KLASIFIKASI GENRE MUSIK BERBASIS AUDIO DAN LIRIK MENGGUNAKAN CNN DAN MULTILINGUAL BERT

Dhany Ramadhan[1], Ilham Julian Efendi[1], Andi Tenriawaru[*1]
[1]Program Studi Ilmu Komputer, Fakultas MIPA, Universitas Halu Oleo, Kendari, Indonesia
Penulis korespondensi: Andi Tenriawaru (andi.tenriawaru@uho.ac.id)

Peningkatan pesat volume konten musik pada era digital memunculkan urgensi pengembangan sistem pengelompokan genre secara otomatis guna menggantikan metode konvensional yang kurang efisien. Penelitian ini merancang arsitektur klasifikasi genre musik berbasis pendekatan multimodal yang mengombinasikan ekstraksi sinyal audio dan informasi semantik lirik lagu berbahasa Indonesia dan Inggris pada empat genre fokus (Country, Hiphop, Pop, dan Rock). Skema deep learning yang diterapkan mengintegrasikan Convolutional Neural Network (CNN) untuk mengekstraksi representasi Mel-Spectrogram dari sinyal audio dan Multilingual BERT guna menangkap keterhubungan konteks lirik. Data audio bersumber dari GTZAN Genre Collection yang mencakup 10 kelas genre (Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, dan Rock), sementara data lirik dihimpun dari Genius Song Lyrics serta repositori lirik lagu Indonesia. Fitur dari kedua modalitas dilebur melalui teknik feature-level fusion. Hasil evaluasi menunjukkan bahwa model berbasis audio memperoleh akurasi sebesar 90,4%, model berbasis lirik mencapai 61,7%, dan model multimodal memperoleh akurasi sebesar 93,8% pada pengujian internal. Akan tetapi, pengujian terhadap 100 sampel lagu dari Spotify menghasilkan akurasi masing-masing sebesar 68,0% untuk model audio, 72,5% untuk model lirik, dan 65,0% untuk model multimodal. Perbedaan tersebut menunjukkan bahwa kinerja model belum konsisten pada data dari sumber yang berbeda. Sistem kemudian diintegrasikan ke dalam aplikasi web berbasis Flask untuk mendukung klasifikasi genre melalui masukan audio, lirik, maupun kombinasi keduanya.

Keywords - B+ Tree, Full-Text Search, Kepegawaian, Sistem Informasi, Tenaga Kependidikan.

I. PENDAHULUAN

Peningkatan volume konten multimedia pada platform digital menimbulkan tantangan dalam pengorganisasian data secara terstruktur. Pengelompokan genre musik secara manual membutuhkan waktu besar serta berpotensi menghasilkan penilaian yang subjektif dan tidak konsisten. Oleh karena itu, diperlukan pendekatan komputasional berbasis kecerdasan buatan untuk mengotomatisasi klasifikasi genre musik [1].

Pengelompokan genre yang hanya mengandalkan sinyal audio sering kali mengalami kendala, terutama ketika lagu dari genre yang berbeda memiliki pola irama atau bentuk *Mel-Spectrogram* yang mirip [2]. Di sisi lain, lirik lagu memberikan informasi pesan dan tema yang dapat memperjelas identitas genre, khususnya pada lagu berbahasa Indonesia dan Inggris untuk genre Country, Hiphop, Pop, dan Rock [3].

Penelitian ini mengembangkan sistem klasifikasi genre musik dengan memanfaatkan karakteristik audio dan informasi tekstual dari lirik lagu. Ekstraksi fitur audio dilakukan menggunakan *Convolutional Neural Network* (CNN) terhadap representasi *Mel-Spectrogram*, sedangkan pemrosesan lirik memanfaatkan Multilingual BERT untuk memperoleh representasi semantik teks. Fitur dari kedua jalur pemrosesan kemudian digabungkan melalui pendekatan *feature-level fusion* [4], Kinerja pendekatan tersebut dievaluasi melalui pengujian model audio, model lirik, dan model multimodal, termasuk pengujian menggunakan data eksternal [5].

II. METODE PENELITIAN

A. Pengumpulan dan Pemrosesan Data

Tahapan awal dalam penelitian ini berfokus pada akuisisi data serta pemrosesan awal (*preprocessing*) terhadap dua modalitas data yang berbeda, yaitu audio dan lirik lagu. Langkah ini dilakukan untuk memastikan data siap digunakan dalam proses pelatihan model *deep learning*. Prosedur pengolahan kedua data tersebut dijelaskan sebagai berikut:

1. Data Audio: Dataset GTZAN terdiri atas 1.000 berkas audio yang mencakup sepuluh genre musik, yaitu Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, dan Rock [6]. Setiap audio dipotong menjadi durasi 30 detik, dinormalisasi, dan diubah menjadi bentuk citra *Mel-Spectrogram* dengan *sampling rate* 22.050 Hz, 128 *Mel-filters*, ukuran n_fft 2048, dan hop_length 512 [7] [8].

