PENGENALAN EMOSI MUSIK MULTIMODAL BERBASIS LATE FUSION PADA DATASET MULTI-MODAL MIREX

Preliminary Experiment

Kelompok 09

Lois Novel E Gurning	122140098
Sakti Mujahid Imani	122140123
Apridian Saputra	122140143
Joshia Fernandes Sectio Purba	122140170
Sikah Nubuahtul Ilmi	122140208

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INDUSTRI

INSTITUT TEKNOLOGI SUMATERA

2025

DAFTAR ISI

1. Ringkasan Singkat dan Tujuan

Proyek ini berfokus pada pengembangan sistem *Music Emotion Recognition* (MER) berbasis multimodal dengan memanfaatkan tiga modalitas utama yaitu audio, lirik, dan MIDI yang tersedia pada dataset MIREX hasil akuisisi otomatis dari AllMusic. Audio cenderung merepresentasikan energi serta karakter akustik, lirik mengandung makna semantik yang berkaitan langsung dengan ekspresi emosional, sedangkan MIDI menyimpan struktur musikal yang relevan terhadap nuansa dan progresi melodi. Temuan EDA sebelumnya mengindikasikan bahwa representasi tiap modalitas belum membentuk klaster emosi yang terpisah secara tegas dalam ruang fitur berdimensi rendah, terutama pada MIDI yang sangat terbatas jumlah datanya. Hal ini memperkuat asumsi bahwa pendekatan unimodal akan kesulitan menembus *glass ceiling effect* yang dilaporkan dalam literatur.

Berdasarkan dari temuan tersebut, proyek ini ditujukan untuk merancang dan mengevaluasi model baseline multimodal menggunakan strategi *Late Fusion*, di mana tiap modalitas diproses oleh model pembelajaran mendalam yang optimal untuk karakteristik datanya, CRNN untuk audio, BERT untuk lirik, dan BiGRU untuk MIDI. Hasil prediksi dari masing-masing model kemudian digabungkan untuk menghasilkan keputusan akhir. Pendekatan ini diharapkan mampu meningkatkan akurasi klasifikasi dengan memanfaatkan keunggulan tiap modalitas secara terpisah.

Hipotesis awal kelompok menyatakan bahwa:

1. Model multimodal akan memberikan performa yang lebih stabil dan akurat dibandingkan model unimodal, terutama dalam kelas-kelas yang sulit dibedakan menggunakan satu modalitas saja.
2. Strategi *Late Fusion* merupakan pilihan terbaik mengingat distribusi data antar-modalitas yang tidak seimbang dan keterbatasan jumlah sampel MIDI.
3. Peningkatan performa terbesar diprediksi datang dari kombinasi audio dan lirik, karena keduanya memiliki ketersediaan data yang lebih besar serta korelasi yang lebih kuat terhadap anotasi emosi dibandingkan MIDI.

Kami juga telah melakukan percobaan unimodal di Collab

Lirik : Processing Lyrics

Audio : audio-testing.ipynb

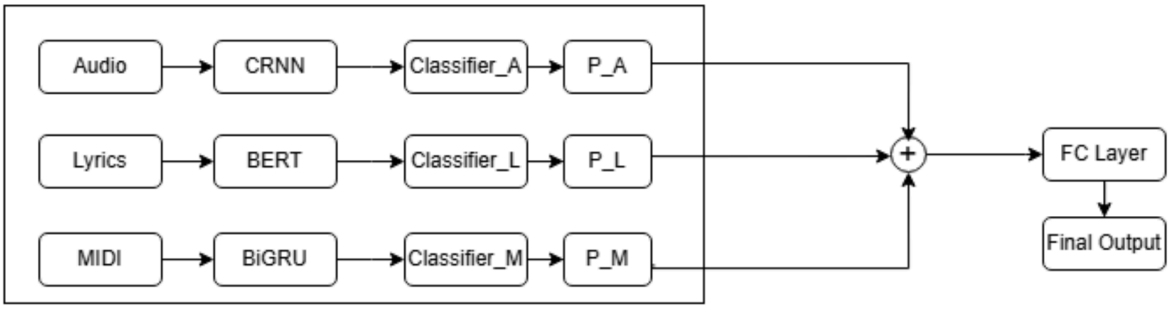
MIDI : midi-adding.ipynb

2. Arsitektur Model Baseline

Model baseline yang digunakan dalam proyek ini dirancang untuk memanfaatkan kekuatan masing-masing modalitas secara terpisah sebelum digabungkan melalui mekanisme *Late Fusion*. Strategi ini dipilih berdasarkan temuan EDA yang menunjukkan ketidakseimbangan ketersediaan data antar modalitas, audio: 903, lirik: 764, MIDI: 193, serta hasil analisis bahwa penggabungan awal (*early fusion*) akan memaksa jumlah sampel efektif turun mengikuti modalitas paling sedikit (MIDI). Oleh karena itu, *Late Fusion* memungkinkan setiap model unimodal belajar secara optimal tanpa kehilangan data berharga dari modalitas lain.

2.1 Arsitektur Umum Baseline Multimodal

Arsitektur baseline mengikuti struktur tiga cabang (audio, lirik, MIDI) yang masing-masing memproses input dengan model yang berbeda sesuai karakteristik datanya. Setiap cabang mengeluarkan vektor representasi atau probabilitas kelas (*logit*) yang kemudian digabungkan pada tahap akhir. Diagram arsitektur secara sederhana berupa skema dengan tiga jalur pemrosesan paralel (CRNN–BERT–BiGRU) yang bertemu pada satu titik fusi sebelum masuk ke *Fully Connected Layer* final



Gambar 2.1 Diagram Arsitektur

2.1.1 Modalitas Audio (CRNN)

Pemilihan Convolutional Recurrent Neural Network (CRNN) didasarkan pada kebutuhan untuk menangkap pola frekuensi waktu dari Mel-Spectrogram secara bersamaan. CNN bertugas mempelajari pola lokal seperti timbre, kontur frekuensi, dan perbedaan spektral antar emosi. RNN (LSTM/GRU) kemudian mengekstraksi dinamika temporal sepanjang 30 detik klip audio, merepresentasikan perubahan intensitas dan energi yang berkaitan erat dengan arousal. CRNN menjadi baseline kuat karena mampu mengombinasikan representasi spasial dan temporal secara efektif.

2.1.2 Modalitas Lirik (BERT)

Lirik diproses dengan *Bidirectional Encoder Representations from Transformers* (*BERT*), yang memiliki kemampuan memahami konteks dua arah.

Keunggulan BERT sebagai baseline antara lain:

1. Mampu memahami relasi semantik mendalam yang muncul dalam lirik
2. Menangkap nuansa emosi secara implisit (misalnya *longing*, *anger*, *nostalgia*),
3. Lebih unggul dibanding model sekuensial murni seperti LSTM dalam tugas klasifikasi teks.

Tahap pengolahan mencakup tokenisasi, padding hingga panjang maksimal tertentu, dan *fine-tuning* BERT pada task klasifikasi lima kelas emosi.

