

Klasifikasi Genre Musik Menggunakan CNN Dengan Arsitektur Resnet-50 Dan Gradient Boost LightGBM

DOI: <http://dx.doi.org/10.35889/jutisi.v15i1.3458>

Creative Commons License 4.0 (CC BY – NC)

Roberto Alessandro^{1*}, Tinaliah²

Informatika, Universitas Multi Data Palembang, Palembang, Indonesia

*e-mail Corresponding Author: robertoalexandro_2226250030@mhs.mdp.ac.id

Abstract

The rapid growth of digital music has driven the need for accurate and efficient automated music genre classification systems. This study evaluates a hybrid approach that integrates the ResNet-50 architecture as a feature extractor through transfer learning and LightGBM as a classifier. Using the GTZAN dataset represented as Mel-spectrograms, the research compares the effectiveness of hyperparameter optimization using Random Search and Grid Search methods. Based on performance evaluation, the hybrid scenario optimized with Grid Search yielded the best performance with an accuracy of 81.20%, outperforming the Random Search method. Nevertheless, the overall experimental results reveal that the end-to-end ResNet-50 model still provides superior performance compared to the hybrid approach. This indicates that the deep features from ResNet-50 are highly representative for separating genre classes, such that the addition of an external ensemble classifier does not yield significant improvements, although the hybrid approach still offers valuable empirical insights as a stable alternative model.

Keywords: Convolutional Neural Network; ResNet-50; LightGBM; Mel-Spectrogram; Klasifikasi Genre Musik;

Abstrak

Pertumbuhan pesat musik digital mendorong kebutuhan akan sistem klasifikasi genre musik otomatis yang akurat dan efisien. Penelitian ini mengevaluasi pendekatan hibrida yang mengintegrasikan arsitektur ResNet-50 sebagai pengekstraksi fitur melalui teknik *transfer learning* dan LightGBM sebagai *classifier*. Menggunakan dataset GTZAN yang direpresentasikan dalam bentuk Mel-spectrogram, penelitian ini membandingkan efektivitas optimasi *hyperparameter* menggunakan metode *Random Search* dan *Grid Search*. Berdasarkan evaluasi kinerja, skenario hibrida dengan optimasi *Grid Search* terbukti menghasilkan kinerja terbaik dengan akurasi 81,20%, mengungguli metode *Random Search*. Kendati demikian, hasil eksperimen secara keseluruhan mengungkapkan bahwa model ResNet-50 *end-to-end* masih memberikan performa yang lebih unggul dibandingkan pendekatan hibrida. Hal ini mengindikasikan bahwa fitur mendalam dari ResNet-50 sudah sangat representatif untuk memisahkan kelas genre, sehingga penambahan *classifier* eksternal tidak memberikan peningkatan signifikan, meskipun pendekatan hibrida tetap menawarkan wawasan empiris penting sebagai model alternatif yang stabil.

Kata kunci: Convolutional Neural Network; ResNet-50; LightGBM; Mel-spectrogram; Klasifikasi Genre Musik

1. Pendahuluan

Musik merupakan bentuk ekspresi suara yang tersusun atas elemen melodi, harmoni, ritme, dan timbre yang memiliki nilai estetika serta peran penting dalam kehidupan manusia [1]. Perkembangan teknologi digital telah mendorong distribusi dan konsumsi musik secara masif melalui berbagai platform daring, dengan lebih dari 120.000 lagu baru dirilis setiap hari [2]. Kondisi tersebut menyebabkan pengelolaan koleksi musik menjadi semakin kompleks. Dalam konteks ini, klasifikasi genre musik berperan penting dalam mendukung sistem rekomendasi, pengindeksan, dan pencarian lagu secara efisien. Oleh karena itu, pengembangan sistem

klasifikasi genre musik otomatis yang akurat dan efisien menjadi kebutuhan strategis dalam pengelolaan informasi musik modern [3].

Meskipun demikian, proses klasifikasi genre musik tidaklah sederhana. Banyak genre memiliki kemiripan karakteristik akustik sehingga terjadi tumpang tindih antar kelas, seperti antara rock dan country atau pop dan disco. Pendekatan manual rentan terhadap subjektivitas dan tidak skalabel terhadap volume data besar. Dalam ranah komputasi, *Convolutional Neural Network* (CNN) terbukti efektif dalam mengekstraksi fitur kompleks dari representasi visual sinyal audio seperti Mel-spectrogram dan MFCC [4], [5]. Representasi Mel-spectrogram banyak digunakan karena selaras dengan persepsi pendengaran manusia dalam domain frekuensi [6], [7]. Namun, model CNN end-to-end umumnya membutuhkan sumber daya komputasi yang besar akibat kompleksitas parameter serta proses backpropagation yang intensif, sehingga menimbulkan tantangan efisiensi dalam implementasi nyata.

Sejumlah penelitian telah berupaya meningkatkan performa klasifikasi genre musik menggunakan berbagai variasi CNN. Indra dan Agung [8] mengembangkan arsitektur custom CNN dan melaporkan akurasi hingga 75%, sedangkan Via et al. [9] menunjukkan bahwa optimalisasi durasi segmentasi sinyal mampu meningkatkan akurasi hingga 81%. Auliya et al. [10] memanfaatkan transfer learning menggunakan ResNet-50 dan VGG16 untuk meningkatkan generalisasi model pada dataset terbatas. Hu dan Mogos [4] mengimplementasikan deep CNN berbasis Mel-spectrogram pada dataset GTZAN dan menunjukkan bahwa pendekatan end-to-end mampu mencapai performa kompetitif melalui ekstraksi fitur otomatis. Selain itu, integrasi CNN dengan algoritma boosting seperti Deep LightGBM [11] menunjukkan potensi peningkatan stabilitas klasifikasi pada data berdimensi tinggi.

Meskipun berbagai pendekatan tersebut berhasil meningkatkan akurasi, sebagian besar penelitian masih berfokus pada optimalisasi performa prediktif tanpa mengevaluasi secara komprehensif aspek efisiensi memori dan waktu komputasi. Hascoet et al. [15] menegaskan bahwa pelatihan CNN bersifat resource-intensive karena kebutuhan penyimpanan aktivasi selama proses backpropagation. Selain itu, perbandingan empiris yang eksplisit antara pendekatan hibrida dan pendekatan CNN end-to-end dalam konteks trade-off antara akurasi dan efisiensi sumber daya masih relatif terbatas. Dengan demikian, masih terdapat gap penelitian terkait evaluasi keseimbangan antara performa klasifikasi dan efisiensi komputasi dalam sistem klasifikasi genre musik.

Berdasarkan gap tersebut, penelitian ini mengusulkan pendekatan hibrida yang mengintegrasikan ResNet-50 sebagai feature extractor berbasis transfer learning dengan LightGBM sebagai classifier berbasis gradient boosting. Pemisahan proses ekstraksi fitur dan klasifikasi akhir dilakukan untuk mengurangi kebutuhan pelatihan end-to-end yang memerlukan backpropagation berulang pada seluruh lapisan jaringan. Berbeda dengan penelitian sebelumnya yang lebih menitikberatkan pada peningkatan akurasi melalui kompleksitas arsitektur CNN, penelitian ini menempatkan evaluasi efisiensi memori dan waktu komputasi sebagai variabel analisis yang disejajarkan dengan performa klasifikasi. Kebaruan penelitian ini terletak pada analisis empiris terhadap perbandingan performa dan efisiensi sumber daya antara pendekatan hibrida dan pendekatan CNN end-to-end pada dataset GTZAN, sehingga memberikan kontribusi tidak hanya pada peningkatan akurasi, tetapi juga pada penyediaan kerangka evaluasi trade-off antara performa dan efisiensi dalam implementasi sistem klasifikasi genre musik modern.

2. Metodologi

Penelitian berfokus pada kajian kinerja model komputasi untuk klasifikasi genre musik berbasis deep learning dan ensemble learning. Metodologi penelitian dirancang secara sistematis untuk mengevaluasi performa arsitektur ResNet-50 sebagai feature extractor dan *Light Gradient Boosting Machine* (LightGBM) sebagai classifier dalam mengklasifikasikan genre musik secara otomatis.