Eksperimen multimodal menggunakan empat genre musik, yaitu Country, Hiphop, Pop, dan Rock, dengan masing-masing 100 berkas audio asli. Dengan demikian, dataset awal mencakup 400 berkas audio. Setiap berkas selanjutnya dibagi menjadi sepuluh segmen berdurasi tiga detik sehingga diperoleh 4.000 segmen audio. Dataset dibagi menjadi data pelatihan sebanyak 2.880 segmen, validasi sebanyak 320 segmen, dan pengujian sebanyak 800 segmen. Pembagian dilakukan berdasarkan berkas

audio asli sebelum segmentasi, sehingga segmen yang berasal dari satu berkas yang sama tetap berada dalam subset yang sama. Oleh karena itu, 800 sampel pada pengujian internal merujuk pada segmen audio, bukan 800 lagu yang berbeda.

2. Data Lirik: Diambil dari Genius Song Lyrics [9] dan kumpulan lirik lagu Indonesia yang difokuskan pada 4 genre (Country, Hiphop, Pop, dan Rock). Tahap pembersihan teks meliputi pembersihan tanda baca (*case folding* dan penghapusan karakter khusus), penyetaraan label genre, serta penyeimbangan jumlah data. Total data yang dihasilkan sebanyak 11.616 lirik lagu berbahasa Indonesia dan Inggris. Teks lirik kemudian diolah menggunakan *tokenizer* Multilingual BERT dengan batas panjang teks (*max sequence length*) 256 [10].
3. Penyiapan data multimodal dilakukan dengan mengintegrasikan representasi fitur audio dan lirik menggunakan pendekatan *feature-level fusion*. Fitur audio diekstraksi oleh *Convolutional Neural Network* (CNN) dari citra *Mel-Spectrogram*, sedangkan fitur tekstual diperoleh melalui Multilingual BERT (mBERT) yang memproses lirik lagu. Pada eksperimen internal, lirik dipasangkan dengan data audio berdasarkan kesamaan label genre, bukan berdasarkan kecocokan identitas lagu. Oleh sebab itu, kedua modalitas yang digabungkan tidak selalu berasal dari lagu yang sama. Berbeda dengan prosedur tersebut, pengujian eksternal menggunakan data Spotify dengan audio dan lirik yang berasal dari lagu yang sama. Perbedaan mekanisme pemasangan ini merupakan aspek metodologis yang perlu diperhatikan ketika menafsirkan hasil pengujian internal dan eksternal, khususnya dalam menilai manfaat penggabungan kedua modalitas.

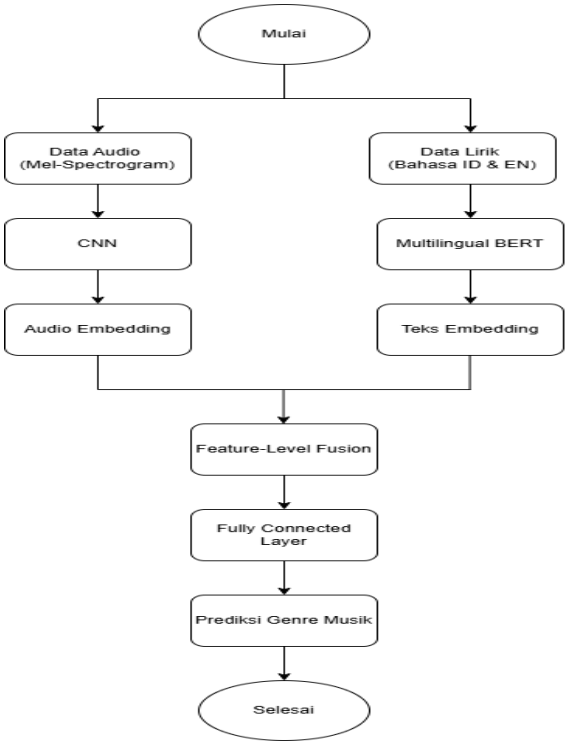
Kesesuaian data audio dan lirik perlu diperhatikan dalam proses penggabungan kedua modalitas. Pada pengujian menggunakan lagu dari Spotify, audio dan lirik dipilih dari lagu yang sama agar hasil klasifikasi merepresentasikan dua modalitas dari objek musik yang sama. Sementara itu, untuk data pelatihan, dasar pemasangan audio dan lirik harus dijelaskan sesuai dengan prosedur yang benar-benar diterapkan dalam eksperimen. Jika pemasangan dilakukan berdasarkan kesamaan genre dan bukan berdasarkan identitas lagu, pasangan tersebut tidak dapat dianggap sebagai pasangan audio dan lirik dari lagu yang sama.

