Optimasi Convolutional Neural Network pada Klasifikasi *Mood* Musik Berbasis Fitur Audio

Merry Royanti Manalu[a1], Made Agung Raharja[a2], Ngurah Agus Sanjaya ER[a3] I Dewa Made Bayu Atmaja Darmawan[a4]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]merryroyantimanalu24@gmail.com
[2]made.agung@unud.ac.id
[3]agus_sanjaya@unud.ac.id
[4]dewabayu@unud.ac.id

Abstract

This study develops an audio-based music mood classification system using a Convolutional Neural Network (CNN) implemented independently without deep learning frameworks in the core architecture and model training process. This approach was chosen so that the model learning process, from forward propagation to parameter updates, can be understood more thoroughly. The GTZAN dataset is used, comprising 999 audio files. Each audio sample is processed at a sampling rate of 22050 Hz, converted to mono, and standardized into a 30-second segment. Audio representation is constructed as a 71-dimensional feature vector consisting of MFCC, spectral features, spectral contrast, chroma features, tonnetz, energy-related features, and tempo, and then standardized using StandardScaler. Class labels are defined into three mood categories, namely positive, negative, and neutral. This study evaluates three CNN architecture variants (5 layer, 7 layer, and 9 layer architectures) and applies training parameter settings including class weighting, label smoothing, Gaussian noise, and early stopping by monitoring the validation Macro F1-score. Experimental results indicate that hyperparameter tuning improves validation performance compared to the baseline, while fine-tuning provides performance improvement under certain configurations. The best model is obtained from the 7 layer architecture after fine-tuning, achieving an accuracy of 0,97 and a Macro F1-score of 0,97. These results indicate that the extracted audio features are capable of representing music mood characteristics effectively. Furthermore, architectural variation has a positive impact on classification performance; however, increasing network depth does not necessarily lead to consistently better performance. The system is integrated into a Flask-based web application for inference and result presentation.

Keywords: *Music Mood Classification, Audio Features, Convolutional Neural Network, Model Optimization, Music Information Retrieval*

1. Pendahuluan

Perkembangan ekosistem musik digital meningkatkan kebutuhan akan pengelompokan lagu berdasarkan karakter afektif. Klasifikasi *mood* musik dapat dimanfaatkan untuk personalisasi rekomendasi, kurasi *playlist*, dan pengembangan aplikasi dalam bidang *Music Information Retrieval* (MIR). Dalam penelitian ini, *mood* dipahami sebagai kondisi afektif yang relatif lebih stabil dan bertahan lebih lama dibandingkan emosi [1]. Selain itu, *mood* positif, negatif, dan netral

p-ISSN: 2301-5373
e-ISSN: 2654-5101

digunakan sebagai kategori teknis untuk kepentingan klasifikasi, bukan sebagai penilaian normatif terhadap baik atau buruknya suatu emosi.

Dalam konteks musik, klasifikasi *mood* bertujuan mengidentifikasi kecenderungan afektif yang direpresentasikan oleh sebuah lagu melalui karakteristik musikal yang dapat diukur. Pendekatan berbasis audio menjadi relevan karena memungkinkan informasi afektif diekstraksi langsung dari sinyal musik. Pada penelitian ini, representasi audio dibangun sebagai vektor fitur berdimensi 71 yang mencakup *Mel Frequency Cepstral Coefficients* (MFCC), *spectral features*, *spectral contrast*, *chroma features*, *tonnetz*, fitur energi, dan tempo. Kombinasi fitur tersebut dirancang untuk menangkap timbre, distribusi energi frekuensi, harmoni, dinamika, dan ritme yang berkaitan dengan persepsi *mood* musik. Di dalam sistem, keluaran utama dirancang menjadi tiga kelas, yaitu positif, negatif, dan netral, serta dilengkapi penyempurnaan menjadi sepuluh kategori turunan untuk memberikan interpretasi yang lebih spesifik.

Sejumlah penelitian terdahulu menunjukkan bahwa klasifikasi *mood* atau karakter musik dapat dilakukan dengan berbagai pendekatan. Pendekatan *Convolutional Neural Network* (CNN) telah digunakan untuk klasifikasi suasana hati musik berbasis web dengan masukan berupa representasi citra dari audio [2]. Pendekatan lain memanfaatkan MFCC dengan *Backpropagation Neural Network* untuk klasifikasi *mood* musik [3]. MFCC juga digunakan bersama *K-Nearest Neighbor* untuk tujuan serupa [4]. Selain itu, CNN menunjukkan potensi yang baik pada tugas klasifikasi musik lain, seperti klasifikasi genre musik Indonesia [5]. Atribut audio tingkat tinggi seperti energi, tempo, dan *valence* juga telah digunakan untuk memprediksi *mood* musik pop [6].

Studi lain juga menunjukkan bahwa arsitektur CNN yang dirancang secara khusus dapat memberikan performa yang lebih baik dibandingkan model pralatih pada tugas klasifikasi musik berbasis dataset GTZAN [7]. Penelitian lain pada domain musik juga menunjukkan bahwa informasi konteks dan preferensi pengguna dapat dimanfaatkan dalam pengembangan sistem musik yang lebih adaptif dan aplikatif [8]. Selain itu, kajian *state-of-the-art* pada bidang *Music Emotion Recognition* menunjukkan bahwa kombinasi fitur audio dan pendekatan pembelajaran mesin memiliki peran penting dalam proses pemetaan karakter afektif musik [9].