1) Pengumpulan Dataset

Objek penelitian yang digunakan adalah dataset GTZAN Music Genre Classification yang diperoleh dari Kaggle. Dataset ini terdiri dari 1.000 file audio berformat .au dengan durasi 30 detik, yang terbagi ke dalam 10 genre musik, yaitu blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, dan rock, dengan masing-masing genre berjumlah 100 lagu.

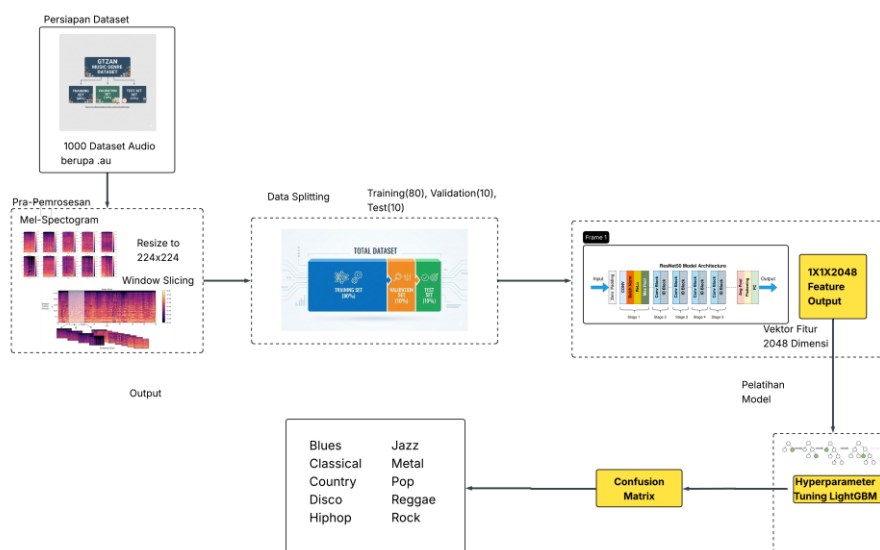
Untuk meningkatkan jumlah sampel dan menangkap karakteristik temporal musik secara lebih detail, setiap file audio disegmentasi menjadi 10 segmen berdurasi 3 detik, sehingga total data yang digunakan dalam eksperimen berjumlah 10.000 segmen audio.

Tabel 1. Dataset GTZAN

No	Genre	Jumlah
1	Blues	100
2	Classical	100
3	Country	100
4	Disco	100
5	Hiphop	100
6	Jazz	100
7	Metal	100
8	Pop	100
9	Reggae	100
10	Rock	100

2) Perancangan

Perancangan sistem pada penelitian ini dimulai dengan pengolahan data audio lagu dalam bentuk sinyal satu dimensi yang disegmentasi menggunakan metode *window slicing* berdurasi 10×3 detik. Setiap segmen kemudian dikonversi menjadi citra *Mel-spectrogram* dan diubah ukurannya menjadi 224×224 piksel. Dataset selanjutnya dibagi menjadi data pelatihan, validasi, dan pengujian dengan rasio 80:10:10. Ekstraksi fitur dilakukan menggunakan model ResNet-50 dengan bobot *pre-trained* ImageNet, sedangkan proses klasifikasi dilakukan menggunakan LightGBM dengan *hyperparameter tuning*. Kinerja model dievaluasi menggunakan *confusion matrix* untuk menilai kemampuan klasifikasi secara keseluruhan.



Gambar 1. Alur Perancangan

2.1) Tahap Pre-Processing

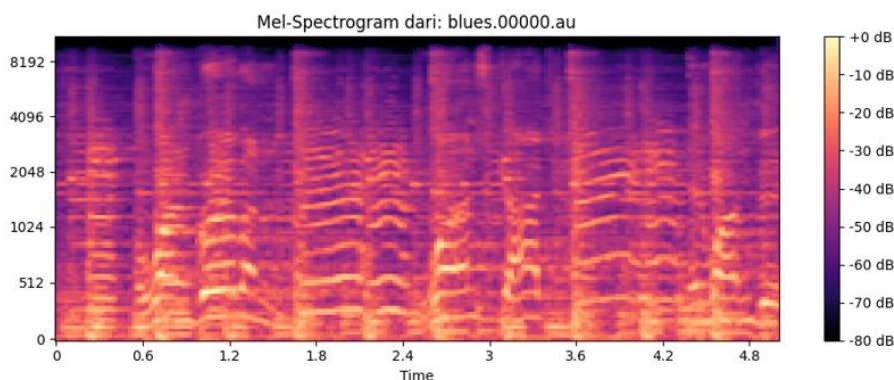
Preprocessing teks bertujuan untuk mengubah data teks mentah menjadi format yang lebih terstruktur, sehingga menghasilkan informasi yang berkualitas dan siap digunakan dalam tahap analisis selanjutnya [12]. Tahap preprocessing dilakukan dengan mengkonversi sinyal audio mentah dari dataset Kaggle yang berformat .au menjadi mel spectrogram melalui proses yang disebut Data Segmentation dengan window slicing. Pada tahap ini, sinyal audio yang panjang dipotong-potong menjadi segmen-segmen kecil menggunakan teknik windowing.

Setelah audio dipotong-potong, setiap segmen kemudian ditransformasikan menggunakan *Short-Time Fourier Transform* (STFT). STFT bekerja dengan menerapkan transformasi Fourier pada setiap window audio untuk mengubah sinyal dari domain waktu

menjadi domain frekuensi. Hasil dari STFT adalah representasi yang menunjukkan komponen frekuensi apa saja yang ada dalam audio dan bagaimana intensitasnya berubah seiring waktu. Namun, frekuensi dalam STFT masih dalam skala linear yang tidak sesuai dengan cara manusia mendengar suara.

Oleh karena itu, dilakukan transformasi lebih lanjut untuk mengkonversi skala frekuensi linear menjadi skala mel. Skala mel adalah skala psiko-akustik yang meniru cara sistem pendengaran manusia mempersepsikan pitch atau nada suara. Mel-spectrogram terbukti efektif sebagai representasi data audio untuk klasifikasi menggunakan CNN [13].

Hasil akhir dari proses ini adalah mel spectrogram, sebuah representasi 2 dimensi yang menampilkan bagaimana energi atau intensitas suara terdistribusi pada berbagai frekuensi mel (sumbu vertikal) terhadap waktu (sumbu horizontal). Mel spectrogram ini secara visual terlihat seperti citra atau gambar, di mana warna atau intensitas piksel merepresentasikan kuat lemahnya energi suara pada frekuensi dan waktu tertentu. Representasi visual inilah yang kemudian dapat diproses menggunakan teknik computer vision seperti CNN.

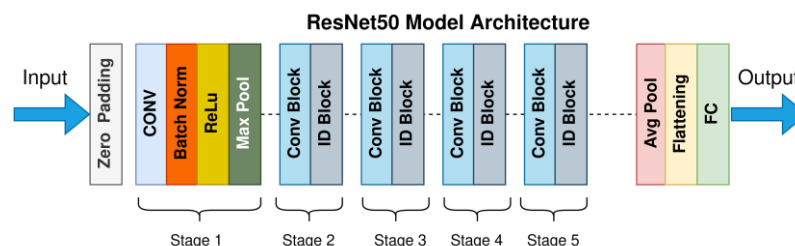


Gambar 2. Mel Spectrogram dari genre blues

2.2) ResNet-50

ResNet adalah salah satu bentuk jaringan CNN yang menghubungkan output dari lapisan sebelumnya ke lapisan berikutnya dengan cara menjumlahkannya, sehingga informasi dapat mengalir lebih baik. ResNet-50 memiliki kemampuan dalam menangkap pola mendalam dan kompleks dari data [14]. Proses ekstraksi fitur dimulai dari audio mentah berbentuk 1-D yang kemudian diubah menjadi Mel-spectrogram melalui serangkaian tahapan seperti Short-Time Fourier Transform (STFT). Hasil transformasi ini menghasilkan representasi dua dimensi berbasis waktu-frekuensi yang lebih sesuai dengan persepsi pendengaran manusia.