B. Arsitektur Model Multimodal dan Web Service

Sistem multimodal menggunakan dua jalur pemrosesan, yaitu jalur audio dan jalur lirik. Data audio diperoleh dari GTZAN, sedangkan data lirik dikumpulkan dari *Genius Song Lyrics with Language Information* dan dataset lirik lagu Indonesia. Pada implementasi yang digunakan, data lirik dipilih secara acak dari kumpulan lirik dengan label genre yang sama dengan data audio. Dengan demikian, kesesuaian pasangan ditentukan berdasarkan kesamaan label genre, bukan berdasarkan verifikasi bahwa kedua data berasal dari lagu yang identik. Kondisi ini perlu dipertimbangkan dalam menafsirkan hasil evaluasi karena informasi audio dan lirik yang digabungkan belum tentu merepresentasikan lagu yang sama [4]. Selanjutnya, hasil penggabungan tersebut diintegrasikan ke dalam aplikasi web berbasis Flask untuk mendukung proses prediksi [5]. Secara keseluruhan, sistem yang dikembangkan terdiri atas tiga bagian utama:

1. Jalur Audio Berbasis CNN: Data audio direpresentasikan dalam bentuk *Mel-Spectrogram* [8] yang selanjutnya digunakan sebagai masukan pada model Convolutional Neural Network (CNN) [11], [12] . Arsitektur model terdiri atas beberapa *Convolutional 2D Layers*, *Batch Normalization*, fungsi aktivasi ReLU, serta *Max Pooling 2D* yang berfungsi mengurangi dimensi fitur. Selain itu, *Dropout Layer* diterapkan untuk meminimalkan risiko *overfitting*. Tahap akhir pada jalur ini menghasilkan representasi fitur dalam bentuk vektor *audio embedding*.
2. Jalur Lirik Berbasis Multilingual BERT: Data lirik lagu dalam bahasa Indonesia dan Inggris diproses menggunakan model Multilingual BERT (mBERT) [10], [13]. Representasi semantik dari teks diperoleh melalui keluaran token [CLS] (*Classification Token*), yang selanjutnya digunakan sebagai representasi vektor *text embedding* untuk menggambarkan karakteristik lirik [14].
3. Penggabungan Fitur dan Implementasi Aplikasi Web: Representasi fitur yang dihasilkan oleh jalur audio dan jalur lirik selanjutnya diintegrasikan menggunakan metode konkatenasi (*feature concatenation*) [4], [15] . Vektor hasil penggabungan kemudian diproses melalui *Fully Connected Layer* (*Dense Layer*) dan lapisan Softmax untuk menghasilkan probabilitas prediksi terhadap empat kelas genre, yaitu Country, Hiphop, Pop, dan Rock. Model multimodal yang telah diperoleh selanjutnya diintegrasikan ke dalam aplikasi web berbasis Flask [5], [16], sehingga pengguna dapat memberikan masukan berupa file audio dan lirik lagu untuk memperoleh hasil prediksi genre.

Secara keseluruhan, ketiga bagian tersebut membentuk arsitektur multimodal yang memungkinkan informasi dari audio dan lirik diproses secara terpisah sebelum diintegrasikan pada tahap *feature-level fusion* [4], [15]. Alur pemrosesan data serta hubungan antar-komponen dalam arsitektur model multimodal yang dikembangkan ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur Model Multimodal dengan Pendekatan Feature-Level Fusion menggunakan CNN dan Multilingual BERT

C. Konfigurasi Pelatihan Model

Pelatihan model dilakukan melalui beberapa skenario eksperimen untuk menentukan konfigurasi yang digunakan pada masing-masing pendekatan. Variasi pengujian mencakup pembagian dataset, ukuran *batch*, jumlah *epoch*, serta parameter lain yang relevan dengan arsitektur model. Evaluasi dilakukan terhadap model CNN untuk modalitas audio, mBERT untuk modalitas lirik, dan model multimodal yang mengintegrasikan kedua representasi fitur. Pada model multimodal, proses optimasi menggunakan optimizer Adam dan fungsi kerugian CrossEntropyLoss, sedangkan *early stopping* digunakan untuk membantu mengendalikan risiko *overfitting*. Rincian konfigurasi yang diterapkan pada setiap model disajikan pada Tabel 1.

Tabel 1. Konfigurasi Pelatihan Model Skenario Terbaik

Model	Parameter Utama	Pengaturan / Nilai Terbaik
Audio (CNN)	Pembagian Data, Ukuran *Batch*, *Epoch*	90:10, Batch 8, Epoch 30
Lirik (mBERT)	*Epoch*, *Batch*, *Max sequence length*	2×10^{-5}, Epoch 10, Batch 16, 256
Multimodal	*Split*, *Optimizer*, *Loss Function*, *Batch*, *learning rate*	80:20, Adam, *CrossEntropy*, Batch 8, 0,001

D. Persamaan Komputasi

Proses komputasi pada sistem meliputi operasi konvolusi 2D pada *Mel-Spectrogram*, konversi frekuensi ke skala Mel, serta perhitungan akurasi evaluasi model. Persamaan yang digunakan pada setiap proses dijelaskan sebagai berikut.