Penelitian lain juga menunjukkan bahwa fitur audio yang diekstraksi menggunakan pendekatan *Short-Time Fourier Transform* (STFT) dan diproses menggunakan algoritma *Random Forest* dapat digunakan untuk klasifikasi genre musik [10]. Pada konteks yang berbeda, metode *K-Means Clustering* telah diterapkan untuk mengelompokkan lagu populer berdasarkan karakteristik tertentu [11], sedangkan teknik pengolahan sinyal audio juga dimanfaatkan dalam pengembangan aplikasi sintesis suara instrumen tradisional menggunakan metode *Double Frequency Modulation* (DFM) [12]. Temuan-temuan tersebut menunjukkan bahwa fitur audio, pengolahan sinyal, dan model pembelajaran mesin dapat diterapkan pada berbagai permasalahan dalam domain musik, termasuk klasifikasi genre, pengelompokan lagu, sintesis audio, dan pemetaan karakter afektif musik.

Meskipun berbagai penelitian telah menunjukkan keberhasilan penggunaan CNN, MFCC, maupun metode pembelajaran mesin lainnya untuk klasifikasi *mood* dan karakteristik musik, masih terdapat beberapa keterbatasan yang belum banyak dibahas. Sebagian besar penelitian terdahulu menggunakan representasi data berbentuk citra spektrum atau berfokus pada pemanfaatan satu jenis fitur audio tertentu, sehingga kontribusi kombinasi fitur audio yang lebih komprehensif terhadap klasifikasi *mood* belum dievaluasi secara mendalam.

Selain itu, implementasi CNN umumnya memanfaatkan *framework deep learning* sehingga proses pembelajaran model cenderung diperlakukan sebagai *black box* dan kurang memberikan gambaran rinci mengenai mekanisme pembelajaran yang terjadi. Oleh karena itu, masih terdapat kebutuhan untuk mengevaluasi efektivitas CNN yang diimplementasikan secara mandiri pada representasi vektor fitur audio berdimensi ringkas serta memanfaatkan pendekatan *weak labeling* untuk membangun sistem klasifikasi *mood* musik berbasis audio.

Berdasarkan kondisi tersebut, penelitian ini mengembangkan sistem klasifikasi *mood* musik berbasis audio menggunakan vektor fitur berdimensi 71. Sistem ini memanfaatkan pendekatan *weak labeling* melalui pemetaan genre musik ke dalam kategori *mood*, serta membandingkan tiga variasi arsitektur CNN yang diimplementasikan secara mandiri, yaitu arsitektur *5 layer*, 7 *layer*, dan 9 *layer*, untuk mempelajari pola pada vektor fitur audio. Selanjutnya, penelitian ini mengevaluasi pengaruh *hyperparameter tuning* dan *fine-tuning* terhadap performa model menggunakan *accuracy* dan *Macro F1-score*, dengan penekanan pada *Macro F1-score* sebagai metrik utama karena evaluasi dilakukan pada data dengan komposisi kelas yang tidak sepenuhnya seimbang.

2. Metodologi Penelitian

Dataset utama yang digunakan dalam penelitian ini adalah GTZAN. Dari total 1000 berkas audio, satu berkas yang rusak dikeluarkan dari proses pengolahan sehingga jumlah data yang digunakan menjadi 999 audio. Seluruh audio diproses pada *sampling rate* 22050 Hz, dikonversi menjadi mono, dan diseragamkan menjadi segmen analisis berdurasi 30 detik. Meskipun seluruh audio pada dataset GTZAN memiliki durasi 30 detik, proses penyeragaman tetap dilakukan untuk memastikan konsistensi panjang sinyal audio pada seluruh data masukan. Jika durasi audio sedikit lebih pendek dari 30 detik, sistem menerapkan *zero padding*, sedangkan jika durasi audio melebihi 30 detik, sistem hanya menggunakan 30 detik pertama. Tahap ini dilakukan untuk memastikan konsistensi masukan pada proses ekstraksi fitur dan pelatihan model.

Tahap berikutnya adalah ekstraksi fitur audio untuk menghasilkan representasi numerik yang dapat dipelajari oleh model. Fitur yang digunakan disusun dalam vektor berdimensi 71 dan dirancang untuk merepresentasikan timbre, distribusi energi frekuensi, harmoni, dinamika, dan ritme. Vektor ini kemudian menjadi masukan utama bagi model CNN dalam proses pelatihan dan evaluasi klasifikasi *mood* musik. Komposisi fitur audio yang digunakan dalam penelitian ini disajikan pada Tabel 1.

Tabel 1. Komposisi Fitur Audio

Kelompok Fitur	Agregasi	Dimensi
MFCC	Nilai rata-rata dan standar deviasi dari 20 koefisien MFCC	40
Spectral features	Nilai rata-rata spectral centroid, spectral bandwidth, dan spectral roll-off	3
Spectral contrast	Nilai rata-rata pada setiap pita frekuensi	7
Chroma features	Nilai rata-rata distribusi chroma	12
Tonnetz	Nilai rata-rata representasi tonal centroid	6
Energi dan Dinamika	Nilai rata-rata RMS Energy dan Zero Crossing Rate	2
Ritme	Nilai rata-rata tempo	1
Total		**71**

Karena dataset GTZAN tidak menyediakan label *mood* secara langsung, penelitian ini menerapkan pendekatan *weak labeling* dengan memetakan informasi genre musik ke dalam tiga kategori *mood* utama, yaitu positif, negatif, dan netral. Pemetaan dilakukan berdasarkan

p-ISSN: 2301-5373
e-ISSN: 2654-5101

kecenderungan emosi yang umum diasosiasikan dengan masing-masing genre musik, sehingga label yang dihasilkan dapat digunakan sebagai target pelatihan model CNN. Sebagai dasar pendukung, penyusunan kategori *mood* dan pemetaan emosi mengacu pada konsep psikologi afektif yang relevan.