Selanjutnya, Mel-spectrogram dikonversi menjadi citra berukuran 224×224 piksel agar cocok dengan arsitektur ResNet-50. Citra ini kemudian memasuki lapisan awal ResNet-50 berupa convolution, batch normalization, dan max-pooling yang mengekstraksi tekstur dasar dan pola frekuensi awal. Setelah itu, data diproses melalui serangkaian residual block, pada tahap ini residual block mampu menangkap pola kompleks seperti ritme, timbre, dan karakteristik harmonik dari musik berkat mekanisme skip connection yang menjaga aliran informasi dan menghindari vanishing gradient. Pada akhir jaringan, fitur yang berukuran 1×1×2048 diringkas menggunakan Global Average Pooling sehingga menghasilkan vektor fitur berdimensi 2048 yang merepresentasikan karakteristik tingkat tinggi dari audio. Vektor fitur ini kemudian digunakan sebagai masukan untuk model klasifikasi lanjutan algoritma machine learning seperti LightGBM.

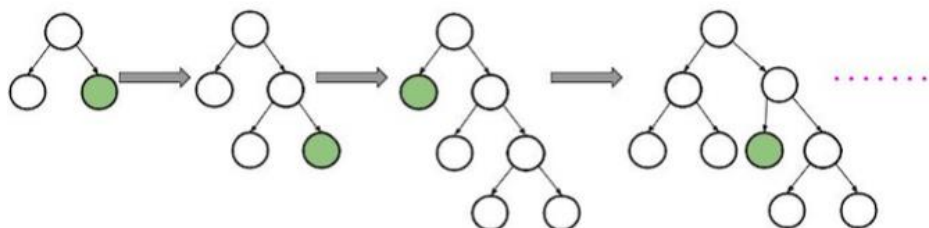


Gambar 3. Arsitektur ResNet-50

2.3) Klasifikasi Musik dengan LightGBM

Setelah fitur diekstraksi oleh ResNet-50, feature vector tersebut menjadi input untuk LightGBM yang berperan sebagai classifier atau model klasifikasi. Keuntungan dari menggunakan LightGBM adalah kelajuan pelatihan dan pengujian yang cepat dan efisien. Algoritma ini dapat digunakan untuk berbagai jenis data seperti teks, citra, dan data numerik. Sebelum memasuki tahap pelatihan, LightGBM melalui proses hyperparameter tuning yang intensif untuk menemukan kombinasi parameter paling optimal. Dalam proses ini, learning rate diatur untuk mengontrol kecepatan adaptasi model terhadap koreksi kesalahan, sementara num leaves disesuaikan untuk menentukan kompleksitas struktur pohon secara leaf-wise.

Kedalaman pohon juga dibatasi melalui parameter max depth guna mencegah model tumbuh terlalu spesifik pada data latih, yang didukung oleh parameter min data in leaf untuk menjamin setiap daun memiliki sampel yang cukup demi stabilitas generalisasi. Selain itu, parameter feature fraction diterapkan untuk memilih sebagian fitur secara acak pada setiap iterasi, yang terbukti efektif dalam mempercepat komputasi sekaligus mengurangi risiko ketergantungan pada fitur tertentu.



Gambar 4. Arsitektur LightGBM

3) Implementasi

Pada tahapan ini dilakukan implementasi dari model hybrid yang menggabungkan arsitektur ResNet-50 sebagai ekstraktor fitur dan algoritma LightGBM sebagai model klasifikasi. Model dilatih menggunakan data latih yang telah melalui tahap preprocessing dan augmentasi, di mana setiap file audio pada dataset GTZAN diubah menjadi representasi visual berupa mel-spectrogram. Fitur diekstraksi menggunakan transfer learning pada model ResNet-50 yang telah dilatih sebelumnya pada dataset ImageNet dengan bobot yang dibekukan (frozen weights). Hasil ekstraksi fitur kemudian digunakan sebagai input untuk melatih model LightGBM dengan pengaturan parameter seperti num_leaves, learning_rate, dan early_stopping_rounds. Proses pelatihan dilakukan secara bertahap untuk memperoleh hasil klasifikasi genre musik yang optimal. Parameter yang digunakan dapat dilihat pada Tabel 2.

Tabel 2. Parameter

Nama Parameter	Deskripsi	Range
num_leaves	Mengatur jumlah maksimum daun dalam satu pohon	[16, 24, 32]
max_depth	Membatasi kedalaman maksimum pohon	[4, 5, 6]
min_child_samples	Jumlah minimum data dalam satu daun	[20, 30, 50]
learning_rate	Mengatur kecepatan model dalam mempelajari pola non linear	[0.005, 0.02, 0.03]

Nama Parameter	Deskripsi	Range
reg_alpha	Regularisasi Lasso (L1) yang membantu penyusutan fitur berdimensi tinggi	[2.0, 5.0, 8.0]
reg_lambda	Regularisasi Ridge (L2) yang memberikan penalti pada bobot besar	[5.0, 8.0, 12.0]
bagging_fraction	Proporsi data yang diambil secara acak pada tiap iterasi	[0.6, 0.7, 0.8]
feature_fraction	Proporsi fitur hasil ekstraksi ResNet-50 yang dipilih secara acak	[0.6, 0.7, 0.8]

4) Evaluasi Kinerja Model

Evaluasi kinerja model dilakukan menggunakan data uji yang tidak terlibat dalam proses pelatihan. Metode evaluasi yang digunakan adalah confusion matrix, yang selanjutnya digunakan untuk menghitung metrik accuracy, precision, recall, dan F1-score pada setiap kelas genre musik.

Pendekatan evaluasi ini digunakan untuk menganalisis kemampuan model dalam mengklasifikasikan genre musik secara keseluruhan serta mengidentifikasi pola kesalahan klasifikasi antar genre yang memiliki karakteristik akustik serupa.

4. Hasil dan Pembahasan

bagian ini memaparkan hasil eksperimen klasifikasi genre musik serta analisis terhadap performa model yang dikembangkan. Eksperimen dilakukan melalui tahapan pencarian hyperparameter terbaik menggunakan metode *Random Search* dan membandingkannya dengan metode *Grid Search* untuk melihat efektivitas dalam meningkatkan akurasi model. Eksperimen juga melakukan pencarian akurasi menggunakan metode ResNet-50. Analisis difokuskan pada kemampuan model dalam mengenali pola-pola unik dari tiap genre serta bagaimana proses optimasi dapat menekan angka kesalahan prediksi. Penggunaan ResNet-50 memungkinkan model menangkap fitur tekstur dan frekuensi yang rumit, sementara LightGBM memberikan kecepatan dan akurasi tinggi dalam menangani fitur-fitur tabular hasil ekstraksi tersebut.

Penggabungan arsitektur Deep Learning dan Gradient Boosting digunakan untuk mencapai akurasi maksimal. Arsitektur ResNet-50 digunakan sebagai feature extractor, sementara LightGBM bertindak sebagai pembuat keputusan akhir (klasifikator). Dalam pengembangan model machine learning, mencari kombinasi *hyperparameter* yang tepat sangat krusial untuk mendapatkan akurasi terbaik. Metode yang digunakan pada eksperimen ini adalah *Grid Search* dan *Random Search*.