1. Operasi Konvolusi 2D

Proses konvolusi 2D digunakan pada model CNN untuk mengekstraksi pola dan karakteristik dari citra *Mel-Spectrogram* [11]. Operasi konvolusi dinyatakan menggunakan Persamaan (1).

$$(I * K)(i-m, j-n)K(m,n) \quad (1)$$

Keterangan rumus:

- I: Matriks *input* (seperti nilai intensitas pada spektrogram).
- K: *Kernel* atau *filter* yang digunakan untuk ekstraksi fitur.
- i, j: Koordinat lokasi pada hasil peta fitur (*output feature map*).
- m, n: Dimensi spasial dari *kernel*.

2. Konversi Frekuensi ke Skala Mel

Sebelum pembentukan Mel-Spectrogram, frekuensi audio dalam satuan Hertz dikonversi ke skala *Mel* [7]. Konversi tersebut dilakukan menggunakan Persamaan (2).

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

Keterangan rumus:

- $Mel(f)$ = nilai frekuensi pada skala Mel.
- f = frekuensi dalam satuan Hertz (Hz).
- 2595 = konstanta pengali dalam formula konversi skala Mel.
- 700 = konstanta frekuensi yang digunakan dalam formula konversi.
- $\log_{10}$ = fungsi logaritma berbasis 10.

3. Perhitungan Akurasi

Akurasi menunjukkan tingkat ketepatan model dalam mengklasifikasikan seluruh data yang diuji. Nilai akurasi dihitung berdasarkan perbandingan antara jumlah prediksi yang benar terhadap keseluruhan data pengujian menggunakan Persamaan (3).

$$Akurasi = \frac{\text{Jumlah Prediksi Benar}}{\text{Total Data Uji}} \quad (3)$$

III. HASIL DAN PEMBAHASAN

Kinerja model klasifikasi genre musik dievaluasi berdasarkan hasil pengujian pada masing-masing modalitas dan model multimodal. Evaluasi mencakup perbandingan performa pada data pengujian, analisis *Confusion Matrix* untuk melihat distribusi hasil klasifikasi, pengujian reliabilitas menggunakan data eksternal dari Spotify, serta evaluasi implementasi model melalui aplikasi web berbasis Flask.

A. Evaluasi Model pada Data Pengujian

Tahap evaluasi awal mencakup perbandingan kinerja antara model *baseline* berbasis audio, model *baseline* berbasis lirik, dan model multimodal yang menggabungkan kedua modalitas. Analisis dilakukan untuk mengetahui perbedaan kinerja klasifikasi yang dihasilkan oleh masing-masing pendekatan serta pengaruh integrasi fitur audio dan lirik terhadap hasil prediksi. Perbandingan hasil pengujian ketiga model tersebut ditampilkan pada Tabel 2.

Tabel 2. Perbandingan Hasil Pengujian Model

No	Model	Akurasi	F1-Score
1	Audio (CNN)	0.904	0.903
2	Lirik (Multilingual BERT)	0.617	0.619
3	Multimodal *(Feature-level Fusion)*	0.938	0.938

Berdasarkan hasil evaluasi pada Tabel 2, model multimodal mencatatkan akurasi sebesar 93,8% dan F1-Score sebesar 0,938. Sementara itu, model berbasis audio memperoleh akurasi 90,4% dengan F1-Score 0,903, sedangkan model berbasis lirik menghasilkan akurasi 61,7% dan F1-Score 0,619. Meskipun model multimodal menunjukkan nilai evaluasi tertinggi, hasil tersebut perlu ditafsirkan dengan mempertimbangkan perbedaan konfigurasi pengujian. Model audio dievaluasi pada sepuluh kelas genre, sedangkan model lirik dan multimodal menggunakan empat kelas. Perbedaan ini membatasi kesetaraan perbandingan sehingga hasil yang diperoleh belum dapat memastikan bahwa penggabungan fitur audio dan lirik merupakan faktor yang secara langsung meningkatkan kinerja klasifikasi. Oleh karena itu, perbandingan yang lebih objektif memerlukan evaluasi ketiga model menggunakan kelas dan data pengujian yang sama. Evaluasi lebih lanjut terhadap kinerja model multimodal pada setiap kelas genre disajikan pada Tabel 3.

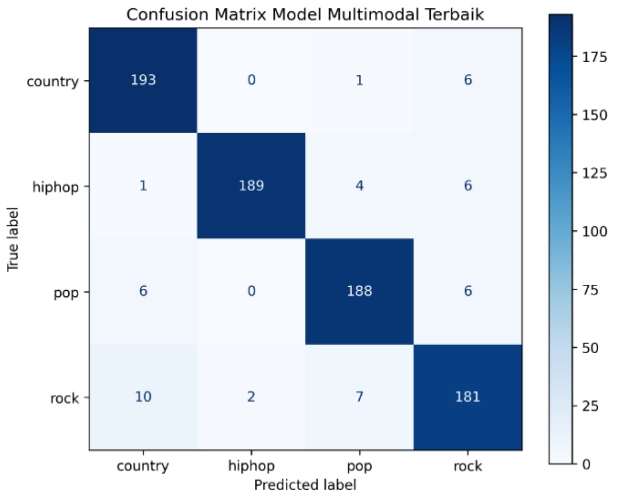
Tabel 3. Performa Model Multimodal pada Setiap Genre