Pemetaan genre ke dalam kategori *mood* dilakukan berdasarkan kecenderungan karakteristik musikal dan persepsi afektif yang secara umum diasosiasikan dengan masing-masing genre musik. Genre dengan dominasi tempo cepat, energi tinggi, dan nuansa musikal yang cenderung ceria dipetakan ke kategori positif. Genre dengan karakter musikal yang lebih agresif, intens, atau bernuansa emosional kuat dipetakan ke kategori negatif. Sementara itu, genre dengan karakter musikal yang relatif tenang, seimbang, dan stabil dipetakan ke kategori netral. Rincian pemetaan genre terhadap kategori *mood* disajikan pada Tabel 2.

Tabel 2. Pemetaan Genre terhadap Kategori Mood

Genre Musik	Kategori Mood
Disco	Positif
Pop	Positif
Reggae	Positif
Country	Positif
Hiphop	Positif
Blues	Negatif
Metal	Negatif
Rock	Negatif
Classical	Netral
Jazz	Netral

Selain tiga kelas utama tersebut, penelitian ini juga menerapkan penyempurnaan (*refinement*) kategori *mood* turunan menjadi sepuluh kategori sebagai informasi interpretatif tambahan pada tahap inferensi. Kelas positif terdiri atas *happy*, *energetic*, *romantic*, dan *joyful*. Kelas negatif terdiri atas *sad*, *angry*, *desperate*, dan *anxious*. Kelas netral terdiri atas *calm* dan *nostalgic*. Kategori turunan ini tidak mengubah keluaran utama model, tetapi digunakan untuk memberikan interpretasi yang lebih spesifik terhadap hasil klasifikasi. Jika lirik tersedia, sistem melakukan pencocokan *keyword* untuk menentukan kategori turunan yang paling sesuai, sedangkan jika lirik tidak tersedia, sistem menggunakan kategori *default* berdasarkan kelas *mood* utama, yaitu *happy* untuk kelas positif, *sad* untuk kelas negatif, dan *calm* untuk kelas netral. Daftar kategori dan *keyword* utama tersebut disusun oleh peneliti serta ditinjau dan divalidasi oleh pakar psikologi agar tetap sesuai dengan konsep psikologi afektif.

Setelah proses pelabelan selesai, data dibagi menjadi 90% data latih dan 10% data validasi, kemudian distandardisasi menggunakan StandardScaler. StandardScaler dilatih pada data latih dan digunakan kembali untuk mentransformasikan data validasi agar proses pelatihan lebih stabil dan konsisten. Model CNN menerima masukan dalam bentuk tensor berukuran (1, 1, 71). Representasi ini dipilih karena fitur audio direpresentasikan sebagai vektor berdimensi 71 yang kemudian dibentuk menjadi tensor agar dapat diproses oleh lapisan konvolusi dua dimensi. Pendekatan ini memungkinkan model mempelajari hubungan lokal antarfitur audio melalui operasi konvolusi tanpa memerlukan representasi citra spektrum. Model diimplementasikan secara mandiri tanpa *framework deep learning* pada inti arsitektur dan proses pelatihan, sehingga

mekanisme komputasi, rancangan lapisan, dan proses pembelajaran model dapat dikendalikan secara penuh. Pendekatan ini sejalan dengan prinsip dasar kecerdasan buatan dan pembelajaran mesin yang menekankan pemahaman terhadap mekanisme pembelajaran model secara menyeluruh [13]. Penelitian ini menguji tiga variasi arsitektur, yaitu arsitektur 5 *layer*, 7 *layer*, dan 9 *layer*.

Pengaturan parameter pelatihan menerapkan *weighted cross entropy*, *class weighting*, *Gaussian noise*, *label smoothing*, dan *early stopping*, dengan *Macro F1-score* validasi sebagai metrik pemantauan utama. Pemilihan *Macro F1-score* sebagai metrik utama dilakukan karena evaluasi multi-kelas pada penelitian ini mempertimbangkan perbedaan proporsi sampel antar kelas, sehingga performa tiap kelas tetap diperhitungkan secara seimbang. Rancangan eksperimen dilakukan secara bertahap melalui pelatihan baseline untuk setiap varian, *hyperparameter tuning* untuk memperoleh konfigurasi yang lebih sesuai, serta *fine-tuning* sebagai evaluasi lanjutan pada konfigurasi terpilih.

Evaluasi model dilakukan menggunakan *accuracy*, *Macro F1-score*, dan *confusion matrix*. Selain itu, *hyperparameter tuning* dilakukan pada setiap varian dengan menguji kombinasi *learning rate*, *weight decay*, *dropout pada dense layer*, dan *label smoothing*. Setelah konfigurasi terbaik diperoleh, *fine-tuning* tambahan diuji untuk menilai apakah penyesuaian parameter lebih lanjut masih dapat meningkatkan kemampuan generalisasi model.

