202210370311072
Ahmad Naufal Luthfan Marzuqi
Prodi Informatika

BAB II

TINJAUAN PUSTAKA

2.1. Music Genre Classification

Music Genre Classification (MGC) merupakan salah satu pilar utama dalam bidang *Music Information Retrieval* (MIR) yang memiliki peran krusial dalam meningkatkan pengalaman pengguna melalui sistem rekomendasi yang lebih presisi serta pengorganisasian perpustakaan musik yang lebih terstruktur [15]. Seiring dengan perkembangan teknologi, metode MGC telah mengalami transformasi fundamental, mulai dari penggunaan teknik tradisional yang mengandalkan ekstraksi fitur manual seperti *Mel-Frequency Cepstral Coefficients* (MFCCs) dan *Support Vector Machines* (SVMs) [7], hingga adopsi arsitektur *deep learning* yang mampu mempelajari representasi fitur tingkat tinggi secara otomatis. Evolusi ini mencakup penggunaan *Convolutional Neural Networks* (CNNs) [5] dan *Recurrent Neural Networks* (RNNs) yang dirancang untuk mengatasi kompleksitas sinyal audio musik yang bersifat dinamis.

2.2. Long Short-Term Memory

Long Short-Term Memory (LSTM) merupakan varian khusus dari Recurrent Neural Networks (RNNs) yang dirancang untuk mengatasi kendala vanishing gradient dan sangat kompeten dalam mengenali pola temporal serta ketergantungan jangka panjang pada data sekuensial musik [16]. Dalam penelitian ini, model LSTM yang dioptimalkan tidak menggunakan audio mentah secara langsung, melainkan memanfaatkan transformasi Mel-spectrogram sebagai input utama. Pendekatan ini dipilih karena Mel-spectrogram mampu merepresentasikan karakteristik audio dalam domain frekuensi yang selaras dengan persepsi pendengaran manusia, sehingga memungkinkan LSTM untuk secara efektif menangkap pola ritme dan melodi yang unik pada setiap genre musik.

2.3. Audio Large Language Models

Kemunculan model fondasi audio berbasis self-supervised learning, atau yang secara luas dikenal sebagai Audio Large Language Models (LLMs), telah membuka paradigma baru dalam pengolahan audio dengan melatih model pada data wicara berskala besar guna menghasilkan representasi audio yang bersifat generalis [16]. Penelitian ini secara khusus mengevaluasi tiga model LLM dengan pendekatan pembelajaran yang berbeda untuk tugas MGC:

a. HuBERT: Menggunakan metode masked prediction untuk mempelajari struktur laten dari sinyal audio.
b. WavLM: Mengintegrasikan fitur denoising dan pelatihan yang bersifat speaker-aware untuk meningkatkan ketahanan representasi audio.
c. Wav2Vec 2.0: Mengandalkan arsitektur berbasis contrastive learning guna membangun representasi yang kuat langsung dari sinyal audio mentah.

2.4. Ekstraksi Fitur

Representasi data merupakan faktor penentu dalam performa klasifikasi genre musik. Penggunaan Mel-spectrogram pada model berbasis CNN dan LSTM telah terbukti menjadi standar yang efektif karena kemampuannya dalam memetakan frekuensi audio ke dalam skala Mel yang logaritmik, sehingga mempermudah model dalam mengidentifikasi komponen akustik yang relevan [5]. Di sisi lain, kemajuan pada model LLM memungkinkan proses fine-tuning dilakukan langsung menggunakan audio mentah (raw waveform). Pendekatan ini meminimalkan kebutuhan akan rekayasa fitur manual, meskipun memerlukan kapasitas komputasi yang jauh lebih besar dan memiliki tantangan tersendiri dalam menjaga stabilitas validasi agar tidak terjadi overfitting.

2.5. Dataset

Dataset yang digunakan dalam penelitian ini adalah GTZAN Dataset, yang merupakan salah satu dataset paling umum digunakan dalam penelitian Music

Information Retrieval (MIR), khususnya pada tugas klasifikasi genre musik. Dataset ini pertama kali diperkenalkan oleh Tzanetakis dan Cook pada tahun 2002 dan hingga saat ini masih menjadi tolok ukur standar dalam pengembangan sistem klasifikasi genre musik secara otomatis. Dalam penelitian ini, dataset GTZAN diperoleh melalui platform Kaggle yang dapat diakses secara publik [27]. GTZAN Dataset terdiri dari 1.000 klip audio dengan durasi masing-masing 30 detik, yang terbagi secara merata ke dalam 10 genre musik, yaitu *blues, classical, country, disco, hiphop, jazz, metal, pop, reggae,* dan *rock*. Setiap genre diwakili oleh 100 klip audio, sehingga distribusi kelas dalam dataset ini bersifat seimbang (*balanced*) dan dapat meminimalkan potensi bias genre dalam proses pelatihan maupun evaluasi model. File audio dalam dataset ini memiliki format WAV dengan *sample rate* 22.050 Hz dan *bit depth* 16-bit mono. Karakteristik tersebut menjadikan GTZAN Dataset kompatibel dengan berbagai pipeline pemrosesan audio, baik untuk ekstraksi fitur berbasis Mel-Spectrogram yang digunakan pada model LSTM maupun untuk pemrosesan *raw audio* yang digunakan pada model Audio Large Language Models (LLM) dalam penelitian ini.