Genre	Precision	Recall	F1-Score	Support
Country	0.92	0.97	0.94	200
Hiphop	0.99	0.95	0.97	200
Pop	0.94	0.94	0.94	200
Rock	0.91	0.91	0.91	200
Akurasi		0.938		

Hasil evaluasi pada Tabel 3 memperlihatkan bahwa model multimodal memiliki kinerja yang berbeda pada setiap genre musik. Genre *Hiphop* mencatatkan nilai *Precision* tertinggi, yaitu 0,99, dengan *F1-Score* sebesar 0,97. Pada genre *Country*, nilai *Recall* mencapai 0,97 dan menjadi yang tertinggi di antara keempat kelas. Sementara itu, genre *Pop* memperoleh nilai *Precision*, *Recall*, dan *F1-Score* masing-masing sebesar 0,94. Genre *Rock* mencatatkan nilai 0,91 pada ketiga metrik tersebut [17].

B. Analisis Confusion Matrix

Analisis *Confusion Matrix* digunakan untuk mengevaluasi ketepatan prediksi model pada setiap kelas serta mengamati kesalahan klasifikasi yang terjadi. Evaluasi ini mencakup 800 sampel data uji yang terbagi secara merata ke dalam empat genre, yaitu *Country*, *Hiphop*, *Pop*, dan *Rock*, dengan masing-masing kelas terdiri atas 200 sampel. Visualisasi hasil klasifikasi tersebut disajikan pada Gambar 2.



Gambar 2. Confusion Matrix Multimodal.

Berdasarkan Gambar 2, model berhasil mengklasifikasikan sebagian besar sampel ke dalam genre yang sesuai dengan label sebenarnya. Jumlah prediksi yang benar mencapai 193 sampel pada *Country*, 189 sampel pada *Hiphop*, 188 sampel pada *Pop*, dan 181 sampel pada *Rock*. Di antara kesalahan yang tercatat, sebanyak 10 sampel *Rock* diprediksi sebagai *Country*. Kesalahan prediksi juga ditemukan pada hubungan antarkelas *Country*, *Pop*, dan *Rock*. Temuan ini menunjukkan bahwa meskipun model multimodal mampu mengenali sebagian besar sampel dengan benar, masih terdapat kesalahan dalam membedakan genre tertentu. Pola tersebut menggambarkan bahwa kinerja klasifikasi tidak sepenuhnya seragam pada seluruh kelas music.

C. Pengujian pada Data Baru (Sportify)

Pengujian menggunakan data baru dilakukan untuk mengevaluasi kinerja model ketika diterapkan pada data yang berasal dari sumber eksternal. Sebanyak 100 sampel lagu dari platform Spotify digunakan dalam pengujian, yang terdiri atas 50 lagu berbahasa Indonesia dan 50 lagu berbahasa Inggris. Data tersebut tidak digunakan pada tahap pelatihan maupun validasi model. Hasil pengujian pada data eksternal disajikan pada Tabel 4.

Tabel 4. Hasil Pengujian Model pada Data Baru dari Spotify

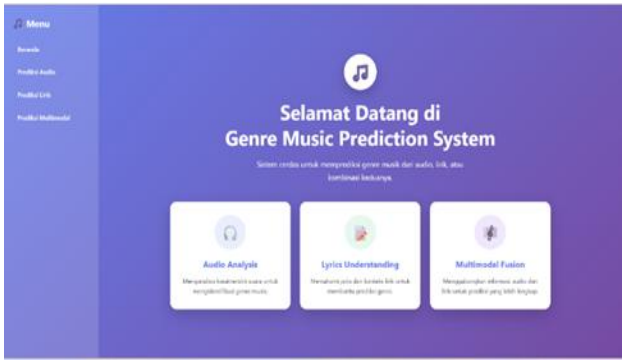
Model	Jumlah Genre	Akurasi
Audio (CNN)	10 Genre	0,680
Lirik (Multilingual BERT)	4 Genre (Country, Hiphop, Pop, Rock)	0,725
Multimodal *(Feature-level Fusion)*	4 Genre (Country, Hiphop, Pop, Rock)	0,650

Berdasarkan hasil evaluasi pada Tabel 4, model berbasis audio mencapai akurasi sebesar 68,0%, sedangkan model berbasis lirik memperoleh 72,5% dan model multimodal menghasilkan 65,0% pada pengujian menggunakan data Spotify. Hasil ini memperlihatkan bahwa kinerja model pada data eksternal berbeda dari hasil evaluasi internal. Model audio dan multimodal mengalami penurunan akurasi dari 90,4% menjadi 68,0% dan dari 93,8% menjadi 65,0%. Sebaliknya, model berbasis lirik menunjukkan peningkatan akurasi dari 61,7% menjadi 72,5%. Perbedaan tersebut perlu ditinjau dengan mempertimbangkan karakteristik data serta prosedur pengujian yang digunakan. Selain itu, pada eksperimen internal, pemasangan audio dan lirik dilakukan berdasarkan kesamaan label genre, sedangkan pengujian eksternal menggunakan audio dan lirik dari lagu yang sama. Perbedaan prosedur ini menjadi salah satu keterbatasan dalam membandingkan hasil kedua pengujian. Dengan demikian, hasil pengujian eksternal menunjukkan bahwa model multimodal belum memberikan kinerja terbaik pada data Spotify, meskipun memperoleh nilai evaluasi tertinggi pada pengujian internal.