Mekanisme utama CNN dalam penelitian ini dijelaskan melalui formulasi matematis operasi konvolusi, fungsi aktivasi ReLU, dan fungsi *softmax* pada lapisan keluaran. Formulasi ini digunakan untuk memperjelas proses komputasi dasar pada arsitektur yang diterapkan.

Secara matematis, operasi konvolusi pada CNN dalam penelitian ini mengikuti formulasi konvolusi dua dimensi. Meskipun fitur audio awal direpresentasikan sebagai vektor berdimensi 71, pada tahap masukan ke model vektor tersebut dibentuk menjadi tensor berukuran (1 × 1 × 71) sehingga dapat diproses menggunakan lapisan konvolusi dua dimensi. Untuk masukan dua dimensi x dan kernel w berukuran $(m \times n)$, keluaran konvolusi pada posisi (i, j) dapat dinyatakan sebagai:

$$y(i,j) = \sum(p=1)^{m} \sum(q=1)^{n} X(i+p-1, j+q-1) W(p,q) + b \qquad (1)$$

dengan b sebagai bias. Keluaran konvolusi selanjutnya diproses oleh fungsi aktivasi ReLU yang didefinisikan sebagai:

$$ReLU(z) = max(0, z) \qquad (2)$$

Pada lapisan keluaran, nilai prediksi untuk masing-masing kelas dihitung menggunakan fungsi *softmax*. Untuk vektor logit z dan kelas ke-i, fungsi *softmax* didefinisikan sebagai:

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \qquad (3)$$

dengan C merupakan jumlah kelas. Pada penelitian ini, C = 3.

3. Hasil dan Pembahasan

Pelatihan baseline menunjukkan bahwa seluruh variasi arsitektur CNN mampu mempelajari pola dasar dari fitur audio yang digunakan pada penelitian ini. Nilai *Macro F1-score baseline* pada ketiga arsitektur juga menunjukkan performa awal yang cukup baik, namun masih terdapat ruang peningkatan terutama pada keseimbangan performa antar kelas. Kondisi ini menunjukkan bahwa

p-ISSN: 2301-5373
e-ISSN: 2654-5101

konfigurasi awal model belum sepenuhnya optimal untuk menangani karakteristik data hasil *weak labeling*. Oleh karena itu, dilakukan *hyperparameter tuning* untuk memperoleh konfigurasi parameter pelatihan yang lebih sesuai dengan karakteristik data dan representasi fitur audio yang digunakan.

Selain itu, perbedaan hasil awal antar arsitektur menunjukkan bahwa respons masing-masing model terhadap data tidak sepenuhnya sama. Hal ini menjadi dasar untuk mengevaluasi lebih lanjut apakah penyesuaian parameter dapat meningkatkan stabilitas dan kemampuan generalisasi model pada setiap arsitektur. Dalam konteks penelitian ini, proses *tuning* juga berperan untuk mengidentifikasi konfigurasi yang paling sesuai bagi representasi fitur audio berdimensi 71 yang digunakan sebagai masukan model. Dengan demikian, evaluasi tidak hanya berfokus pada perbandingan antar arsitektur, tetapi juga pada sejauh mana setiap model mampu dioptimalkan melalui pengaturan parameter pelatihan.

Pada tahap *hyperparameter tuning*, penelitian ini mengeksplorasi beberapa parameter pelatihan yang diterapkan secara konsisten pada seluruh variasi arsitektur. Parameter yang diuji meliputi *optimizer* (Adam dan SGD), *learning rate*, *weight decay*, *dropout* pada *dense layer*, serta *label smoothing*. Kombinasi parameter tersebut digunakan untuk mengevaluasi pengaruh konfigurasi pelatihan terhadap stabilitas pembelajaran dan kemampuan generalisasi model. Proses eksplorasi dilakukan secara sistematis dengan menjalankan berbagai kombinasi parameter pada masing-masing arsitektur CNN. Ruang pencarian *hyperparameter* yang digunakan dalam penelitian ini disajikan pada Tabel 3.

Tabel 3. Ruang Pencarian Hyperparameter

Parameter	Nilai yang Diuji
Optimizer	Adam, SGD
Learning Rate	0,0001 ; 0,0002 ; 0,0003 ; 0,0004 ; 0,0005
Weight Decay	0,0005 ; 0,0010 ; 0,0015
Dropout (Dense)	0,20 ; 0,30 ; 0,40
Label Smoothing	0,01 ; 0,03 ; 0,05

Setelah konfigurasi terbaik pada setiap arsitektur diperoleh melalui proses *baseline*, *hyperparameter tuning*, dan *fine-tuning*, dilakukan pemilihan konfigurasi akhir berdasarkan performa validasi terbaik. Ringkasan performa akhir dari konfigurasi terbaik pada masing-masing arsitektur disajikan pada Tabel 4. Konfigurasi tersebut selanjutnya digunakan sebagai model akhir pada implementasi sistem berbasis web.