2.1.3 Modalitas MIDI (BiGRU)

Data MIDI berbentuk rangkaian peristiwa musikal (pitch, durasi, velocity) sehingga membutuhkan model mampu mempelajari pola sekuensial.

Bidirectional GRU (BiGRU) dipilih karena:

1. Mampu memanfaatkan konteks ke belakang dan ke depan dalam urutan nada,
2. Lebih ringan dari LSTM namun tetap stabil untuk dataset kecil,
3. Efektif menangkap progresi melodi, ritme, dan densitas nada yang berhubungan dengan emosi musik.

Representasi MIDI dikonversi menjadi urutan pitch berukuran tetap sebelum masuk ke embedding layer dan BiGRU.

2.1.4 Mekanisme Late Fusion

Setelah masing-masing model menghasilkan prediksi (logit atau probabilitas), nilai dari tiap modalitas digabungkan menggunakan operasi penjumlahan atau konkatenasi (bergantung implementasi baseline). Lapisan *Fully Connected* terakhir berfungsi sebagai *meta-classifier* yang mempelajari bobot kontribusi optimal dari masing-masing modalitas. Softmax kemudian menghasilkan lima probabilitas kelas emosi akhir.

Keuntungan utama *Late Fusion*:

1. Tidak sensitif terhadap *missing modality*.
2. Memungkinkan pemanfaatan kapasitas penuh tiap modalitas.
3. Memberikan interpretabilitas lebih baik terkait kontribusi masing-masing model.

3. Setup Eksperimen

Perancangan eksperimen dilakukan untuk memastikan bahwa setiap modalitas menghasilkan representasi yang optimal sebelum digunakan dalam proses fusi multimodal. Setiap modalitas memiliki karakteristik data yang berbeda, sehingga pendekatan pemrosesan dan konfigurasi model disesuaikan secara khusus. Setup eksperimen baseline disusun berdasarkan hasil EDA awal, desain arsitektur pada proposal, serta implementasi unimodal yang telah dilakukan untuk audio, lirik, dan MIDI. Bagian ini mencakup tiga aspek yaitu *data splitting*, *preprocessing & feature extraction*, dan *hyperparameter* yang digunakan untuk masing-masing modalitas.

3.1 Modalitas Audio

Modalitas audio menyediakan gambaran langsung terkait karakter akustik musik, seperti intensitas, timbre, dan dinamika frekuensi yang berperan penting dalam pembentukan persepsi emosi. Berdasarkan analisis awal pada dataset, mayoritas file audio memiliki durasi konsisten 30 detik dan menunjukkan perbedaan pola energi yang cukup jelas antar klaster emosi, terutama antara klaster bernuansa tinggi seperti Passionate dan klaster bernuansa rendah seperti Wistful. Konsistensi ini memungkinkan proses pemrosesan dan ekstraksi fitur dilakukan secara seragam pada seluruh sampel.

3.1.1 Data Splitting

Pembagian data pada modalitas audio dilakukan menggunakan pendekatan stratifikasi untuk mempertahankan keseimbangan proporsi label. Dataset berisi 903 sampel audio, dan pembagian dilakukan sebagai berikut:

Train : 80%
Validation : 20%
Metode : StratifiedShuffleSplit

Stratifikasi diterapkan untuk memastikan distribusi kelas tetap stabil di antara dua subset. Ketidakseimbangan kelas diketahui menjadi karakteristik dataset ini sehingga menjaga proporsi kelas menjadi langkah penting agar model tidak bias terhadap kategori dominan

3.1.2 Preprocessing dan Feature Extraction

Pendekatan pemrosesan audio disusun berdasarkan karakteristik dataset MIREX serta kebutuhan model CRNN untuk mendapatkan representasi frekuensi dan waktu yang

informatif.

Preprocessing

Tahapan yang diterapkan meliputi:

1. Konversi menjadi sinyal mono untuk menyatukan seluruh kanal sehingga representasi tidak dipengaruhi perbedaan stereo.
2. Penyeragaman sampling rate ke 22.050 Hz guna memastikan setiap audio dianalisis pada resolusi sinyal yang sama.
3. Normalisasi durasi menjadi 30 detik melalui pemotongan bagian berlebih atau penambahan padding pada sampel yang lebih pendek. Durasi ini dipilih karena sesuai dengan mayoritas panjang file dalam dataset asli.
4. Normalisasi amplitudo menggunakan z-score sehingga intensitas suara antar file menjadi stabil dan tidak menyebabkan dominasi pada proses pembelajaran.

Tahapan ini memastikan bahwa semua audio berada pada kondisi akustik yang seragam sebelum diekstraksi menjadi fitur.

Feature Extraction

Representasi audio dibentuk menggunakan Log-Mel Spectrogram, yang mencerminkan distribusi energi pada berbagai pita frekuensi secara temporal. Pengaturan yang digunakan meliputi:

1. 128 mel bands
2. Ukuran jendela transformasi 1024 sampel
3. Langkah geser 512 sampel
4. Konversi energi menjadi skala logaritmik

Hasil ekstraksi berupa tensor dua dimensi yang menggambarkan pola frekuensi-waktu. Representasi ini terbukti menampilkan perbedaan karakteristik antar klaster emosi; misalnya klaster bernuansa kuat dan energik cenderung memiliki intensitas spektral lebih padat dibandingkan klaster bernuansa lembut.

3.1.3 Hyperparameter

Pelatihan model CRNN menggunakan konfigurasi berikut:

Hyperparameter	Nilai
Batch size	8

Hyperparameter	Nilai
Learning rate	0.001
Optimizer	Adam
Epochs	20
Loss	CrossEntropy

Pengaturan ini dipilih untuk menyeimbangkan stabilitas pelatihan dan kapasitas model dalam mengolah input spektrogram berdimensi besar. Batch size yang kecil membantu menjaga proses komputasi tetap stabil, sementara jumlah epoch disesuaikan agar model memiliki waktu cukup untuk mempelajari pola tanpa risiko overfitting yang berlebihan.

3.2 Modalitas Lirik

Modalitas lirik digunakan untuk menangkap informasi semantik dan ekspresi emosional yang muncul dalam bentuk kata dan frasa. Lirik lagu sering kali memuat ungkapan perasaan, narasi, dan kosakata emosional yang tidak muncul dalam sinyal audio, sehingga pendekatan pemrosesan teks menjadi bagian penting dalam sistem multimodal. Proses pengolahan lirik difokuskan pada pembentukan dataset yang bersih, representatif, serta pemetaan label yang akurat sebelum digunakan untuk melatih model berbasis Transformer.

3.2.1 Data Splitting

Dataset lirik pada awalnya terdiri dari seluruh file teks yang dikaitkan dengan identitas lagu. Namun, sebagian file tidak ditemukan, tidak terbaca, atau berisi teks kosong. Setelah proses penyaringan dilakukan, total 709 sampel lirik dinyatakan valid dan digunakan sebagai dataset akhir.