D. Implementasi Antarmuka Aplikasi Web

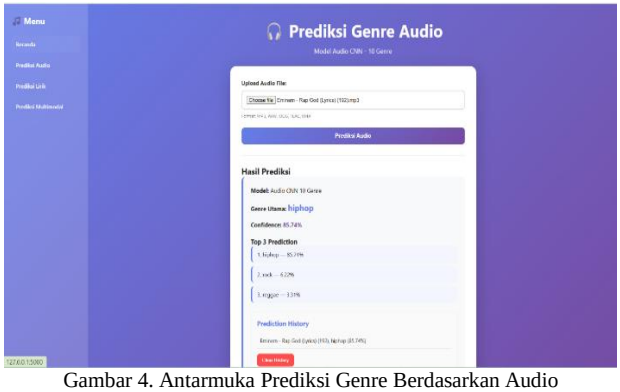
Sistem klasifikasi genre musik diintegrasikan ke dalam aplikasi berbasis web menggunakan *framework* Flask. Implementasi ini bertujuan untuk menyediakan antarmuka yang memungkinkan pengguna melakukan proses klasifikasi secara interaktif berdasarkan jenis data yang tersedia. Aplikasi mendukung tiga skenario klasifikasi, yaitu klasifikasi berdasarkan audio, klasifikasi berdasarkan lirik, dan klasifikasi multimodal yang menggabungkan informasi audio dan lirik.

Halaman utama digunakan sebagai antarmuka awal untuk mengakses sistem dan memilih metode klasifikasi yang akan digunakan. Melalui halaman ini, pengguna dapat menentukan jenis input sesuai dengan kebutuhan proses prediksi. Tampilan halaman utama aplikasi ditunjukkan pada Gambar 3.



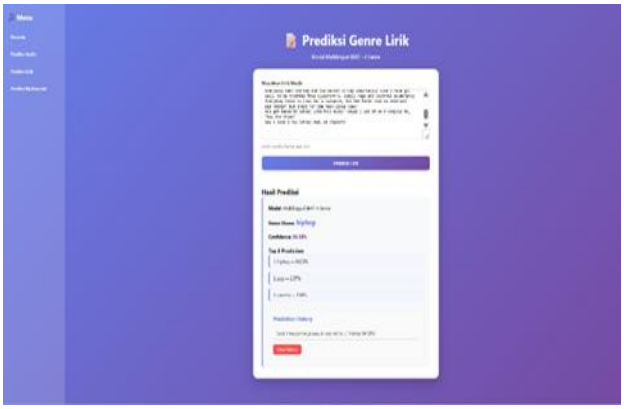
Gambar 3. Halaman Beranda Sistem

Pada klasifikasi berbasis audio, pengguna dapat mengunggah berkas audio dengan format yang didukung oleh sistem, seperti .mp3 atau .wav. Berkas audio yang diberikan kemudian diproses melalui tahapan *preprocessing* dan diubah menjadi representasi *Mel-Spectrogram* sebagai input bagi model CNN. Setelah proses inferensi dilakukan, sistem menampilkan hasil prediksi genre beserta informasi probabilitas dari masing-masing kelas. Antarmuka klasifikasi berbasis audio ditunjukkan pada Gambar 4.



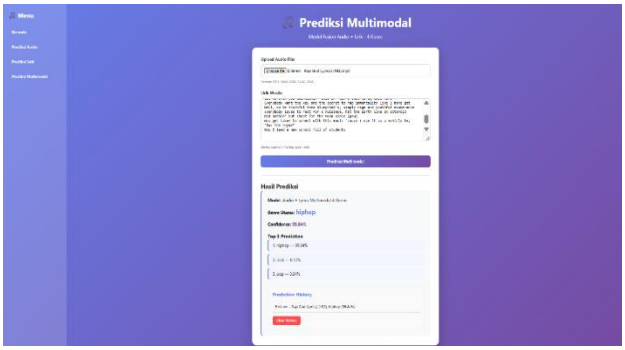
Gambar 4. Antarmuka Prediksi Genre Berdasarkan Audio

Klasifikasi berbasis lirik digunakan untuk melakukan prediksi berdasarkan informasi tekstual dari lagu. Pengguna memasukkan lirik pada antarmuka yang tersedia, kemudian teks tersebut diproses melalui tahapan *preprocessing* sebelum diberikan kepada model Multilingual BERT. Model selanjutnya melakukan proses klasifikasi berdasarkan representasi semantik dari teks yang diberikan. Hasil prediksi ditampilkan pada antarmuka beserta informasi probabilitas untuk setiap kelas. Tampilan antarmuka klasifikasi berbasis lirik ditunjukkan pada Gambar 5.