Tabel 4. Performa Konfigurasi Terbaik pada Setiap Arsitektur

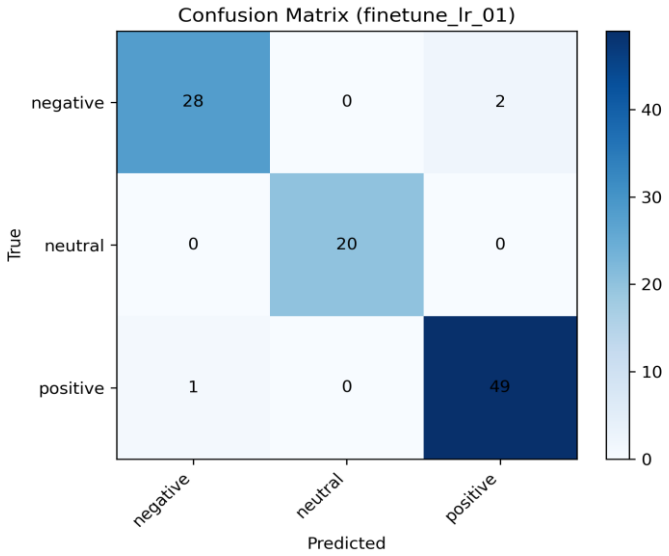
Varian Arsitektur	Konfigurasi terbaik	Best epoch	Accuracy validasi	Macro F1-score validasi
5 Layer	Hyperparameter Tuning	114	0,93	0,9364
7 Layer	Fine-Tuning	145	0,97	0,9731
9 Layer	Hyperparameter Tuning	30	0,95	0,9456

Berdasarkan hasil evaluasi, arsitektur 7 *layer* hasil *fine-tuning* memberikan performa terbaik dengan *accuracy* validasi sebesar 0,97 dan *Macro F1-score* sebesar 0,9731. Pada tahap *fine-tuning*, penyesuaian parameter dilakukan dengan memvariasikan nilai *learning rate* sebesar 0,00005, 0,00010, 0,00015, dan 0,00020 pada konfigurasi terbaik hasil *hyperparameter tuning*. Performa terbaik diperoleh pada *learning rate* 0,00020 dengan epoch terbaik ke-145. Sementara itu, arsitektur 5 *layer* dan 9 *layer* menunjukkan konfigurasi terbaik pada tahap *hyperparameter tuning* dengan *Macro F1-score* masing-masing sebesar 0,9364 dan 0,9456.

Perbedaan hasil akhir pada setiap arsitektur menunjukkan bahwa kedalaman jaringan memengaruhi kemampuan representasi model terhadap fitur audio yang digunakan. Dalam penelitian ini, arsitektur 7 *layer* memberikan performa terbaik karena mampu menghasilkan keseimbangan yang lebih baik antara kapasitas model dan kemampuan generalisasi pada data validasi. Sementara itu, arsitektur yang lebih dalam belum menunjukkan peningkatan performa yang signifikan, yang kemungkinan dipengaruhi oleh ukuran dataset yang relatif terbatas serta bentuk representasi masukan yang telah diringkas ke dalam vektor fitur berdimensi 71. Kondisi tersebut menyebabkan penambahan kompleksitas model belum sepenuhnya memberikan keuntungan tambahan pada proses pembelajaran.

Selain itu, perbedaan antara nilai *accuracy* dan *Macro F1-score* pada beberapa konfigurasi menunjukkan bahwa evaluasi model tidak cukup hanya ditinjau dari *accuracy* saja. Penggunaan *Macro F1-score* sebagai metrik utama membantu memberikan gambaran yang lebih seimbang terhadap performa setiap kelas, terutama pada data hasil *weak labeling* yang tidak sepenuhnya memiliki distribusi kelas yang seimbang. Dengan demikian, hasil penelitian menunjukkan bahwa pengaturan parameter pelatihan dan pemilihan arsitektur yang sesuai berperan penting dalam memperoleh performa klasifikasi *mood* musik yang stabil.

Confusion matrix pada Gambar 1 menunjukkan bahwa model mampu mengenali seluruh kelas *mood* dengan distribusi prediksi yang sangat baik. Kelas netral berhasil dikenali secara sempurna tanpa kesalahan klasifikasi, sedangkan kelas positif dan negatif juga menunjukkan tingkat prediksi yang tinggi dengan jumlah kesalahan yang relatif kecil. Kesalahan klasifikasi yang masih muncul terutama terjadi antara kelas negatif dan positif, namun jumlahnya tidak signifikan dibandingkan total prediksi benar pada masing-masing kelas.



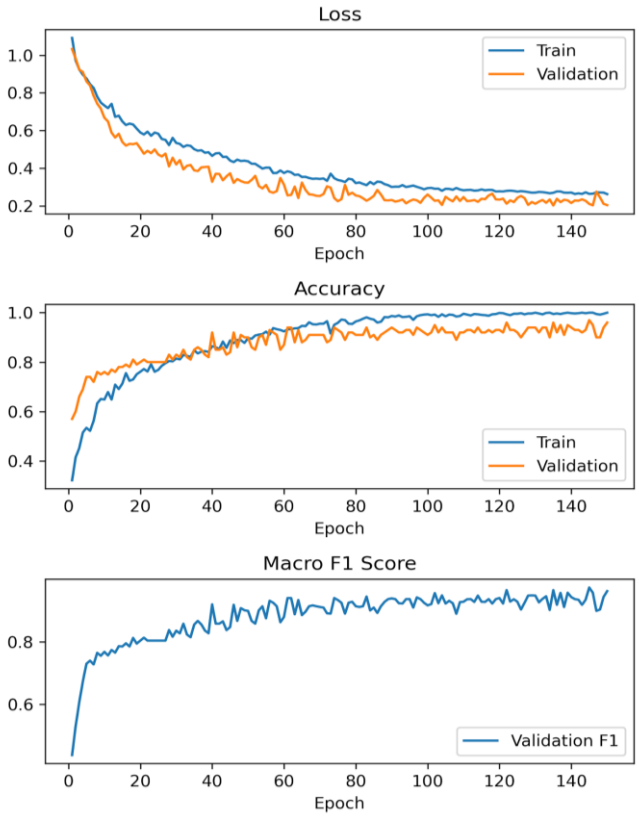
Gambar 1. Confusion Matrix Model Terbaik Arsitektur 7 Layer