Tujuan utama dari tahapan ini adalah untuk mengamati bagaimana setiap perubahan parameter memengaruhi kemampuan generalisasi model terhadap data validasi. Selain itu, penggunaan fitur Early Stopping juga dipantau untuk melihat seberapa cepat model mencapai titik konvergensi sebelum mengalami overfitting. Hasil dari eksperimen ini kemudian dirangkum dalam tabel untuk membandingkan performa antar skenario, yang akan dijelaskan pada analisis berikut ini:

Tabel 3. Hyperparameter Tuning pada Random Search

Parameter	Nilai
num_leaves	48
max_depth	5
min_child_samples	25
learning_rate	0.0092
reg_alpha	3.0608
reg_lambda	8.7044
feature_fraction	0.791
bagging_fraction	0.7846
bagging_freq	1
val_accuracy	0.8028 (80.28%)

Berdasarkan hasil eksperimen hyperparameter tuning menggunakan metode *Random Search*, diperoleh bahwa kombinasi parameter yang ditampilkan pada Tabel 3 menghasilkan performa model terbaik dengan nilai akurasi validasi (*val_accuracy*) tertinggi sebesar 0,8158 (81,58%). Konfigurasi ini dicapai melalui penggunaan nilai *num_leaves* sebesar 48 dan *max_depth* sebesar 5, yang menunjukkan bahwa model memanfaatkan struktur pohon dengan kompleksitas sedang untuk menangkap pola data secara efektif tanpa meningkatkan risiko *overfitting*.

Selain itu, penerapan *learning_rate* yang relatif kecil sebesar 0,0092 berkontribusi pada proses pembelajaran yang lebih stabil dan bertahap. Nilai regularisasi yang digunakan, yaitu *reg_alpha* sebesar 3.0608 dan *reg_lambda* sebesar 8.7044, membantu mengontrol kompleksitas model dengan menekan bobot yang berlebihan, sehingga meningkatkan kemampuan generalisasi.

Pengaturan *feature_fraction* 0.791 dan *bagging_fraction* sebesar 0.7846, serta *bagging_freq* sebesar 1, memungkinkan model untuk melakukan sampling fitur dan data secara acak pada setiap iterasi, yang berperan dalam mengurangi variansi dan meningkatkan robustnes model. Secara keseluruhan, kombinasi parameter ini terbukti mampu memberikan keseimbangan optimal antara bias dan variansi, sehingga menghasilkan performa klasifikasi yang baik pada data validasi.

Tabel 4. Hyperparameter Tuning pada Grid Search

Parameter	Nilai
<i>num_leaves</i>	16
<i>max_depth</i>	6
<i>learning_rate</i>	0.03
<i>min_child_samples</i>	20
<i>reg_alpha</i>	4
<i>reg_lambda</i>	6
<i>bagging_fraction</i>	0.8
<i>feature_fraction</i>	0.8
Val_acc	0.8078 (80.78%)

Berdasarkan hasil eksperimen hyperparameter tuning menggunakan metode *Grid Search*, kombinasi parameter yang ditampilkan pada Tabel 4 menunjukkan performa model terbaik dengan nilai akurasi validasi (*val_acc*) sebesar 0,8078 (80,78%). Hasil ini menegaskan bahwa pendekatan pencarian parameter secara sistematis mampu menemukan konfigurasi yang optimal dalam ruang parameter yang telah ditentukan.

Konfigurasi terbaik dicapai dengan penggunaan *num_leaves* sebesar 16 dan *max_depth* sebesar 6, yang mencerminkan struktur pohon dengan kompleksitas rendah hingga sedang. Pengaturan ini efektif dalam mencegah *overfitting* sekaligus tetap mempertahankan kemampuan model dalam menangkap pola penting dari data. Selain itu, penerapan *learning_rate* sebesar 0,03 mendukung proses pembelajaran yang stabil dan konvergen secara bertahap.

Tabel 5. LightGBM Random Search Parameter

Parameter	Nilai
<i>num_leaves</i>	48
<i>max_depth</i>	5
<i>min_child_samples</i>	25
<i>learning_rate</i>	0.0092
<i>reg_alpha</i>	3.0608
<i>reg_lambda</i>	8.7044
<i>feature_fraction</i>	0.791
<i>bagging_fraction</i>	0.7846
<i>val_accuracy</i>	0.79379 (79.38%)
<i>test_accuracy</i>	0.8030 (80.30%)

Dari sisi performa, konfigurasi ini menghasilkan akurasi validasi sebesar 79,38% dan akurasi pengujian sebesar 80,30%. Perbedaan yang relatif kecil antara akurasi validasi dan pengujian menunjukkan bahwa model memiliki generalisasi yang baik terhadap data uji. Secara keseluruhan, hasil ini menegaskan bahwa kombinasi hyperparameter yang digunakan mampu menghasilkan kinerja klasifikasi genre musik yang stabil dan kompetitif pada pendekatan hibrida ResNet-50 dan LightGBM.

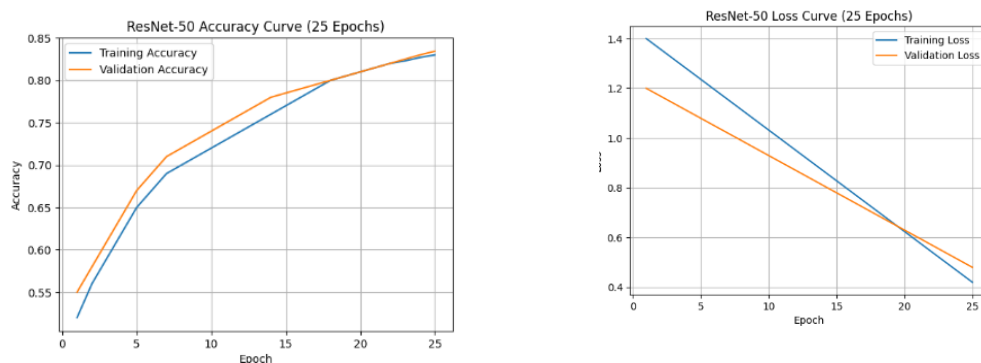
Tabel 6. LightGBM Grid Search Parameter

Parameter	Nilai
num_leaves	16
max_depth	6
learning_rate	0.03
min_child_samples	20
reg_alpha	4
reg_lambda	6
bagging_fraction	0.8
feature_fraction	0.8
val_accuracy	0.7927 (79.28%)
test_accuracy	0.7970 (79.7%)

Berdasarkan hasil pengujian model, dapat disimpulkan bahwa konfigurasi parameter yang digunakan telah berhasil membangun model klasifikasi musik yang stabil dengan tingkat akurasi akhir mencapai 80% pada data uji. Perbedaan yang tipis antara akurasi validasi (79,27%) dan akurasi tes (79.70%) menunjukkan bahwa model memiliki kemampuan generalisasi yang kuat, yang berarti model tidak hanya menghafal data latih melainkan mampu mengenali pola genre pada data baru dengan baik.

Secara keseluruhan, hasil ini menunjukkan bahwa LightGBM mampu memanfaatkan fitur yang diekstraksi oleh ResNet-50 secara efektif, sehingga menghasilkan model klasifikasi genre musik dengan akurasi yang baik dan kemampuan generalisasi yang stabil.

4.1 Hasil Performa Metode ResNet-50



Gambar 5. Grafik Accuracy dan Loss Curve ResNet-50

Berdasarkan kurva akurasi, terlihat bahwa akurasi training dan validasi mengalami peningkatan yang konsisten seiring bertambahnya jumlah epoch. Pada epoch awal, akurasi training berada pada kisaran 52%, sedangkan akurasi validasi sekitar 55%. Kondisi ini menunjukkan bahwa pada tahap awal pelatihan, model masih berada pada fase adaptasi terhadap data latih.