Pembagian dataset dilakukan menggunakan:

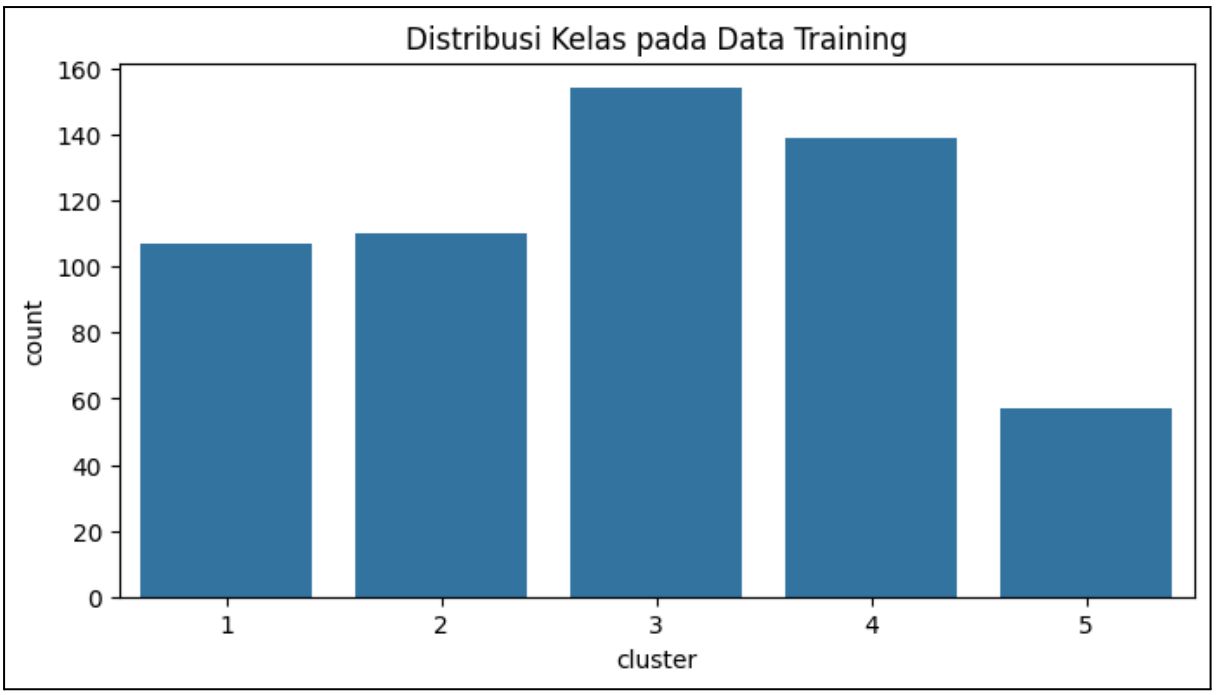
Train : 80%

Validation : 10%

Test : 10%

Metode : Stratifikasi berdasarkan lima label emosi

Stratifikasi diperlukan karena distribusi label lirik tidak seimbang, di mana beberapa kategori emosi memiliki jumlah sampel jauh lebih sedikit dibanding kategori lainnya. Dengan menjaga proporsi label tetap konsisten di semua subset, model dapat mempelajari pola semantik secara lebih stabil dan evaluasi menjadi lebih representatif.



Gambar 3.2 Distribusi kelas pada data *training* lirik

3.2.2 Preprocessing dan Feature Extraction

Pemrosesan lirik dilakukan melalui serangkaian tahapan yang memastikan bahwa teks dalam kondisi bersih dan dapat diproses oleh model berbasis Transformer.

Preprocessing

Proses awal difokuskan pada penyusunan dataset dan pembersihan teks:

1. Pemetaan dan pembacaan file lirik

   Setiap file teks dicari berdasarkan nama file yang terhubung dengan label emosi. Hanya file yang berhasil dibaca dan memiliki isi yang valid yang dimasukkan ke dataset.

2. Penyaringan file tidak valid

   File yang hilang, rusak, atau berisi teks kosong dikeluarkan dari dataset untuk menjaga kualitas pelatihan.

3. Normalisasi teks

   Pembersihan dilakukan secara ringan, termasuk:

   a. Penyesuaian huruf menjadi bentuk seragam,

   b. Pengurangan whitespace berlebih,

   c. Menjaga struktur kalimat tanpa menghilangkan tanda baca penting.

   Cara ini dipilih agar makna emosional yang terkandung dalam kalimat tetap terjaga.

Tahapan ini menghasilkan korpus teks bersih yang siap diproses lebih lanjut.

Feature Extraction

Representasi lirik dibentuk melalui proses tokenisasi dan embedding:

1. Tokenisasi menggunakan tokenizer BERT

   Teks diubah menjadi rangkaian token subword yang dapat mewakili kosakata umum maupun kata unik dalam lirik.
2. Padding dan truncation pada panjang tetap 256 token

   Panjang ini dipilih karena sesuai dengan distribusi panjang lirik dalam dataset serta efisien untuk pelatihan model.
3. Penyusunan input untuk model

   Hasil tokenisasi menghasilkan dua komponen:

   a. *input ids*

   b. *attention mask*

   kemudian digabungkan dengan label untuk membentuk dataset tensor.
4. Ekstraksi fitur kontekstual

   Pada tahap pelatihan, representasi fitur tidak diekstraksi secara manual. Model mengambil embedding dari token [CLS], yang mewakili makna keseluruhan teks dalam konteks model Transformer.

Pendekatan ini memungkinkan model memahami nuansa semantik seperti kata bermakna emosional (*pain*, *hurt*, *love*, *empty*), narasi, maupun pola bahasa khas tiap kategori.

3.2.3 Hyperparameter

Pelatihan model lirik menggunakan konfigurasi hyperparameter berikut:

Hyperparameter	Nilai
Batch size	16
Learning rate	2e-5
Optimizer	AdamWW
Epochs	10
Loss	CrossEntropy

Konfigurasi ini dipilih untuk memastikan proses fine-tuning berlangsung stabil. Learning rate yang kecil memungkinkan model beradaptasi tanpa mengubah bobot representasi dasar secara berlebihan, sementara batch size 16 memberikan keseimbangan antara kualitas gradien

dan penggunaan memori. Jumlah epoch ditetapkan berdasarkan ukuran dataset serta kompleksitas representasi teks.

3.3 Modalitas MIDI

Modalitas MIDI memberikan representasi simbolik dari musik, berupa rangkaian nada dan durasi yang tidak dipengaruhi kualitas rekaman. Berbeda dari audio yang memuat informasi gelombang suara, MIDI menyimpan struktur musikal seperti urutan pitch, perubahan nada, dan kepadatan not. Struktur ini berpotensi memuat pola-pola yang berkaitan dengan emosi, seperti kontur melodi yang naik-turun, repetitif, atau padat. Proses pemrosesan MIDI berfokus pada mengekstraksi urutan nilai pitch yang dapat digunakan oleh model sekuensial.

3.3.1 Data Splitting

Dataset MIDI memiliki jumlah sampel paling sedikit dibandingkan modalitas lain. Dari seluruh daftar lagu, hanya sebagian yang memiliki file MIDI lengkap dan dapat diproses, dan setelah penyaringan hanya 193 sampel MIDI yang valid. File lain dikeluarkan karena hilang, tidak dapat dibaca, atau tidak memuat informasi pitch.

