkatkan interpretabilitas model dengan menampilkan area spektrum yang paling berkontribusi terhadap keputusan klasifikasi. Dengan demikian, penelitian ini diharapkan tidak hanya menghasilkan sistem deteksi lagu AI yang akurat dan *robust*, tetapi juga transparan serta memiliki nilai interpretabilitas yang tinggi.

1.2. Rumusan Masalah

Berdasarkan penjelasan latar belakang yang telah disampaikan, rumusan masalah dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana pengaruh kompleksitas arsitektur CNN berbasis ResNet dan penerapan data *augmentation* terhadap performa dan *robustness* model dalam mendeteksi lagu AI?
2. Bagaimana metode *Explainable Artificial Intelligence* (XAI) dapat digunakan untuk memberikan interpretasi terhadap proses dan hasil klasifikasi model?
3. Bagaimana merancang dan mengimplementasikan antarmuka pengguna (GUI) untuk menampilkan hasil deteksi dan interpretasi model secara interaktif?

1.3. Batasan Masalah

Batasan masalah menjadi aspek penting dalam penelitian untuk menghindari ruang lingkup yang terlalu luas. Berikut beberapa pembatasan masalah yang diterapkan dalam penelitian ini:

1. Data utama yang digunakan dalam penelitian ini terdiri dari lagu *human-generated* dan lagu *AI-generated* dengan jumlah masing-masing 50 lagu setiap label sebagai data pelatihan dan pengujian utama. Selain itu, terdapat 12 lagu tambahan pada setiap label sebagai data uji *robust.*
2. Penelitian ini difokuskan pada peningkatan *robustness model* dalam mendeteksi lagu AI. terhadap variasi kondisi audio dan perbedaan karakteristik data.
3. Arsitektur CNN yang digunakan berbasis *Residual Network* (ResNet-18, ResNet-34, dan ResNet-50).
4. Ekstraksi fitur audio dilakukan menggunakan representasi spektogram berbasis STFT (*Short-Time Fourier Transform*).
5. Teknik *Explainable AI* (XAI) menggunakan Grad-CAM hanya menginterpretasikan *heatmap* spektogram dari hasil segmentasi lagu setelah melalui proses *trimming.*
6. Teknik *data augmentation* dibatasi dalam pemrosesan audio, seperti *pitch shifting*, *time stretching*, dan *gaussian noise.*
7. Evaluasi performa model dibatasi pada penggunaan metrik *accuracy*, *precision*, *recall*, dan *F1-score.*
8. Implementasi hasil penelitian dilakukan dalam bentuk prototipe berbasis *web application* menggunakan Streamlit.

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah yang telas dijelaskan, tujuan penelitian ini disusun untuk memberikan arahan yang lebih jelas serta sebagai upaya untuk menjawab permasalahan yang dirumuskan. Adapun tujuan dari penelitian sebagai berikut:

1. Menganalisis dan membandingkan performa beberapa arsitektur CNN berbasis *Residual Network* dalam mendeteksi lagu AI, baik tanpa maupun dengan penerapan *data augmentation*, serta mengevaluasi pengaruh kompleksitas arsitektur *dan augmentasi* terhadap peningkatan *robustness* dan kemampuan generalisasi model pada skenario manipulasi sinyal dan pengujian lintas genre.
2. Menerapkan metode *Explainable AI* (XAI) untuk memberikan interpretasi dari hasil kinerja model
3. Merancang dan mengimplementasikan GUI berbasis Streamlit untuk menampilkan hasil deteksi dan interpretasi model secara interaktif.

1.5. Manfaat Penelitian

Penelitian yang dilakukan mempunyai manfaat untuk masyarakat, baik masyarakat akademik maupun non akademik. Adapun beberapa manfaat dari penelitian ini.

1. Manfaat Teoritis
   - Menjadi referensi penelitian di bidang pengolahan sinyal audio, penerapan *deep learning*, dan *explainable AI* (XAI) untuk deteksi lagu berbasis AI.
   - Memberikan kontribusi terhadap pengembangan metode klasifikasi CNN berbasis *Residual Network* yang *robust* melalui data *augmentation*, sekaligus memperluas kajian tentang bagaimana XAI dapat meningkatkan interpretabilitas model dalam mendeteksi lagi generasi AI
2. Manfaat Praktis
   - Membantu industri musik Indonesia dalam mendeteksi keaslian lagu dengan menyediakan sistem yang tidak hanya akurat, tetapi juga dapat dijelaskan hasil keputusannya (*explainable*), sehingga meningkatkan kepercayaan pengguna terhadap sistem AI.
   - Memberikan solusi awal untuk mengidentifikasi perbedaan antara lagu asli dengan lagu buatan AI secara otomatis sekaligus menyediakan visualisasi atau penjelasan fitur audio yang menjadi dasar deteksi.