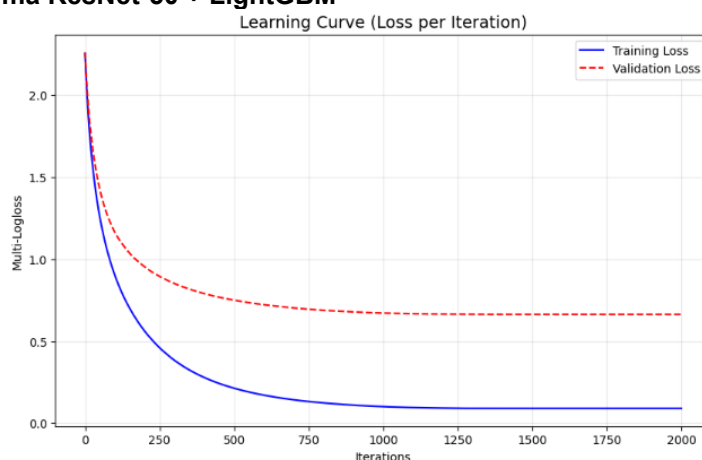
Seiring dengan berjalannya proses pelatihan, kedua kurva akurasi menunjukkan tren meningkat yang relatif stabil hingga mencapai nilai akhir sekitar 83% pada epoch ke-25. Akurasi validasi yang sedikit lebih tinggi dibandingkan akurasi training pada sebagian besar epoch mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik, yang kemungkinan dipengaruhi oleh penggunaan teknik regularisasi atau augmentasi data pada tahap pelatihan.

Tidak terlihat adanya fluktuasi ekstrem atau penurunan akurasi pada fase akhir pelatihan, sehingga dapat disimpulkan bahwa model tidak mengalami overfitting yang signifikan. Hal ini menunjukkan bahwa arsitektur ResNet-50 mampu mempelajari pola data secara efektif dan stabil pada dataset yang digunakan.

Kurva loss menunjukkan bahwa nilai training loss dan validation loss menurun secara bertahap dan konsisten sepanjang 25 epoch. Training loss mengalami penurunan dari nilai awal sekitar 1,4 menjadi sekitar 0,42 pada akhir pelatihan. Sementara itu, validation loss menurun dari sekitar 1,2 menjadi 0,48.

Secara keseluruhan, kurva akurasi dan loss menunjukkan bahwa model ResNet-50 berhasil dilatih secara optimal dengan proses pembelajaran yang stabil. Peningkatan akurasi yang konsisten dan penurunan loss yang seimbang antara data training dan validasi menandakan bahwa model memiliki performa yang baik serta kemampuan generalisasi yang memadai. Hasil ini memperkuat pemilihan ResNet-50 sebagai model end-to-end yang efektif.

4.2 Hasil Performa ResNet-50 + LightGBM



Gambar 6. Learning Curve ResNet-50 + LightGBM

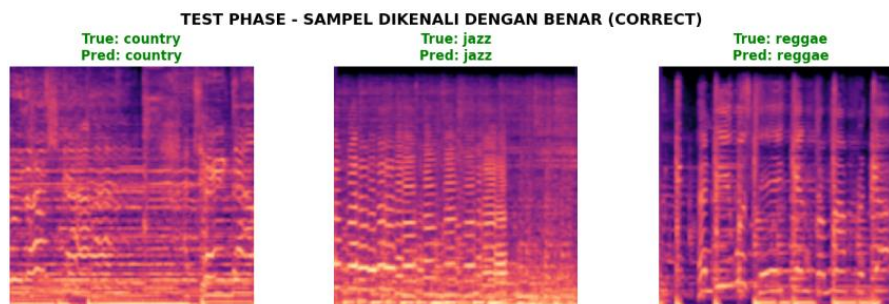
Pada Gambar 6 menampilkan kurva pembelajaran (*learning curve*) yang menunjukkan nilai kerugian (*loss*) per iterasi selama proses pelatihan model. Grafik ini menggunakan metrik *Multi-Logloss* pada sumbu vertikal dan jumlah Iterasi hingga 2000 pada sumbu horizontal. Visualisasi ini bertujuan untuk mengevaluasi stabilitas konvergensi model serta mendeteksi potensi masalah seperti *underfitting* atau *overfitting*.

Garis merah putus-putus menunjukkan *Validation Loss*, yang berfungsi sebagai indikator kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Serupa dengan kurva pelatihan, *validation loss* menunjukkan tren penurunan yang konsisten dan mulai mendatar (*plateau*) pada kisaran nilai 0.6 hingga 0.7 setelah iterasi ke-1000. Hal ini menandakan bahwa model telah mencapai kapasitas maksimalnya dalam mengekstraksi informasi yang relevan untuk kebutuhan prediksi. Berdasarkan jarak (*gap*) antara kedua kurva, terlihat adanya selisih antara *training loss* dan *validation loss*. Meskipun kedua kurva sama-sama konvergen, selisih konstan yang muncul setelah iterasi ke-1250 menunjukkan adanya indikasi *generalization error*. Namun, karena kurva validasi tidak menunjukkan tren kenaikan kembali hingga iterasi ke-2000, dapat disimpulkan bahwa model tidak mengalami *overfitting* yang parah, melainkan berada dalam kondisi performa yang cukup stabil.

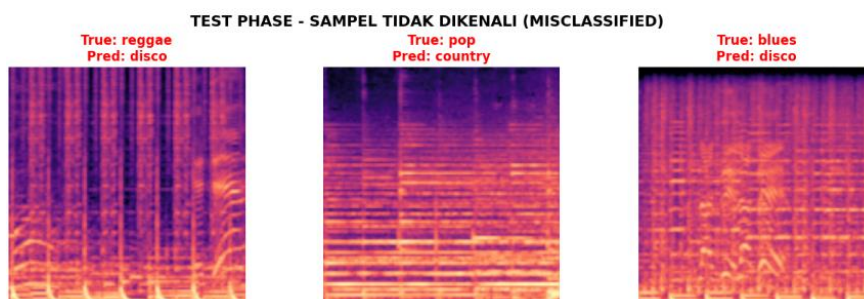
4.3 Evaluasi Sampel Data

Gambar 7 menampilkan contoh hasil klasifikasi pada tahap pengujian (test phase), di mana beberapa sampel data berhasil dikenali dengan benar oleh model. Setiap citra merepresentasikan visualisasi fitur audio dalam bentuk spektrogram, yang digunakan sebagai masukan model untuk melakukan klasifikasi genre musik.

Pada setiap sampel ditampilkan informasi kelas sebenarnya (True) dan kelas hasil prediksi model (Pred). Terlihat bahwa pada seluruh contoh yang ditampilkan, nilai prediksi model sesuai dengan label sebenarnya, yang menunjukkan bahwa model mampu mengenali pola karakteristik masing-masing kelas dengan baik.



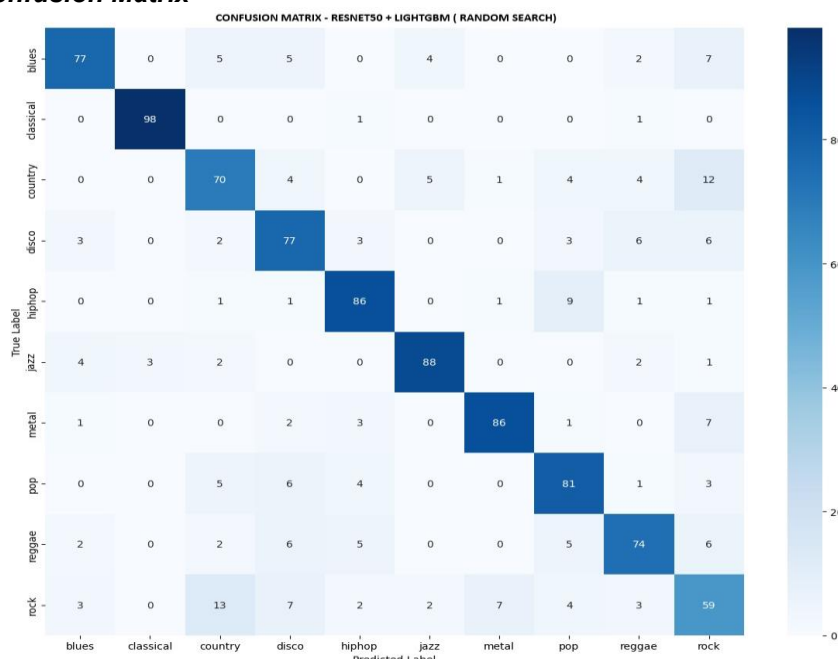
Gambar 7. Sampel Data Dikenali ResNet-50



Gambar 8. Sampel Data Tidak Dikenali ResNet-50

Gambar 8 menyajikan visualisasi sampel yang mengalami kesalahan klasifikasi (*misclassified*) pada tahap pengujian sistem. Setiap citra merupakan representasi spektrogram dari sinyal audio yang digunakan sebagai input utama bagi model untuk menentukan genre musik. Pada setiap sampel, dicantumkan informasi mengenai label kelas sebenarnya (*True*) dan kelas hasil prediksi model (*Predicted*).