Pembagian data dilakukan dengan proporsi:

Train : 80%

Validation : 20%

Metode : Stratified split berdasarkan lima kategori emosi

Meskipun jumlah MIDI relatif kecil, stratifikasi tetap digunakan agar setiap klaster emosi tetap terwakili dalam proporsi yang stabil pada train dan validation. Hal ini penting karena ketidakseimbangan label dapat membuat model kesulitan mengenali pola pada kategori minoritas.

3.3.2 Preprocessing dan Feature Extraction

Pemrosesan MIDI dilakukan dengan mengekstraksi informasi pitch yang menjadi dasar bagi representasi musik simbolik. Tahapannya meliputi:

Preprocessing

1. Pembacaan file MIDI

   Setiap file MIDI dianalisis untuk mengekstrak nada dari seluruh instrumen non-perkusi. Informasi yang diambil mencakup pitch dan urutan kemunculannya.
2. Penyaringan bagian drum atau perkusi

Instrumen dengan saluran perkusi dihapus karena tidak memuat informasi nada.

3. Penyusunan urutan pitch

   Nada dalam file diurutkan berdasarkan waktu kemunculannya sehingga membentuk rangkaian pitch yang merepresentasikan garis melodi atau struktur musik utama.

4. Standarisasi panjang urutan

   Setiap file dikonversi menjadi urutan pitch dengan panjang tetap 500 token:

   a. Jika urutan lebih pendek → *padding*,

   b. Jika lebih panjang → *truncation*.

   Padding dilakukan menggunakan nilai khusus agar tidak dibaca sebagai pitch valid.

Tahapan ini menghasilkan representasi simbolik yang homogen untuk seluruh sampel.

Feature Extraction

Fitur MIDI diproses dalam bentuk urutan embedding pitch untuk digunakan sebagai input model sekuensial. Prosesnya meliputi:

1. Pemetaan pitch ke embedding vektor berdimensi tetap

   Embedding ini memetakan setiap nilai pitch (0–127) ke representasi kontinu yang dapat dipelajari model.

2. Pemodelan urutan menggunakan jaringan BiGRU

   Pendekatan ini memungkinkan model memahami konteks nada dari dua arah, baik dari awal ke akhir maupun sebaliknya, yang penting untuk menangkap struktur melodi.

3. Representasi akhir diambil dari keluaran jaringan sekuensial

   Vektor keluaran digunakan sebagai dasar klasifikasi pada lima kategori emosi.

Pendekatan ini meniru cara analisis linier terhadap melodi, di mana pola interval, repetisi, dan kompleksitas melodi dapat mencerminkan karakter emosi tertentu.

3.3.3 Hyperparameter

Pelatihan model MIDI menggunakan arsitektur BiGRU dengan konfigurasi berikut:

Hyperparameter	Nilai
Batch size	32
Learning rate	0.001
Optimizer	Adam
Epochs	30 (dengan early stopping)
Loss	CrossEntropy

Batch size 32 sesuai dengan ukuran input sekuens pitch yang relatif kecil sehingga efisien untuk pelatihan. Learning rate 0.001 memberikan kecepatan pembaruan parameter yang stabil pada model sekuensial ringan. 30 epoch memberi ruang yang cukup bagi model untuk mempelajari pola, namun penggunaan early stopping mencegah overfitting. Adam dipilih karena performanya yang konsisten pada model berbasis urutan.

4. Hasil Baseline & Analisis Awal

Evaluasi baseline dilakukan secara terpisah pada masing-masing modalitas untuk melihat kemampuan unimodal dalam mengenali emosi musik. Hasil ini menjadi fondasi untuk memahami kontribusi tiap modalitas sebelum dilakukan penggabungan pada tahap multimodal. Pada tahap awal, seluruh model dilatih menggunakan konfigurasi dasar tanpa teknik optimasi lanjutan, sehingga hasil baseline mencerminkan performa awal yang bersifat *straightforward*.

4.1 Modalitas Audio

Evaluasi baseline pada modalitas audio dilakukan menggunakan model CRNN dengan input berupa log-mel spectrogram. Hasil training menunjukkan bahwa model mampu menangkap sebagian pola akustik yang berkaitan dengan emosi, namun performa keseluruhan masih terbatas karena tumpang tindih karakteristik antar-klaster. Bagian ini merangkum matrix evaluasi, perilaku learning curve, serta analisis terhadap kesalahan model.

4.1.1 Matrix Evaluasi dan Confusion Matrix

Model menghasilkan performa berikut pada data validasi:

1. Akurasi validasi ≈ 0.44
2. Performa antarklaster tidak merata, dengan skor tertinggi pada klaster bernuansa "melankolis" dan skor terendah pada klaster energik & ceria
3. Precision dan recall bervariasi antar kelas karena distribusi label tidak seimbang

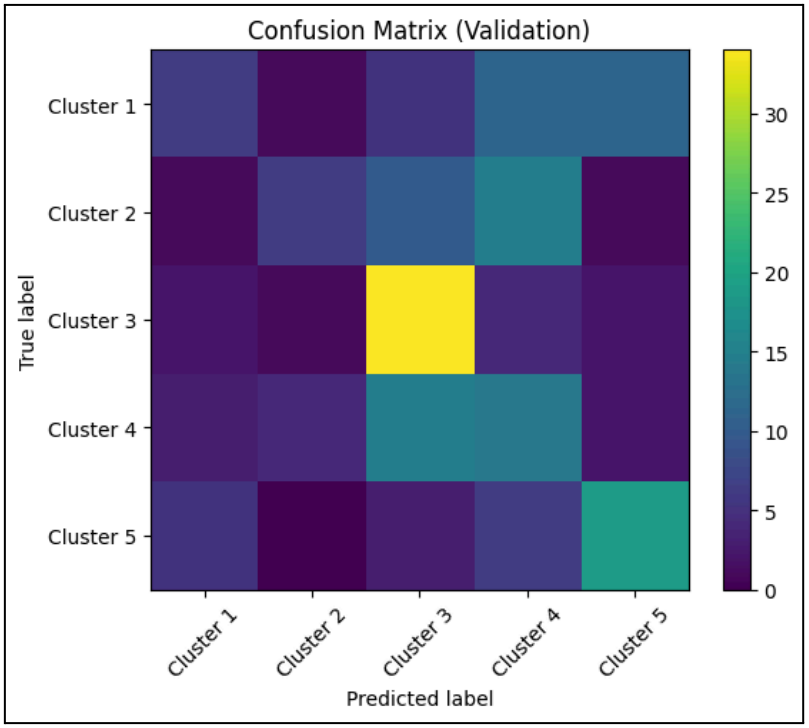
```
Validation accuracy (audio-only CRNN): 0.4365

Classification report (val):
              precision    recall  f1-score   support

   Cluster 1       0.35      0.18      0.24        34
   Cluster 2       0.50      0.18      0.27        33
   Cluster 3       0.51      0.79      0.62        43
   Cluster 4       0.28      0.37      0.32        38
   Cluster 5       0.54      0.58      0.56        33

    accuracy                           0.44       181
   macro avg       0.44      0.42      0.40       181
weighted avg       0.44      0.44      0.41       181
```