Hasil tersebut menunjukkan bahwa representasi fitur audio yang digunakan pada penelitian ini telah mampu membantu model membedakan karakteristik utama antar kelas *mood*. Kombinasi fitur spektral, harmonik, energi, dan ritme memberikan informasi yang cukup representatif untuk mendukung proses klasifikasi. Selain itu, hasil *confusion matrix* juga menunjukkan bahwa

p-ISSN: 2301-5373
e-ISSN: 2654-5101

konfigurasi *fine-tuning* pada arsitektur 7 *layer* mampu meningkatkan kemampuan generalisasi model terhadap data validasi secara lebih konsisten.

Kurva pelatihan pada Gambar 2 menunjukkan bahwa proses pelatihan berlangsung stabil hingga mencapai epoch terbaik. Nilai *loss* mengalami penurunan secara konsisten selama proses pelatihan, sedangkan *accuracy* dan *Macro F1-score* validasi menunjukkan tren peningkatan yang stabil. Pola tersebut menunjukkan bahwa model mampu mempelajari representasi fitur audio dengan baik tanpa mengalami fluktuasi performa yang signifikan pada data validasi.



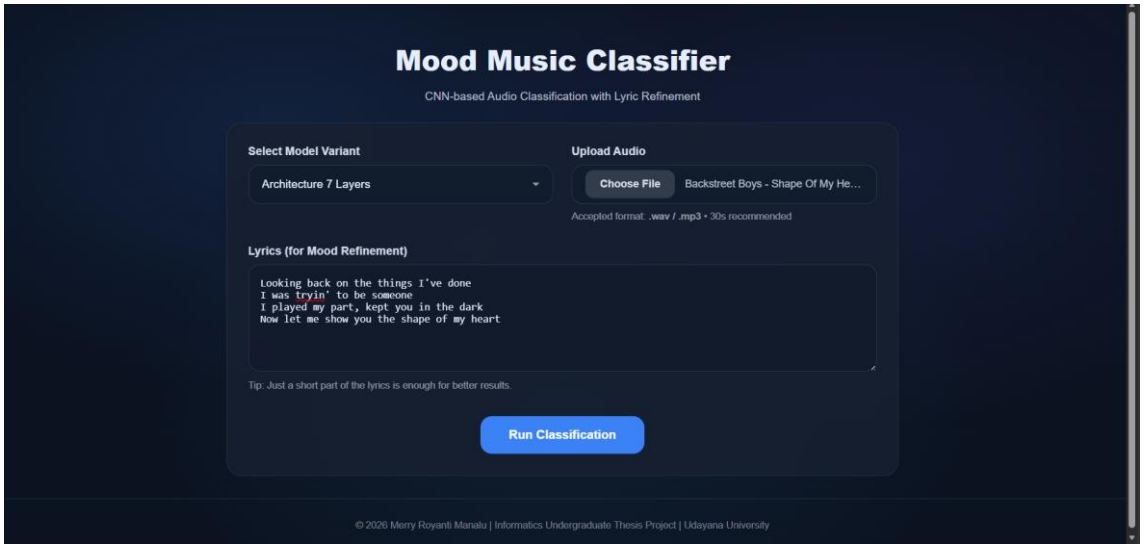
Gambar 2. Kurva Pelatihan Model Terbaik Arsitektur 7 Layer

Selain itu, selisih antara performa pelatihan dan validasi relatif kecil, yang menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data yang digunakan. Kurva *loss* validasi yang tetap stabil hingga akhir pelatihan juga menunjukkan bahwa model tidak mengalami *overfitting* yang signifikan. Hasil ini memperlihatkan bahwa konfigurasi *fine-tuning* pada arsitektur 7 *layer* mampu menghasilkan proses optimasi yang lebih stabil dibandingkan konfigurasi sebelumnya.

Model terbaik selanjutnya diintegrasikan ke dalam aplikasi web berbasis Flask untuk mendukung proses inferensi audio dan penyajian hasil prediksi secara lebih praktis. Pada tahap ini, sistem menyediakan pilihan arsitektur 5 *layer*, 7 *layer*, dan 9 *layer* yang masing-masing menggunakan konfigurasi terbaik hasil pelatihan. Pengguna dapat mengunggah audio, memilih arsitektur model yang akan digunakan, serta menambahkan lirik secara opsional untuk mendukung mekanisme *refinement* kategori *mood* turunan.