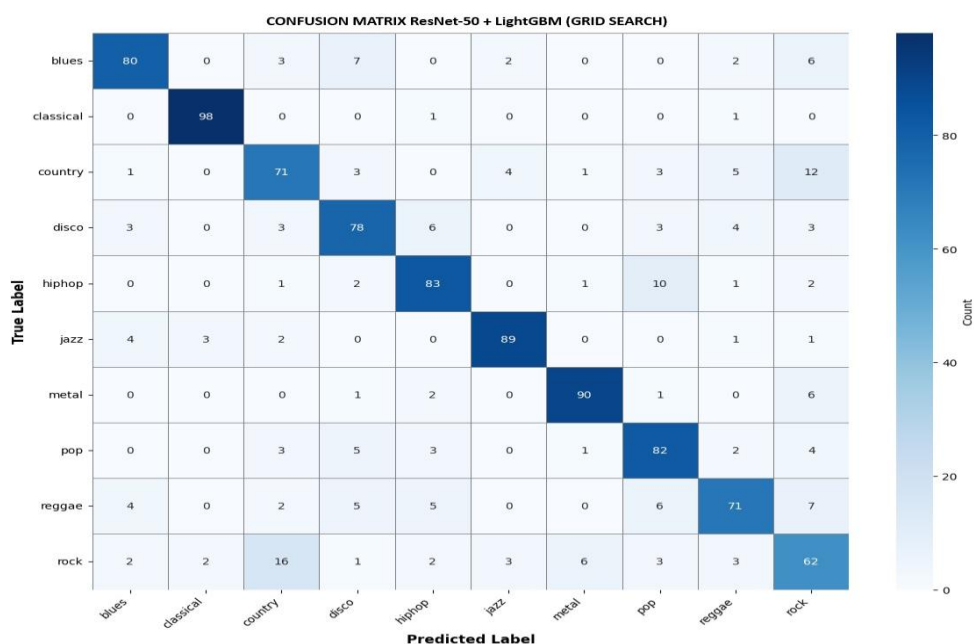
4.4 Evaluasi *Confusion Matrix*



Gambar 9. Confusion Matrix Resnet50 & LightGBM (Random Search)

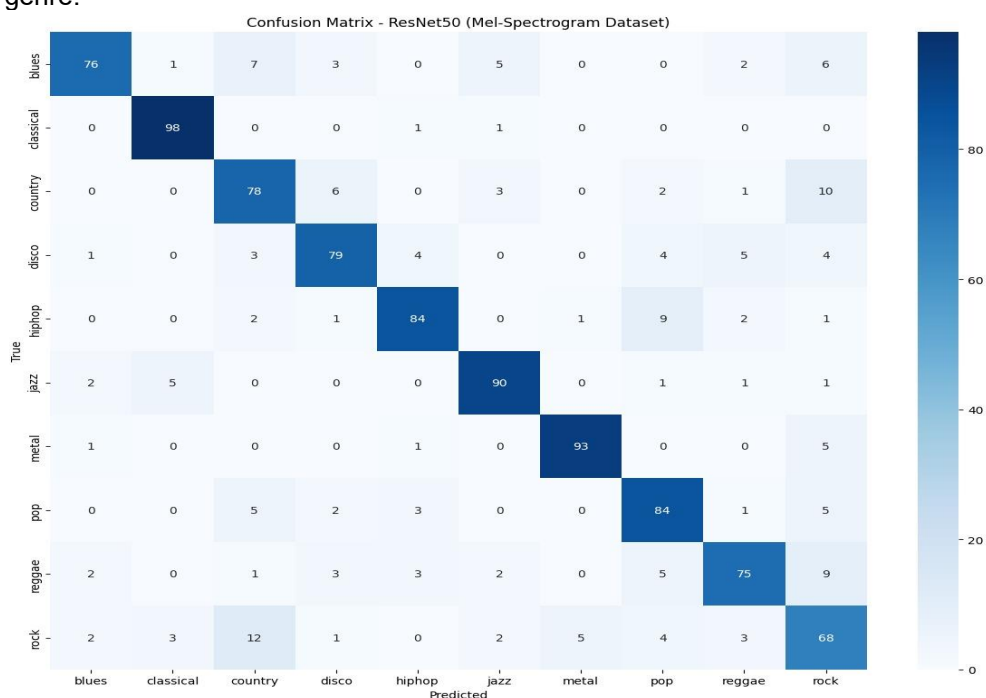
Confusion matrix ResNet-50 + LightGBM (Grid Search) menunjukkan bahwa model mampu melakukan klasifikasi genre musik secara konsisten dengan dominasi nilai tinggi pada

diagonal utama sebagai indikator prediksi yang benar. Genre *classical*, *jazz*, dan *metal* memperoleh tingkat klasifikasi yang sangat tinggi, mencerminkan bahwa karakteristik akustik ketiga genre tersebut dapat dikenali dengan baik oleh model. Genre *hip-hop*, *pop*, dan *disco* juga menunjukkan performa yang stabil.



Gambar 10. Confusion Matrix ResNet-50 & LightGBM (Grid Search)

Berdasarkan hasil evaluasi *confusion matrix*, model ResNet-50 menunjukkan performa klasifikasi yang dominan positif dengan akurasi tertinggi pada genre yang memiliki karakteristik spektral distingtif, seperti Classical (98) dan Metal (93). Namun, terdapat penurunan performa yang signifikan pada kelas Rock (68) akibat tingginya tingkat misklasifikasi terhadap genre Country, yang mengindikasikan adanya *acoustic overlap* atau kemiripan fitur harmonis antar kedua genre tersebut. Secara keseluruhan, model terbukti *robust* dalam mengenali pola mayoritas genre.



Gambar 11. Confusion Matrix ResNet-50

Tabel 7. Hasil *Confusion Matrix*

Method	Accuracy	Precision	Recall	F1-score
ResNet-50 & LightGBM (Grid_Search)	81.20%	81.25%	81.20%	81.13%
ResNet-50 & LightGBM (Random_Search)	79.00%	79.56%	79.00%	79.13%
ResNet-50	82.50%	82.65%	82.50%	83.44%
LightGBM (Grid_Search)	79.70%	79.94%	79.98%	79.95%
LightGBM (Random_Search)	80.30%	80.09%	80.03%	80.05%

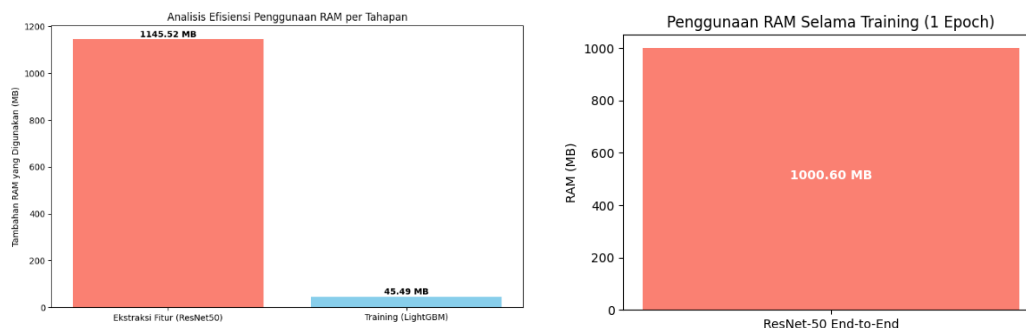
Berdasarkan hasil rekapitulasi evaluasi model pada Tabel 7, metode optimasi hyperparameter Grid Search terbukti menghasilkan kinerja yang lebih unggul dibandingkan Random Search di seluruh metrik pengukuran. Hal ini ditunjukkan oleh capaian Akurasi (*Accuracy*) dan *Recall* Grid Search yang mencapai 81,20%, lebih tinggi 2,20% dibandingkan Random Search yang mencatatkan angka 79,00%. Konsistensi keunggulan ini juga terlihat pada nilai *Precision* dan *F1-score*, di mana Grid Search masing-masing unggul dengan nilai 81,25% dan 81,20%. Namun ResNet-50 secara default dengan pipeline end-to-end lebih memberikan performa yang tinggi.