Gambar 4.1 Matrix Evaluasi Audio

Confusion matrix menunjukkan pola berikut:

1. Klaster bernuansa tinggi cenderung saling tertukar, terutama antara kategori energik dan kategori ceria

2. Klaster mendatar atau low-energy lebih mudah dikenali karena pola spektralnya lebih seragam
3. Mislabeling paling sering terjadi pada pasangan klaster yang memiliki kontur energi yang mirip



Gambar 4.2 Confusion Matrix Audio

4.1.2 *Learning Curve* (*Loss* dan *Accuracy*)

Learning curve pelatihan memperlihatkan pola berikut:

1. *Training loss* menurun secara konsisten sepanjang 20 epoch
2. *Validation loss* berfluktuasi, namun tidak menunjukkan kenaikan drastis yang menandakan overfitting berat
3. *Validation accuracy* cenderung stagnan setelah pertengahan pelatihan, menunjukkan bahwa model mengalami keterbatasan dalam mempelajari pola yang lebih dalam dari fitur akustik
4. Gap antara training dan validation relatif kecil, menandakan bahwa model tidak mengalami underfitting parah, tetapi juga belum mampu membedakan klaster dengan baik

Gambar 4.3 Grafik *Learning Curve (Loss & Metric)* Audio

4.1.3 Analisis *Success* dan *Failure Case*

Hasil evaluasi menunjukkan beberapa pola keberhasilan dan kegagalan yang konsisten pada berbagai sampel:

Success Cases

1. Lagu dengan perubahan energi yang jelas antara verse dan chorus cenderung berhasil dipetakan ke klaster berenergi tinggi.
2. Sampel dengan struktur harmonik sederhana tetapi memiliki distribusi energi yang khas (misal mid-frequency dominan) lebih mudah dikenali.
3. Klaster bernuansa melankolis menunjukkan performa terbaik karena pola spektral yang relatif konsisten di seluruh dataset.

Failure Cases

1. Lagu dengan instrumen homogen (misal hanya gitar atau piano) sulit dibedakan secara emosional karena pola frekuensinya mirip antara klaster.
2. Sampel yang memiliki overlap intensitas dengan klaster tetangga sering tertukar, misalnya lagu energik bertempo lambat diklasifikasikan sebagai ceria.

   Lagu dengan dinamika energi datar sulit diklasifikasikan karena tidak memiliki ciri akustik emosional yang kuat.

```
Epoch 01 | train_loss=1.6257 acc=0.2285 | val_loss=1.5972 acc=0.2762
  ✓ saved best model
Epoch 02 | train_loss=1.5706 acc=0.2950 | val_loss=1.5062 acc=0.2928
  ✓ saved best model
Epoch 03 | train_loss=1.5473 acc=0.3075 | val_loss=1.6710 acc=0.1989
Epoch 04 | train_loss=1.5084 acc=0.3186 | val_loss=1.4390 acc=0.3757
  ✓ saved best model
Epoch 05 | train_loss=1.4975 acc=0.3283 | val_loss=1.4230 acc=0.3757
Epoch 06 | train_loss=1.4587 acc=0.3657 | val_loss=1.4217 acc=0.4088
  ✓ saved best model
Epoch 07 | train_loss=1.4451 acc=0.3670 | val_loss=1.4054 acc=0.3923
Epoch 08 | train_loss=1.4388 acc=0.3684 | val_loss=1.3953 acc=0.3923
Epoch 09 | train_loss=1.4265 acc=0.3795 | val_loss=1.5475 acc=0.2983
Epoch 10 | train_loss=1.4001 acc=0.3850 | val_loss=1.4002 acc=0.3702
Epoch 11 | train_loss=1.4275 acc=0.3864 | val_loss=1.4343 acc=0.3702
Epoch 12 | train_loss=1.3938 acc=0.4003 | val_loss=1.3853 acc=0.4199
  ✓ saved best model
Epoch 13 | train_loss=1.4063 acc=0.3961 | val_loss=1.5915 acc=0.3149
Epoch 14 | train_loss=1.3768 acc=0.4169 | val_loss=1.3669 acc=0.3923
Epoch 15 | train_loss=1.3777 acc=0.4017 | val_loss=1.4700 acc=0.3646
Epoch 16 | train_loss=1.3557 acc=0.4127 | val_loss=1.3439 acc=0.4365
  ✓ saved best model
Epoch 17 | train_loss=1.3673 acc=0.4169 | val_loss=1.3710 acc=0.4033
Epoch 18 | train_loss=1.3218 acc=0.4307 | val_loss=1.3975 acc=0.4199
Epoch 19 | train_loss=1.3517 acc=0.4155 | val_loss=1.3698 acc=0.3923
Epoch 20 | train_loss=1.3269 acc=0.4391 | val_loss=1.3543 acc=0.4088
```

Gambar 4.4 Hasil *Training* Audio

4.1.4 Kesimpulan Awal Modalitas Audio

1. Representasi akustik belum cukup untuk memisahkan seluruh kategori emosi secara kuat.
2. Tumpang tindih karakteristik antar-klaster menjadi tantangan utama.
3. Performanya masih dapat ditingkatkan dengan arsitektur lebih kompleks, regularisasi lebih baik, atau kombinasi modalitas.

4.2 Modalitas Lirik

Evaluasi baseline pada modalitas audio dilakukan menggunakan model CRNN dengan input berupa log-mel spectrogram. Hasil training menunjukkan bahwa model mampu menangkap sebagian pola akustik yang berkaitan dengan emosi, namun performa keseluruhan masih terbatas karena tumpang tindih karakteristik antar-klaster. Bagian ini merangkum matrix evaluasi, perilaku learning curve, serta analisis terhadap kesalahan model.

4.2.1 Matrix Evaluasi dan Confusion Matrix

Model menghasilkan performa berikut pada data validasi:

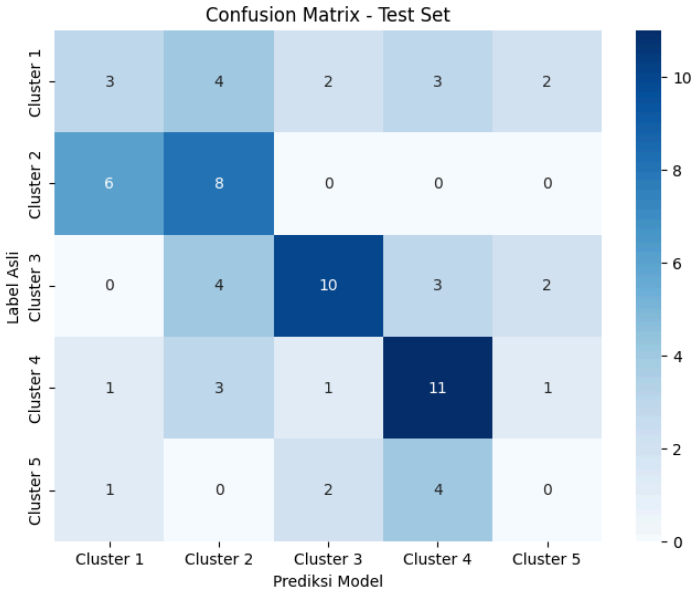
1. Akurasi validasi berada pada rentang ≈ 0.38
2. Kelas yang memiliki kosakata emosional eksplisit (misalnya sedih, rindu, perpisahan) menunjukkan performa lebih tinggi
3. Kelas dengan makna ambigu atau konteks bercampur lebih sering tertukar

	precision	recall	f1-score	support
Cluster 1	0.27	0.21	0.24	14
Cluster 2	0.42	0.57	0.48	14
Cluster 3	0.67	0.53	0.59	19
Cluster 4	0.52	0.65	0.58	17
Cluster 5	0.00	0.00	0.00	7
accuracy			0.45	71
macro avg	0.38	0.39	0.38	71
weighted avg	0.44	0.45	0.44	71

Gambar 4.5 Matrix Evaluasi Lirik

Confusion matrix menunjukkan pola berikut:

1. Klaster bernuansa melankolis dan reflektif lebih mudah dikenali karena banyak menggunakan kata-kata emosional yang jelas
2. Klaster bernuansa positif atau energik sering tertukar satu sama lain karena kosakata yang digunakan terkadang serupa (misal kata “love” dapat muncul dalam konteks ceria maupun sedih)
3. Misclass terbesar berasal dari klaster dengan lirik bersifat naratif, panjang, atau mengandung makna ganda



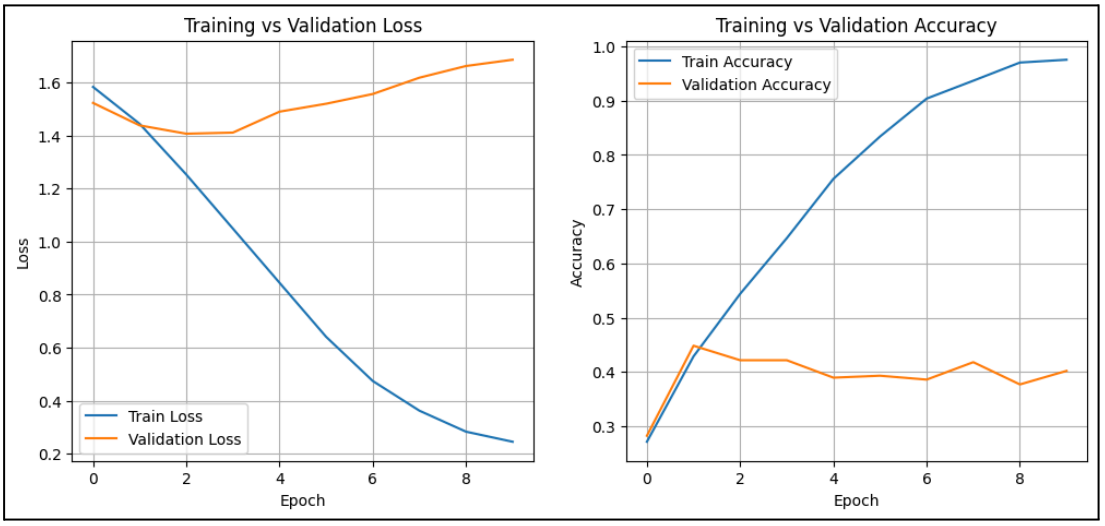
Gambar 4.6 Confusion Matrix Lirik

4.2.2 *Learning Curve* (*Loss* dan *Accuracy*)

Learning curve pada model lirik menunjukkan pola pelatihan yang relatif stabil:

1. Training loss menurun secara konsisten pada hampir seluruh epoch
2. Validation loss memiliki bentuk yang mirip dengan training loss, menandakan bahwa model tidak mengalami overfitting signifikan
3. Validation accuracy naik secara bertahap dan mencapai plateau setelah beberapa epoch
4. Perbedaan antara train dan validation relatif kecil, menunjukkan generalisasi yang baik pada dataset yang seimbang setelah stratifikasi

Pola ini mencerminkan bahwa representasi kontekstual dari BERT mampu menangkap nuansa semantik yang relevan bagi klasifikasi emosional.



Gambar 4.7 Grafik *Learning Curve (Loss & Metric)* Lirik

4.2.3 Analisis *Success* dan *Failure Case*

Hasil evaluasi menunjukkan beberapa pola keberhasilan dan kegagalan yang konsisten pada berbagai sampel:

Success Cases

Model cenderung berhasil mengklasifikasikan lirik yang

1. Mengandung kata-kata emosional eksplisit, seperti *sad*, *lonely*, *forever*, *broken*, *heart*, *empty*
2. Menggunakan pola bahasa yang jelas menggambarkan suasana tertentu, seperti narasi kehilangan atau kerinduan

3. Memiliki struktur kalimat sederhana dan langsung, sehingga konteks emosinya mudah diambil oleh embedding

Contoh keberhasilan umum:

1. Lirik bernuansa sedih atau reflektif hampir selalu masuk kategori yang tepat
2. Lirik bertema cinta yang romantis tetapi stabil secara emosional sering diklasifikasikan benar

Failure Cases

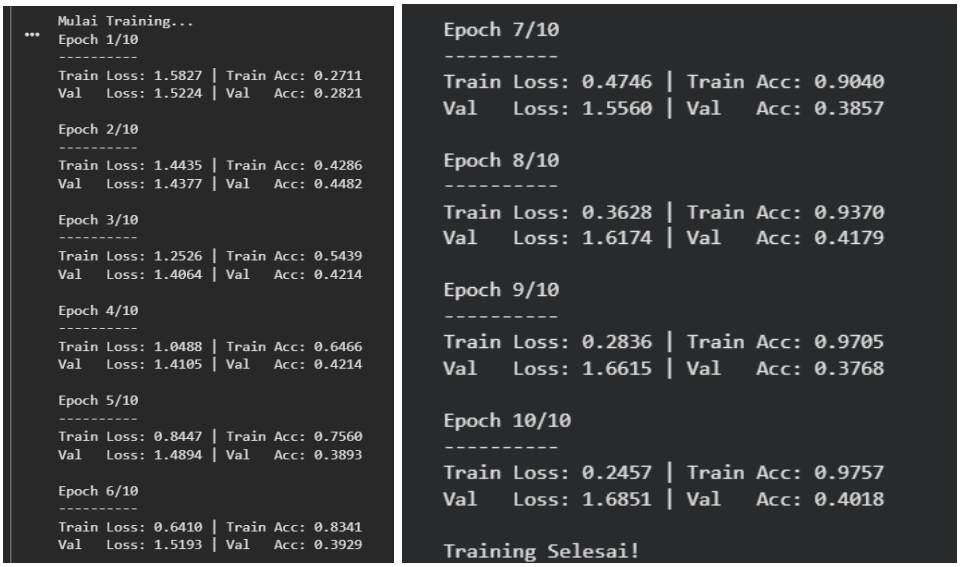
1. Mengandung ambiguitas emosional

   Misalnya lirik yang menceritakan hubungan sulit namun tetap optimis, model sering salah membaca polaritas emosi.
2. Bersifat naratif panjang

   Lirik panjang dengan banyak sub-konteks emosional sering membuat model kesulitan menentukan emosi dominan.
3. Menggunakan metafora berat atau simbolisme

   Kata-kata metaforis yang tidak eksplisit secara emosional mempersulit penarikan makna kontekstual.
4. Memakai kosakata "positif" tetapi bernuansa sedih

   Kata *love*, *hold*, atau *stay* dapat muncul di lirik sedih maupun ceria, hal ini sering menjadi penyebab confusion.