Gambar 5. Antarmuka Prediksi Genre Berdasarkan Lirik

Selain klasifikasi menggunakan satu modalitas, sistem juga mendukung klasifikasi multimodal yang memanfaatkan informasi audio dan lirik secara bersamaan. Pada proses ini, masing-masing modalitas diproses menggunakan model yang sesuai untuk memperoleh representasi fitur. Selanjutnya, fitur dari kedua modalitas digabungkan menggunakan pendekatan *feature fusion* sebelum digunakan dalam proses klasifikasi. Hasil prediksi kemudian ditampilkan pada antarmuka beserta informasi hasil klasifikasi. Tampilan klasifikasi multimodal ditunjukkan pada Gambar 6.



Gambar 6. Antarmuka Prediksi Genre Multimodal Berdasarkan Audio dan Lirik

Pada setiap proses klasifikasi, sistem menampilkan informasi hasil prediksi yang dapat mencakup genre hasil klasifikasi, probabilitas prediksi, serta *Top 3 Prediction* yang menunjukkan tiga kelas dengan nilai probabilitas tertinggi. Selain itu, tersedia *prediction history* yang mencatat riwayat hasil pengujian yang telah dilakukan. Fitur-fitur tersebut digunakan untuk menyajikan informasi hasil inferensi secara lebih terstruktur.

Implementasi antarmuka berbasis Flask menunjukkan bahwa model klasifikasi yang telah dikembangkan dapat diintegrasikan ke dalam sistem berbasis web untuk mendukung proses inferensi pada tiga pendekatan, yaitu audio, lirik, dan multimodal. Integrasi tersebut memungkinkan model yang dikembangkan digunakan melalui antarmuka yang menyediakan alur input dan penyajian hasil prediksi secara terstruktur.

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Penelitian ini mengembangkan sistem klasifikasi genre musik berbasis multimodal dengan memanfaatkan *Convolutional Neural Network* (CNN) untuk mengekstraksi fitur audio dari representasi *Mel-Spectrogram* dan *Multilingual BERT* untuk memproses lirik berbahasa Indonesia dan Inggris. Kedua jenis fitur tersebut digabungkan melalui pendekatan *feature-level fusion*. Pada pengujian internal, model multimodal memperoleh akurasi sebesar 93,8% dengan *F1-Score* 0,938. Sementara itu, model berbasis audio mencapai akurasi 90,4% dengan *F1-Score* 0,903, sedangkan model berbasis lirik memperoleh akurasi 61,7% dengan *F1-Score* 0,619.

Pada pengujian eksternal menggunakan 100 lagu dari Spotify, model audio menghasilkan akurasi sebesar 68,0%, model lirik sebesar 72,5%, dan model multimodal sebesar 65,0%. Hasil tersebut memperlihatkan adanya perbedaan kinerja antara evaluasi internal dan eksternal. Meskipun model multimodal memberikan hasil tertinggi pada pengujian internal, model tersebut belum menunjukkan kinerja terbaik pada data eksternal. Interpretasi perbandingan ini juga perlu mempertimbangkan perbedaan konfigurasi kelas serta mekanisme pemasangan audio dan lirik antara kedua pengujian. Pada eksperimen internal, audio dan lirik dipasangkan berdasarkan kesamaan label genre, sedangkan pada pengujian eksternal keduanya berasal dari lagu yang sama.

Sistem klasifikasi yang dikembangkan telah diimplementasikan dalam aplikasi web berbasis Flask yang mendukung masukan audio, lirik, maupun kombinasi kedua modalitas. Penelitian selanjutnya disarankan untuk menggunakan pasangan audio dan lirik yang memiliki identitas lagu yang sesuai serta menerapkan konfigurasi kelas dan data pengujian yang setara pada seluruh model. Perbaikan tersebut diperlukan agar evaluasi dapat memberikan gambaran yang lebih objektif mengenai kontribusi penggabungan fitur audio dan lirik terhadap kinerja klasifikasi genre musik.

B. Saran

Perluasan dataset dengan menambahkan lagu dari berbagai wilayah dan budaya di Indonesia dapat dilakukan untuk meningkatkan keragaman karakteristik musik dan lirik. Penambahan variasi lirik berbahasa Indonesia dan Inggris dengan karakteristik penggunaan bahasa yang lebih beragam juga dapat dilakukan untuk memperkaya representasi data tekstual. Selain itu, penggunaan arsitektur berbasis *Transformer* untuk pemrosesan audio, seperti Audio *Spectrogram Transformer* (AST), dapat dieksplorasi sebagai alternatif terhadap CNN. Penambahan jumlah dan variasi genre serta penggunaan data eksternal yang lebih beragam juga dapat dilakukan untuk memperoleh evaluasi model yang lebih komprehensif.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Bapak Ilham Julian Efendi, S.T., M.T. dan Ibu Dr. Andi Tenriawaru, S.Si., M.Si. selaku dosen pembimbing atas bimbingan, arahan, dan masukan selama pelaksanaan penelitian. Ucapan terima kasih juga disampaikan kepada seluruh Civitas Akademika Program Studi Ilmu Komputer Universitas Halu Oleo atas dukungan dan fasilitas yang diberikan selama proses penelitian.