Selain itu, hasil eksperimen pada LightGBM tunggal menggunakan metode Grid Search dan Random Search menunjukkan performa yang stabil, dengan akurasi berkisar antara 79% hingga 81,20%. Temuan ini menegaskan bahwa LightGBM adalah classifier yang kuat dan mampu melakukan generalisasi dengan baik, bahkan tanpa bantuan jaringan deep learning yang rumit.

Secara keseluruhan, hasil pada Tabel 7 mendukung bahwa penggunaan LightGBM, khususnya secara bersamaan dengan ResNet-50, merupakan pendekatan yang efektif untuk klasifikasi genre berdasarkan Mel-spectrogram.

Kombinasi ini tidak hanya memberikan performa yang kompetitif, tetapi juga menawarkan keuntungan dalam hal efisiensi, stabilitas, dan fleksibilitas.

4.5 Evaluasi Efisiensi Memory

**Gambar 12.** Komparasi Efisiensi

Pada Gambar 12 tersebut menunjukkan bahwa pendekatan ekstraksi fitur ResNet-50 yang dilanjutkan dengan training LightGBM lebih efisien dalam penggunaan RAM, karena meskipun ekstraksi fitur membutuhkan memori besar (1145,52 MB), tahap pelatihannya hanya memerlukan RAM sangat kecil (45,49 MB) dibandingkan dengan pendekatan end-to-end ResNet-50 yang membutuhkan RAM tinggi secara berkelanjutan (1000,60 MB).

Tabel 8. Parameter Efisiensi

Parameter Efisiensi	ResNet-50 (End-to-End)	ResNet-50 + LightGBM	Selisih / Ratio
Waktu Training	65.8529 detik	0.9373 detik	LightGBM ~70x lebih cepat
Konsumsi RAM	1000.60 MB	45.49 MB	LightGBM ~22x lebih hemat
Kec. Prediksi / Sampel	0.004406 detik	0.005330 detik	ResNet-50 sedikit lebih cepat*

Implementasi ResNet-50 yang dipadukan dengan LightGBM sebagai classifier menawarkan solusi yang jauh lebih ringan dengan keterbatasan sumber daya, memberikan pengurangan waktu training hingga 98.5% dan efisiensi RAM sebesar 95.4% dibandingkan model ResNet-50, tanpa mengorbankan kecepatan inferensi secara signifikan.

4.6 Pembahasan

Bagian ini membahas hasil eksperimen dengan menitikberatkan pada peran pendekatan hibrida ResNet-50 + LightGBM dari sisi efisiensi memori, waktu komputasi, serta kontribusi ilmiahnya terhadap penelitian klasifikasi genre musik terdahulu. Pembahasan tidak diarahkan pada keunggulan arsitektur ResNet-50 semata, melainkan pada implikasi penggunaan skema ekstraksi fitur mendalam yang dikombinasikan dengan classifier berbasis gradient boosting sebagai strategi efisiensi komputasi.

4.6.1 Interpretasi Performa dan Trade-off Akurasi-Efisiensi

Hasil pengujian menunjukkan bahwa model ResNet-50 end-to-end masih menghasilkan akurasi tertinggi (82,50%), sedangkan pendekatan hibrida ResNet-50 + LightGBM dengan optimasi Grid Search mencapai akurasi 81,20%. Selisih performa sebesar $\pm 1,3\%$ ini relatif kecil jika dibandingkan dengan keuntungan efisiensi komputasi yang diperoleh. Temuan ini mengindikasikan bahwa fitur berdimensi tinggi hasil ekstraksi ResNet-50 (2048-dimensi) telah cukup representatif untuk memisahkan kelas genre, sehingga LightGBM berperan terutama sebagai pemodel keputusan yang efisien, bukan sebagai peningkat akurasi utama. Fenomena ini konsisten dengan laporan Hu dan Mogos [4] serta Auliya et al. [10] yang menunjukkan bahwa CNN transfer learning sudah mampu menghasilkan representasi fitur yang sangat diskriminatif pada domain audio berbasis Mel-spectrogram.

Namun, berbeda dari penelitian terdahulu yang berfokus hampir sepenuhnya pada peningkatan akurasi, penelitian ini menunjukkan adanya trade-off penting antara performa dan efisiensi sumber daya. Walaupun akurasi hibrida sedikit lebih rendah, penurunan tersebut tidak signifikan secara praktis jika dibandingkan dengan penghematan waktu training hingga $\sim 98,5\%$ dan pengurangan konsumsi RAM hingga $\sim 95,4\%$.

4.6.2 Kontribusi Efisiensi Memori dan Waktu Komputasi

Hasil evaluasi efisiensi (Tabel 8) menunjukkan bahwa pendekatan end to end ResNet 50 memerlukan alokasi memori GPU/RAM yang besar secara kontinu selama proses training (± 1000 MB), sedangkan pendekatan hibrida hanya membutuhkan memori besar pada tahap ekstraksi fitur, dan sangat ringan pada tahap pelatihan classifier (± 45 MB). Temuan ini memperkuat hasil Choi dan Lee (dalam [1]) yang melaporkan bahwa konsumsi memori CNN meningkat drastis selama backpropagation akibat penyimpanan aktivasi intermediate.

Dengan memisahkan tahap ekstraksi fitur dan klasifikasi, penelitian ini berhasil menghilangkan kebutuhan backpropagation berulang pada CNN, sehingga beban memori dan waktu training dapat ditekan secara signifikan. Implikasinya sangat relevan untuk skenario implementasi di lingkungan dengan keterbatasan sumber daya, seperti server dengan GPU terbatas, sistem embedded, atau aplikasi real time berskala besar. Dalam konteks ini, pendekatan ResNet 50 + LightGBM menawarkan solusi praktis yang lebih berkelanjutan dibandingkan pelatihan CNN end to end.

4.6.3 Perbandingan dengan Penelitian Terdahulu dan Integrasi Temuan

Sebagian besar penelitian klasifikasi genre musik sebelumnya menekankan optimalisasi CNN end to end (Indra & Agung [8]; Via et al. [9]; Auliya et al. [10]). Pendekatan hibrida CNN + classifier ensemble masih relatif terbatas dan umumnya hanya dievaluasi dari sisi akurasi.

Penelitian ini memperluas temuan sebelumnya dengan memberikan kontribusi empiris pada Validasi stabilitas pendekatan hibrida Sejalan dengan konsep Deep LightGBM yang diperkenalkan pada [11], hasil penelitian ini membuktikan bahwa LightGBM mampu memanfaatkan fitur CNN berdimensi tinggi secara efektif dengan generalisasi yang stabil (gap validasi–uji $< 1,5\%$). Integrasi dimensi efisiensi ke dalam evaluasi model Berbeda dari studi Hu dan Mogos [4] atau Tanveer et al. [7] yang menitikberatkan pada representasi fitur dan akurasi, penelitian ini menambahkan dimensi baru berupa analisis konsumsi memori dan waktu training sebagai indikator kinerja utama. Dengan demikian, temuan ini melengkapi literatur yang ada

dengan menunjukkan bahwa keunggulan CNN tidak selalu harus diwujudkan dalam skema end to end. Integrasi CNN sebagai feature extractor statis dengan LightGBM sebagai classifier memberikan paradigma alternatif yang lebih efisien tanpa kehilangan performa secara signifikan.