Gambar 4.8 Hasil *Training* lirik

4.2.4 Kesimpulan Awal Modalitas Lirik

1. Modalitas lirik memberikan performa paling tinggi di antara semua unimodal.

2. Struktur bahasa natural dan kosakata emosional memberikan sinyal kuat bagi model.
3. Tantangan utama muncul dari lirik yang ambiguitas maknanya, naratif kompleks, atau metaforis.
4. Potensinya untuk meningkatkan performa multimodal sangat besar karena melengkapi informasi akustik dan simbolik.

4.3 Modalitas MIDI

Modalitas MIDI memberikan representasi simbolik musik dalam bentuk rangkaian pitch yang mencerminkan struktur melodi dan pergerakan nada. Berbeda dari audio, MIDI tidak memuat informasi timbre atau intensitas akustik, tetapi tetap mengandung pola melodi yang dapat berhubungan dengan emosi. Dataset MIDI dalam penelitian ini merupakan modalitas dengan jumlah sampel paling sedikit, yaitu 193 sampel, sehingga baseline dilakukan dengan model ringan berbasis jaringan sekuensial.

4.3.1 Matrix Evaluasi dan Confusion Matrix

Model BiGRU yang digunakan pada modalitas MIDI menghasilkan performa yang paling rendah di antara ketiga modalitas.

1. Akurasi validasi berkisar 0.29–0.31
2. F1-score rata-rata rendah dan bervariasi antar kelas
3. Kesalahan klasifikasi dominan pada kelas dengan struktur melodi serupa

```
•••  Classification report:
                precision    recall  f1-score   support

     Cluster 1       0.00      0.00      0.00         5
     Cluster 2       0.29      0.40      0.33         5
     Cluster 3       0.12      0.20      0.15         5
     Cluster 4       0.00      0.00      0.00         3
     Cluster 5       0.00      0.00      0.00         2

      accuracy                           0.15        20
     macro avg       0.08      0.12      0.10        20
  weighted avg       0.10      0.15      0.12        20
```

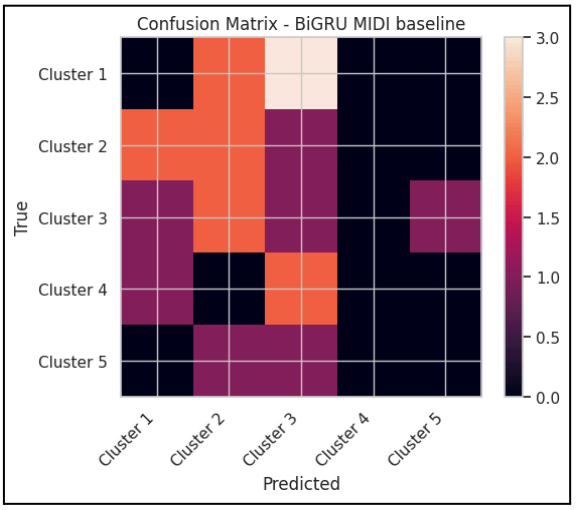
Gambar 4.9 Matrix Evaluasi MIDI

Confusion matrix memperlihatkan pola berikut:

1. Sebagian besar kelas memiliki tingkat *misclassification* tinggi
2. Klaster dengan kontur melodi datar (rentang pitch kecil) sering tertukar dengan klaster

bernuansa lembut atau melankolis

3. Klaster yang memiliki pola pitch lebih variatif cenderung lebih mudah dikenali, meskipun tetap tidak konsisten



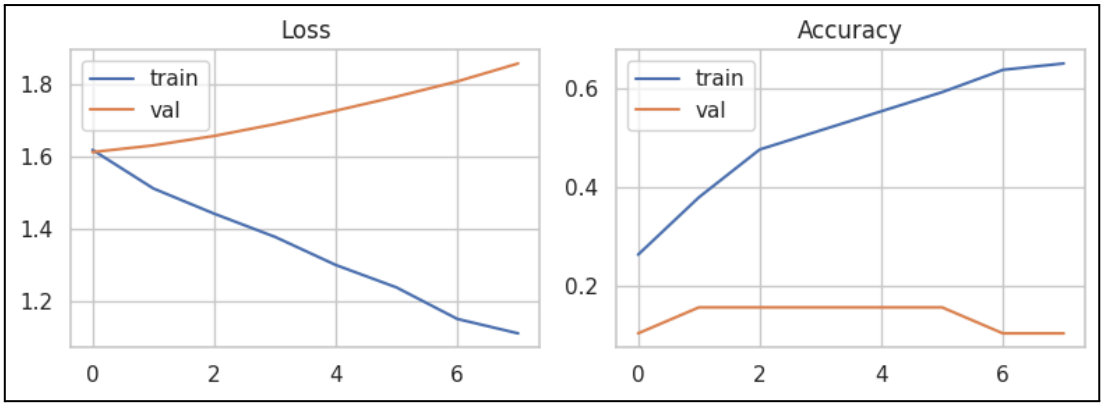
Gambar 4.10 Matrix Evaluasi MIDI

4.3.2 *Learning Curve* (*Loss* dan *Accuracy*)

Learning curve dari pelatihan menunjukkan pola yang khas untuk dataset kecil:

1. Training loss menurun secara stabil, menandakan model mampu mempelajari pola dasar pitch sequence
2. Validation loss stagnan lebih awal, kemudian berfluktuasi, menunjukkan bahwa model kesulitan menggeneralisasi pada data baru
3. Validation accuracy cenderung datar, dengan peningkatan sangat kecil setelah beberapa epoch pertama
4. Early stopping mencegah overfitting, karena setelah titik tertentu model hanya menghafal pola dari training set

Kurva ini menunjukkan bahwa keterbatasan jumlah data menjadi hambatan utama bagi model MIDI untuk membedakan klaster emosi.



Gambar 4.11 Grafik *Learning Curve (Loss & Metric)* MIDI

4.3.3 Analisis *Success* dan *Failure Case*

Success Cases

Beberapa sampel berhasil diklasifikasikan dengan benar karena memiliki pola pitch yang lebih khas, seperti:

1. Perubahan nada yang besar dan jelas sepanjang melodi
2. Progresi pitch yang konsisten memperlihatkan kontur emosi tertentu
3. Struktur ritmis yang tidak terlalu repetitif

Sampel dengan dinamika melodi yang jelas, misalnya naik-turun secara bertahap, lebih mudah dipetakan ke klaster emosional yang sesuai.