DAFTAR PUSTAKA

[1] A. Alamsyah, F. Ardiansyah, and A. Kholiq, "Music Genre Classification Using Mel Frequency Cepstral Coefficients and Artificial Neural Networks: A Novel Approach," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 937–948, Dec. 2024, doi: 10.15294/sji.v11i4.13660.

[2] S. M. Fardhani, Y. Wihardi, and E. Piantari, "Klasifikasi Genre Musik Dengan Mel Frequency Cepstral Coefficient Dan Spektogram Menggunakan Convolutional Neural Network," 2021. [Online]. Available: https://ejournal.upi.edu/index.php/JATIKOM

[3] S. P. Martinez, M. Zimmermann, M. S. Offermann, and F. Reither, "Exploring Genre and Success Classification through Song Lyrics using DistilBERT: A Fun NLP Venture," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.21068

[4] R. Liyanarachchi and A. Joshi, "A Survey on Multimodal Music Emotion Recognition,"arXiv preprintarXiv:2504.18799,2025, doi:10.48550/arXiv.2504.18799.

[5] Elan Suherlan, Shindy Arti, Siti Nabilah, and Zahwah Hazimah, "Penerapan Flask Framework untuk Deployment Model Machine Learning dalam Mendukung Analisis Adaptasi Mahasiswa Pada Pembelajaran Daring," *Pinter : Jurnal Pendidikan Teknik Informatika dan Komputer*, vol. 9, no. 1, pp. 110–117, Jun. 2025, doi: 10.21009/pinter.9.1.15.

[6] K. Xu, "Recognition and Classification Model of Music Genres and Chinese Traditional Musical Instruments Based on Deep Neural Networks," *Sci. Program.*, vol. 2021, 2021, doi: 10.1155/2021/2348494.

[7] S. M. Al Sasongko, S. Tsaury, S. Ariessaputra, and S. Ch, "Mel Frequency Cepstral Coefficients (MFCC) Method and Multiple Adaline Neural Network Model for Speaker Identification," *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 4, p. 2306, Dec. 2023, doi: 10.62527/joiv.7.4.1376.

[8] T C. Ba, T. D. T. Le, and L.T. Van, "Music genre classification using deep neural network and data augmentation," Entertainment Computing, vol. 53, Art. no. 100929, 2025, doi: 10.1016/j.entcom.2025.100929.

[9] T. Alcántara, O. García-Vázquez, M. Hernandez, H. Calvo, and A. Desiderio, "LyricScraper: A Dataset of Spanish Song Lyrics Created via Web Scraping and Dual-labeling for LLM Classification," *Computacion y Sistemas*, vol. 28, no. 4, pp. 2251–2260, 2024, doi: 10.13053/CyS-28-4-5292.

[10] T. F. Tavares and F. J. Ayres, "*Multi-label Cross-lingual automatic music genre classification from lyrics with Sentence* BERT," Jan. 2025, [Online]. Available: http://arxiv.org/abs/2501.03769.

[11] M. Krichen, "Convolutional Neural Networks: A Survey," *Computers*, vol. 12, no. 8, Aug. 2023, doi: 10.3390/computers12080151.

[12] C. R. Wairata, E. R. Swedia, and M. Cahyanti, "Pengklasifikasian Genre Musik Indonesia Menggunakan *Convolutional Neural Network*," *Sebatik*, vol. 25, no. 1, pp. 255–261, Jun. 2021, doi: 10.46984/sebatik.v25i1.1286.

[13] S. Aum and S. Choe, "srBERT: automatic article classification model for systematic review using BERT," *Syst. Rev.*, vol. 10, no. 1, Dec. 2021, doi: 10.1186/s13643-021-01763-w.

[14] O. Galal, A. H. Abdel-Gawad, and M. Farouk, "Federated Freeze BERT for text classification," *J. Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00885-x.

[15] X. Zhang, G. Ru, J. Wang, N. Cheng, and J. Xiao, "Meningkatkan Klasifikasi Genre Musik dari Sifat Multi-Modal Musik dan Perspektif Korelasi Genre," in Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, doi: 10.1109/ICASSP49357.2023.10097241.

[16] D. J. Evan and P. O. N. Saian, "Implementasi Python Framework Flask pada Modul Transfer Out Toko di Pt Xyz," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 4, pp. 1121–1131, Nov. 2023, doi: 10.29100/jipi.v8i4.4020.

[17] S. N. Luqman *et al.*, "Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM," *Jurnal Sistem Cerdas*, vol. 4, no. 2, pp. 114-125, 2021, doi: 10.37396/jsc.v4i2.162.