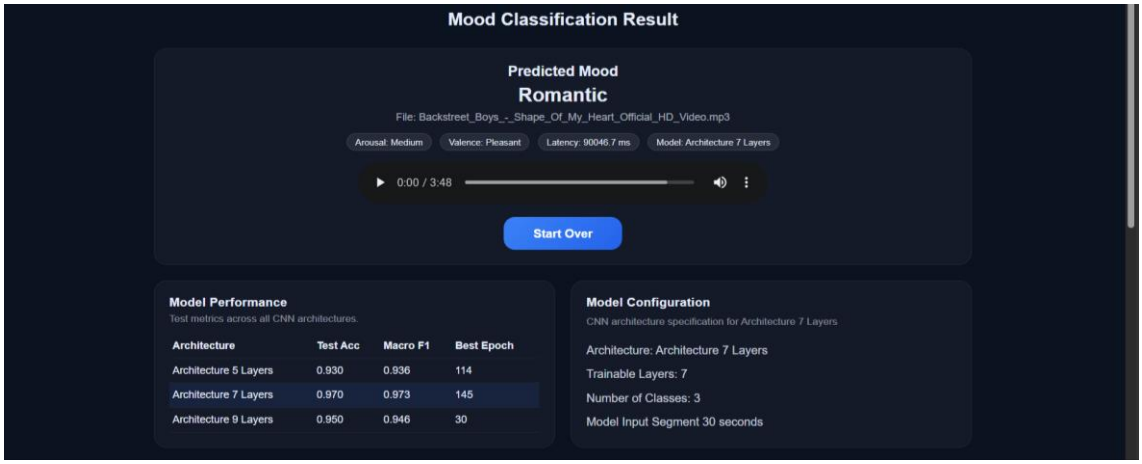
Hasil yang ditampilkan tidak hanya mencakup kelas *mood* utama, tetapi juga informasi interpretatif tambahan dan elemen visual pendukung yang membantu pembacaan hasil prediksi. Integrasi ini menunjukkan bahwa model yang dihasilkan tidak hanya memberikan performa validasi yang baik, tetapi juga dapat diterapkan pada sistem yang mendukung penggunaan secara lebih aplikatif.

Gambar 3 menampilkan implementasi antarmuka sistem berbasis web yang digunakan untuk proses inferensi. Pengguna dapat memilih arsitektur model, mengunggah file audio, serta menambahkan lirik secara opsional. Sistem kemudian menampilkan hasil prediksi *mood* beserta informasi pendukung lainnya. Implementasi ini menunjukkan bahwa model klasifikasi *mood* musik yang dikembangkan dapat diterapkan dalam alur penggunaan yang lebih praktis dan informatif.



Gambar 3. Antarmuka Sistem Klasifikasi Mood Musik

Gambar 4 dan Gambar 5 menampilkan implementasi antarmuka hasil prediksi pada sistem klasifikasi *mood* musik berbasis web. Pada bagian atas halaman, sistem menampilkan hasil prediksi *mood* utama, nama file audio, *audio player*, serta informasi tambahan seperti *arousal*, *valence*, *latency*, dan arsitektur model yang digunakan. Selain itu, sistem juga menampilkan ringkasan performa model dan konfigurasi arsitektur untuk memberikan konteks tambahan terhadap proses inferensi yang dilakukan.

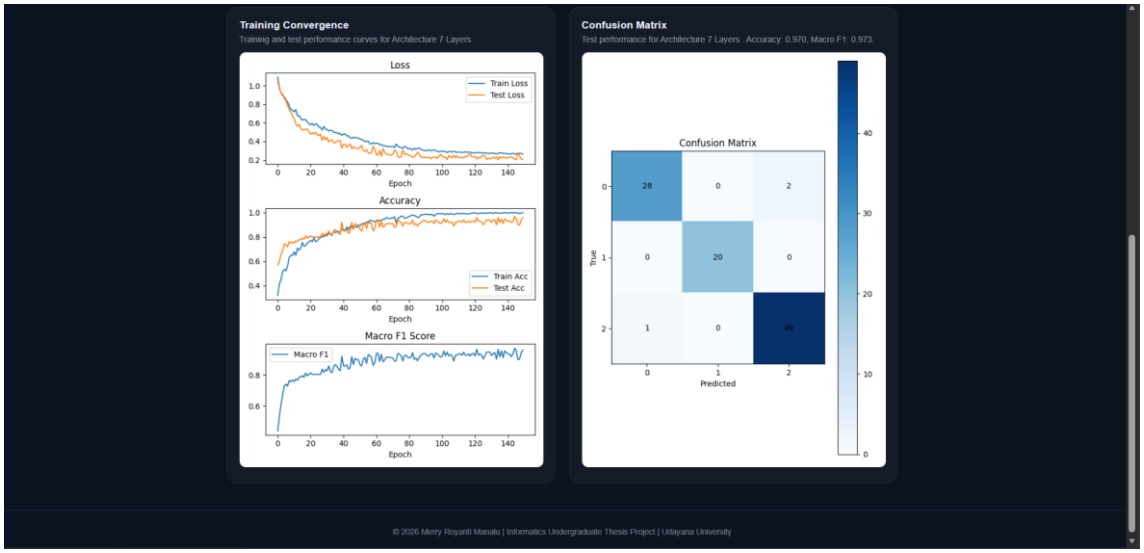


Gambar 4. Antarmuka Halaman Hasil Prediksi Bagian Atas

Pada bagian bawah halaman, sistem menampilkan visualisasi pendukung berupa kurva pelatihan dan *confusion matrix* dari model yang digunakan pada proses inferensi. Visualisasi tersebut membantu pengguna memahami performa model secara lebih informatif, termasuk kestabilan proses pelatihan dan distribusi hasil klasifikasi antar kelas *mood*. Integrasi antara hasil prediksi, informasi model, dan visualisasi evaluasi menunjukkan bahwa sistem tidak hanya berfungsi

p-ISSN: 2301-5373
e-ISSN: 2654-5101

sebagai antarmuka inferensi, tetapi juga sebagai sarana interpretasi performa model secara lebih komprehensif.