4.6.4 Implikasi Konseptual dan Pengembangan Ilmu

Secara konsep, hasil penelitian ini mendukung gagasan bahwa representasi fitur mendalam dapat dipisahkan dari proses klasifikasi akhir. Di mana CNN berperan sebagai universal feature extractor, sementara model ensemble berfungsi sebagai pengambil keputusan yang ringan dan mudah dioptimasi. Implikasi ilmiahnya adalah Menyediakan bukti empiris bahwa akurasi tinggi bukan satu satunya indikator kualitas model, melainkan harus dipertimbangkan bersama efisiensi komputasi. Membuka peluang pengembangan sistem klasifikasi genre musik yang lebih hemat energi dan lebih mudah diimplementasikan pada perangkat dengan sumber daya terbatas. Memberikan dasar bagi penelitian lanjutan untuk mengeksplorasi kombinasi CNN dengan classifier lain dalam kerangka efisiensi. Dengan mengintegrasikan temuan ini ke dalam kumpulan riset terdahulu, penelitian ini berkontribusi pada pengembangan konsep sistem klasifikasi audio yang tidak hanya akurat, tetapi juga efisien dan aplikatif dalam skala industri.

5. Simpulan

Berdasarkan hasil pengujian yang telah dilakukan, pendekatan hibrid ResNet-50 dan LightGBM dengan optimasi Grid Search mampu menghasilkan kinerja terbaik dengan akurasi pengujian sebesar 81.20%, lebih tinggi dibandingkan dengan optimasi menggunakan Random Search. Hasil ini menunjukkan bahwa LightGBM mampu memanfaatkan fitur hasil ekstraksi mendalam ResNet-50 secara efektif dan stabil, terutama ketika dikombinasikan dengan konfigurasi hyperparameter yang tepat.

Namun demikian, hasil eksperimen juga memperlihatkan bahwa model ResNet-50 yang digunakan secara end-to-end memberikan performa yang lebih tinggi dibandingkan pendekatan hibrid. Hal ini mengindikasikan bahwa fitur hasil pembelajaran mendalam pada ResNet-50 telah cukup representatif untuk memisahkan kelas genre musik, sehingga penambahan classifier berbasis ensemble tidak selalu menghasilkan peningkatan akurasi yang signifikan.

Meskipun demikian, kelebihan utama dari eksperimen yang dilakukan terletak pada penyusunan pipeline evaluasi yang sistematis dan komprehensif, mulai dari segmentasi audio, representasi Mel-spectrogram, ekstraksi fitur berbasis transfer learning, hingga optimasi hyperparameter pada model ensemble. Pendekatan ini memungkinkan analisis yang lebih mendalam terhadap peran masing-masing komponen dalam sistem klasifikasi genre musik.

Selain itu, eksperimen ini memberikan wawasan empiris bahwa penggunaan classifier lanjutan seperti LightGBM dapat menghasilkan performa yang relatif stabil dengan tingkat generalisasi yang baik. Oleh karena itu, pendekatan ResNet-50 dan LightGBM dapat dipertimbangkan sebagai alternatif atau model pembanding, khususnya dalam skenario yang mengutamakan efisiensi komputasi, fleksibilitas pemodelan, atau kebutuhan untuk menghindari pelatihan model deep learning secara end-to-end.

Daftar Referensi

- [1] S. Surahman, U. Bhayangkara, and J. Raya, *Musik Di Indonesia Tiga Masa : Klasik , Modern , dan Postmodern*, no. August. 2024.
- [2] P. Hendra, S. S. Sekolah, T. Teologi, and P. Kebenaran, "Musik dalam Dinamika Pujian Penyembahan," *Pneum. J. Teol. Kependetaan*, vol. 10, no. 2, pp. 176–199, 2020, doi: 10.56438/pneuma.v10i2.26.
- [3] Kim, A.J. Differential Effects of Musical Expression of Emotions and Psychological Distress on Subjective Appraisals and Emotional Responses to Music. *Behav. Sci.* 2023, 13, 491. [https://doi.org/ 10.3390/bs13060491](https://doi.org/10.3390/bs13060491)
- [4] Y. Hu and G. Mogos, "Music genres classification by deep learning," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 2, pp. 1186–1198, 2022, doi: 10.11591/ijeecs.v25.i2.pp1186-1198.
- [5] T. Tinaliah *et al.*, "Penerapan Convolutional Neural Network Untuk Klasifikasi Citra Ekspresi Wajah Manusia Pada MMA Facial Expression Dataset," vol. 8, no. 4, pp. 2051–2059, 2021, doi: <https://doi.org/10.35957/jatisi.v8i4.1437>.

- [6] R. F. Fadhilah and R. Sumiharto, "Klasifikasi Suara Untuk Memonitorid Hutan Berbasis Convolutional Neural Network," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, vol. 13, no. 1, pp. 13–22, 2023, doi: 10.22146/ijeis.79536.
- [7] M. H. Tanveer, H. Zhu, W. Ahmed, A. Thomas, B. M. Imran, and M. Salman, "Mel-spectrogram and Deep CNN Based Representation Learning from Bio-Sonar Implementation on UAVs," *2021 Int. Conf. Comput. Control Robot. ICCCR 2021*, pp. 220–224, 2021, doi: 10.1109/ICCCR49711.2021.9349416.
- [8] K. Indra and M. Agung, "Analisis Komparatif Arsitektur CNN dan VGG16 pada Klasifikasi Genre Musik," *J. Nas. Teknol. Inf. dan Apl.*, vol. 3, pp. 889–898, 2025, doi: <https://doi.org/10.24843/JNATIA.2025.v03.i04.p19>.
- [9] Y. V. Via, I. Y. Purbasari, and A. P. Pratama, "Analisa Algoritma Convolution Neural Network (Cnn) Pada Klasifikasi Genre Musik Berdasar Durasi Waktu," *Scan J. Teknol. Inf. dan Komun.*, vol. 17, no. 1, pp. 35–41, 2022, doi: 10.33005/scan.v17i1.3251.
- [10] Y. A. Auliya, D. I. Swasono, and P. P. Harwanto, "Classification of Indonesian Music Genres Using Transfer Learning with ResNet-50 and Mel-Frequency Cepstral Coefficient Feature Extraction," *Math. Model. Eng. Probl.*, vol. 12, no. 3, pp. 840–850, 2025, doi: 10.18280/MMEP.120310.
- [11] Z. R. LI Xingjian, TANG Xinyi, "Deep LightGBM Sound classification using Deep LightGBM algorithm," *Tech. Acoust.*, vol. 41, no. 6, pp. 871–877, 2022, doi: 10.16300/j.cnki.1000-3630.2022.06.012.
- [12] I. R. Huriyah, A. Ismania, and S. Widianingrum, "Optimization of Data Augmentation Based on Synonym Replacement in News Text Classification Using Neural Network," *Komputa J. Ilm. Komput. dan Inform.*, vol. 14, no. 1, pp. 100–107, 2025, doi: 10.34010/komputa.v14i1.
- [13] R. F. Junaidi *et al.*, "Baby Cry Sound Detection: A Comparison of Mel Spectrogram Image on Convolutional Neural Network Models," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 4, pp. 355–369, 2024, doi: 10.35882/jeeemi.v6i4.465.
- [13] R. F. Junaidi *et al.*, "Baby Cry Sound Detection: A Comparison of Mel Spectrogram Image on Convolutional Neural Network Models," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 4, pp. 355–369, 2024, doi: 10.35882/jeeemi.v6i4.465.
- [14] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," *Appl. Sci.*, vol. 12, no. 18, pp. 1–43, 2022, doi: 10.3390/app12188972.
- [15] T. Hascoet, Q. Febvre, W. Zhuang, Y. Ariki, and T. Takiguchi, "Reversible designs for extreme memory cost reduction of CNN training," *EURASIP J. Image Video Process.*, vol. 2023, no. 1, p. 1, 2023, doi: 10.1186/s13640-022-00601-w.