Failure Cases

Sebagian besar kesalahan model disebabkan oleh karakteristik berikut:

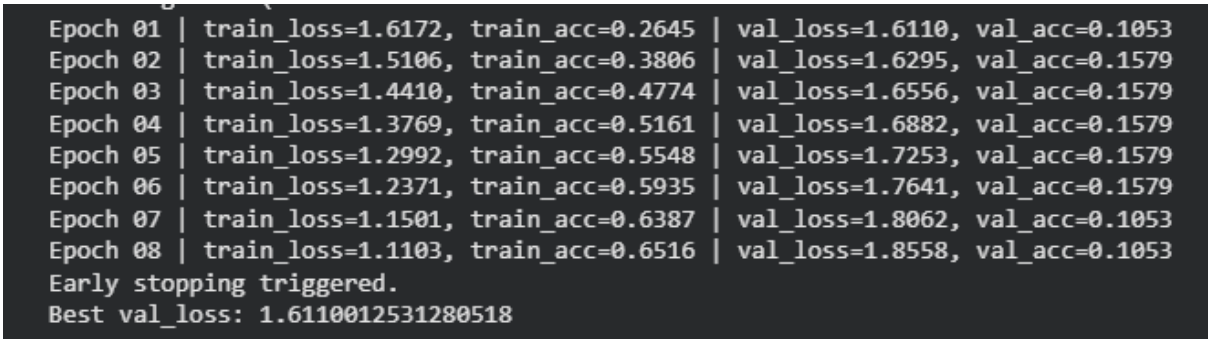
1. Pola pitch repetitif

   Nada yang berulang-ulang membuat representasi dalam ruang embedding hampir sama pada banyak lagu dari berbagai klaster.
2. Rentang pitch yang sempit

   Lagu dengan variasi pitch rendah tidak memberikan sinyal emosional yang kuat, sehingga model menebak secara acak.
3. Ukuran dataset yang kecil

   Jumlah sampel MIDI yang terbatas (193 saja) membuat distribusi pola melodi per kelas tidak cukup representatif.
4. Tidak adanya informasi dinamika atau timbre

   MIDI hanya menyimpan informasi nada, bukan ekspresi musikal seperti penekanan, vibrato, atau kekuatan pukulan yang sering menentukan emosi.



Gambar 4.12 Hasil *Training* MIDI

4.3.4 Kesimpulan Awal Modalitas MIDI

1. Performa MIDI adalah yang paling rendah dari ketiga modalitas.
2. Struktur pitch pada dataset ini tidak selalu mencerminkan emosi secara eksplisit.
3. Model BiGRU mampu mempelajari pola dasar, tetapi tidak cukup kuat untuk memisahkan klaster emosi yang memiliki struktur melodi serupa.
4. Perbaikan performa MIDI sangat mungkin dilakukan melalui:
   a. Penambahan fitur musikal lain (duration, velocity),
   b. Embedding yang lebih kaya,
   c. Arsitektur sekuensial yang lebih kompleks.

5. Rencana untuk Optimalisasi

Hasil baseline menunjukkan bahwa potensi peningkatan performa multimodal sangat bergantung pada perbaikan representasi unimodal, terutama pada modalitas MIDI yang saat ini masih menghasilkan akurasi paling rendah. Oleh karena itu, rencana optimalisasi berikut difokuskan pada memperkaya fitur musik, memperbaiki arsitektur model, dan menyempurnakan mekanisme fusi multimodal agar informasi dari berbagai modalitas dapat saling melengkapi secara lebih efektif.

Pada modalitas audio, peningkatan dilakukan melalui pendalaman arsitektur CRNN dan eksplorasi fitur akustik yang lebih kaya seperti MFCC atau chroma. Pendekatan ini diharapkan mampu menangkap karakter harmonik dan timbre yang tidak sepenuhnya tercermin pada mel-spectrogram saja. Teknik augmentasi seperti time-stretch atau pitch-shift berpotensi menyeimbangkan distribusi kelas yang tidak merata, sementara penyesuaian hyperparameter dan regularisasi tambahan akan diupayakan untuk meningkatkan stabilitas model.

Modalitas lirik yang sudah memberikan performa tertinggi tetap dapat dioptimalkan melalui fine-tuning bertahap pada BERT agar model lebih sensitif terhadap karakteristik bahasa dalam lirik musik. Eksperimen dengan model bahasa alternatif seperti RoBERTa atau DistilBERT akan dilakukan untuk melihat apakah ada konfigurasi yang dapat memberikan peningkatan performa dengan biaya komputasi yang lebih rendah. Selain itu, penambahan fitur linguistik sederhana, seperti pola n-gram atau skor sentimen yang dapat memperkaya sinyal emosional yang tidak tertangkap langsung oleh embedding.

Optimalisasi paling signifikan direncanakan pada modalitas MIDI. Representasi pitch-only terbukti terlalu sederhana untuk menangkap dinamika emosi dalam musik. Oleh karena itu, fitur MIDI akan diperluas mencakup durasi nada, velocity, serta inter-onset interval yang merepresentasikan ritme. Selain itu, pendekatan encoding berbasis kejadian (event-based encoding), seperti REMI atau MIDI-Like, akan dipertimbangkan untuk memodelkan struktur musik secara lebih fleksibel, sementara piano-roll multi-channel dapat digunakan untuk memetakan lapisan polifoni dengan lebih baik. Dari sisi model, BiGRU akan ditingkatkan ke versi yang lebih efektif dengan konfigurasi dua lapisan yang lebih ramping, hidden size yang lebih kecil, dropout yang lebih tinggi (sekitar 0.5), serta penambahan attention layer untuk memperkuat mekanisme penangkapan pola melodi. Langkah-langkah ini dipilih karena sesuai untuk dataset kecil yang tetap membutuhkan kemampuan representasi yang mendalam.

Pada tingkat multimodal, mekanisme fusi akan ditingkatkan dari pendekatan

sederhana berbasis skor menjadi strategi hybrid fusion yang menggabungkan representasi pertengahan dari masing-masing modalitas. Lapisan fusi tambahan akan digunakan agar model dapat belajar bobot kontribusi tiap modalitas dengan lebih baik. Normalisasi skala antar-modalitas juga akan diterapkan untuk memastikan bahwa tidak ada modalitas yang mendominasi hasil akhir karena perbedaan magnitudo nilai. Proses pelatihan akan diperkuat dengan penyesuaian learning rate scheduling, penggunaan class weighting, dan regularisasi tambahan agar model tetap stabil dalam menghadapi ketidakseimbangan label.

Secara keseluruhan, kombinasi perbaikan pada tingkat unimodal, arsitektur fusi, dan strategi pelatihan diharapkan mampu menghasilkan peningkatan signifikan pada performa multimodal. Target minggu depan adalah memperoleh versi pertama model multimodal dengan arsitektur fusion yang telah diperbarui, disertai evaluasi lengkap dan komparasi terhadap baseline yang ada.