Gambar 5. Antarmuka Halaman Hasil Prediksi Bagian Bawah

Secara keseluruhan, implementasi berbasis web menunjukkan bahwa model klasifikasi *mood* musik yang dikembangkan tidak hanya memberikan performa evaluasi yang baik, tetapi juga dapat diterapkan dalam alur penggunaan yang lebih praktis dan aplikatif. Integrasi antara proses inferensi, visualisasi evaluasi, dan penyajian informasi model memperlihatkan bahwa sistem mampu mendukung penggunaan model klasifikasi secara lebih utuh dalam satu antarmuka.

4. Kesimpulan

Penelitian ini berhasil mengembangkan sistem klasifikasi *mood* musik berbasis audio menggunakan *Convolutional Neural Network* (CNN) yang diimplementasikan secara mandiri tanpa *framework deep learning* pada inti arsitektur dan proses pelatihan. Sistem memanfaatkan representasi fitur audio berdimensi 71 untuk mengklasifikasikan *mood* ke dalam tiga kelas utama, yaitu positif, negatif, dan netral. Berdasarkan hasil evaluasi, konfigurasi terbaik diperoleh pada arsitektur 7 *layer* hasil *fine-tuning* dengan *accuracy* validasi sebesar 0,97 dan *Macro F1-score* sebesar 0,9731. Hasil *confusion matrix* dan kurva pelatihan menunjukkan bahwa model mampu mempelajari pola pada fitur audio dengan performa dan kemampuan generalisasi yang baik.

Selain menghasilkan performa klasifikasi yang baik, model terbaik juga berhasil diintegrasikan ke dalam aplikasi web berbasis Flask untuk mendukung proses inferensi audio secara lebih praktis dan aplikatif. Sistem mampu menampilkan hasil prediksi *mood*, informasi interpretatif tambahan, serta visualisasi evaluasi model dalam satu antarmuka. Untuk penelitian selanjutnya, penggunaan dataset dengan label *mood* berbasis audio yang memiliki *ground-truth* secara langsung dan jumlah data yang lebih besar dapat dipertimbangkan untuk meningkatkan kemampuan generalisasi model.

Referensi

[1] D. A. Rifani and D. R. Rahadi, “Ketidakstabilan Emosi dan Mood Masyarakat Dimasa Pandemi Covid-19,” *Jurnal Manajemen Bisnis*, vol. 18, no. 1, 2021.

[2] R. Fasti and D. Avianto, “Implementasi Web Klasifikasi Suasana Hati Berdasarkan Potongan Lagu dengan Memanfaatkan Convolutional Neural Network,” *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, vol. 5, no. 1, pp. 881–892, 2024.

[3] P. I. Maulana, A. Aranta, F. Bimantoro, and I. G. Andika, “Klasifikasi Mood Musik Berdasarkan Mel Frequency Cepstral Coefficients dengan Backpropagation Neural Network,” *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, vol. 5, no. 1, pp. 72–85, 2022.

[4] F. F. Surenggana, A. Aranta, and F. Bimantoro, “Klasifikasi Mood Musik Menggunakan K-Nearest Neighbor dan Mel Frequency Cepstral Coefficients,” *Jurnal Teknologi Informasi, Komputer dan Aplikasinya (JTIKA)*, vol. 4, no. 2, pp. 263–276, 2022.

[5] 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, 2021.

[6] L. Nurhalimah, T. I. Hermanto, and I. Kaniawulan, “Analisis Prediksi Mood Genre Musik Pop Menggunakan Algoritma K-Means dan C4.5,” *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 4, pp. 1006–1013, 2022.

[7] K. I. Pradnya and M. A. Raharja, “Analisis Komparatif Arsitektur CNN dan VGG16 pada Klasifikasi Genre Musik,” Jurnal Nasional Teknologi Informasi dan Aplikasinya, vol. 3, no. 4, pp. 889–898, 2025.

[8] G. A. V. M. Giri, M. L. Radhitya, M. A. Raharja, and I. W. Supriana, “Sistem Rekomendasi Musik Berdasarkan Data Konteks Pada Listening History Musik dan Keterkaitan Artis Indonesia,” Jurnal RESISTOR, vol. 5, no. 1, pp. 86–93, 2022.

[9] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, “Music Emotion Recognition: A State of the Art Review,” Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), pp. 255–266, 2010.

[10] M. R. Manalu and M. A. Raharja, “Analisis dan Klasifikasi Genre Musik Menggunakan Algoritma STFT dan Random Forest,” Jurnal Nasional Teknologi Informasi dan Aplikasinya, vol. 3, no. 1, pp. 205–214, 2024.

[11] P. N. W. Wesnawa and M. A. Raharja, “Pengelompokan Lagu Populer untuk Musik Gym Menggunakan Metode K-Means Clustering,” Jurnal Nasional Teknologi Informasi dan Aplikasinya, vol. 2, no. 4, pp. 861–868, 2024.

[12] M. A. Raharja and I. D. M. B. A. Darmawan, “Rancang Bangun Aplikasi Sintesis Suara Gamelan Gerantang Bali Menggunakan Metode Double Frequency Modulation (DFM),” Jurnal RESISTOR (Rekayasa Sistem Komputer), vol. 4, no. 2, pp. 119–126, 2021.

[13] M. A. Raharja et al., Kecerdasan Buatan dan Soft Computing. Bandung: Widina Media Utama, 2025.