BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Musik sudah menjadi bagian dari sejarah panjang manusia, secara tidak langsung musik sudah membimbing manusia kepada peradaban sekarang. Musik dapat mempengaruhi fungsionalitas manusia dan sebagai seni yang mampu mendorong emosional manusia dari sedih sampai bahagia. Hasil studi telah dilakukan sejak 1950 bahwa musik dapat mempengaruhi emosional dari pendengarnya [1]. Penilitian telah dilakukan bahwa musik dapat mempengaruhi kinerja dari otak ketika mendengar musik senang dan sedih, musik yang membangkitkan semangat ataupun ketakutan, pada bagian *amydgala* pada otak bereaksi terhadap musik yang didengar, hal ini memperkuat bahwa musik dapat memicu perubahan aktivitas di wilayah otak inti yang mendasari emosi [2] [3].

Music Emotion Recognition (MER) menjadi cabang penelitian berfokus pada interaksi sistem komputasi dalam mengenali dan memodelkan emosi yang terdapat dalam musik. Emosi dinilai sebagai elemen fundamental pada manusia ketika memilih lagu yang dapat disesuaikan, adaptasi, manipulasi, atau mengekspresikan kondisi psikologis dan *mood* [4]. Dari penelitian yang sudah dilakukan, MER bisa dibagi menjadi 2, yakni *categorical* dan *dimensional* [5]. *Categorical* mengklasifikan lagu ke dalam label emosi yang statis dan diskrit seperti, *sad, longing, lyric, jumping, joyful, warm, majestic, chanting*. MER *dimensional* memakai konsep 2D, emosi di koordinatkan dengan memakai sumbu X dan sumbu Y antara Valensi dan Arousal [5][6][7][8].

Model *categorical* dinilai terlalu kaku dalam mengklasifikasi musik, dikarenakan lagu hanya di batasi oleh suatu emosi, sedangkan lagu bersifat *versatile*, nada yang digunakan tidak selalu senang, pada bagian *intro* dapat menggunakan nada yang pelan atau tenang, dan di bagian *bridge* melodi lagu dapat berubah menjadi *powerful* [9]. Karakteristik tersebut yang menyebabkan banyak peneliti menggunakan model *dimensional*. Meskipun model dimensional dinilai lebih akurat untuk merepresentasikan emosi musik, model ini harus melakukan prediksi regresi *valence* dan *arousal* secara kontinu (frame-by-frame atau second-by-second) [10] menghadirkan tantangan teknis yang kompleks. Dari penelitian yang sudah dilakukan, pemetaan emosi pada sebuah musik sulit dilakukan

dikarenakan banyaknya variable yang tidak bisa dicantumkan seperti gagal dalam menangkap depedensi sekuensial jarak jauh, menggunakan ektraksi fitur tingkat rendah (*low-level features*) dengan terpisah sehingga kehilangan konteks dari spasial-temporal suara [11]. Pemetaan dari sinyal audio 1 dimensi (1D) mentah ke representasi emosi tingkat tinggi yang dinamis memerlukan fitur perantara akustik yang kaya dan pada saat yang sama arsitektur pembelajaran mendalam yang tangguh terhadap data deret waktu (*time series*).

Dalam musik ada beberapa hal krusial yang diperlu untuk diperhatikan seperti nada, melodi, dan tempo. Hal - hal tersebut menjadikan musik sebagai musik. Dalam penelitian yang sudah dilakukan sebelumnya, *Mel-Frequency Cepstral Coefficient* (MFCC) dan *Deep Convolutional Network* (DCNN) digunakan dalam melakukan analisis MER dan audio visual [12] [13]. *Mel-spectrogram* bekerja layaknya telinga manusia yang mengubah audio menjadi impuls listrik dan disalurkan kepada otak agar dapat dimengerti. *Spectrogram* mengubah suara menjadi *time-frequency* dan memasukkannya ke dalam skala *linear-frequency*. Sebaliknya, *Mel-Spectrogram* menggunakan *quasi-logarithmic spacing* yang sangat mereferensi pada sistem pendengaran manusia [14].

Penelitian musik tidak hanya sebatas pada fitur audio, terdapat peran lebih dalam musik yakni emosi yang dapat tersalurkan melalui melodi melodi tersebut. Data yang diekstraksi menggunakan *arousal* dan *valence* yang dikoordinatkan ke dalam dimensional. Penelitian terdahulu memberikan hasil yang cukup bagus dalam memprediksi emosi yang ada dalam musik menggunakan CNN dan CRNN, penelitian [15] menghasilkan RMSE 0.380, MAE 0.068, dan R^2 74.39% dengan menggunakan CNN saja. Dan menghasilkan RMSE 0.043 dengan MAE 0.033 dan R^2 93.2% dengan menggunakan CNN + CRNN (Bi-LSTM). Fitur-fitur tingkat rendah dari audio dimasukkan ke dalam jaringan bidirectional long short-term memory (Bi-LSTM) untuk mendapatkan informasi urutan fitur audio lebih lanjut. Penelitian [16] menunjukkan penggunaan CNN dan Bi-GRU dalam mengevaluasi emosi musik, penelitian tersebut menunjukkan model CNN + Bi-GRU hasil MSE 0.0232, MAE 0.098 dan R^2 0.556 dan dibandingkan dengan hanya CNN yang menghasilkan MSE 0.0269, MAE 0.111, dan R^2 0.521. Berdasarkan penelitian sebelumnya yang sudah dilakukan, membuktikan bahwa CRNN mampu menghasilkan hasil yang lebih baik dibandingkan CNN saja pada pemodelan berbasis MER. CNN sangat baik dalam mengekstraksi dan merepresentasikan informasi spasial dari matriks fitur audio, model ini kehilangan informasi temporal dari urutan waktu sinyal audio. CRNN digunakan sebagai

solusi untuk mengatasi kekurangan CNN tersebut, CRNN mampu mengekstraksi informasi spasial sekaligus merepresentasikan informasi temporal secara efektif [17].

Penelitian terdahulu mayoritas terfokus pada perhitungan jarak galat murni seperti RMSE, MSE, dan MAE serta R^2 sebagai variansi data. Pada konteks pemodelan emosi musik yang bersifat kontinu, metrik jarak tersebut memiliki kelemahan fundamental karena tidak mampu merepresentasikan apakah fluktuasi prediksi model benar-benar selaras dan seirama dengan trajektori perubahan emosi aktual seiring berjalannya waktu. Dan masih terbatas kajian yang mengeksplorasi penggabungan fitur spasial *mel-spectrogram* dengan fitur harmoni *chroma* untuk mendeteksi emosi musik yang ada pada lagu.

Berdasarkan celah penelitian tersebut, penelitian ini mengkaji secara komprehensif perbedaan kemampuan CNN dan CRNN (Bi-GRU) dalam menangkap informasi spasial-temporal menggunakan penggabungan fitur *mel-spectrogram* dan *chroma*. Untuk memastikan bahwa model tidak hanya meminimalisir galat tetapi juga meniru dinamika emosi secara presisi, penelitian ini menetapkan *Concordance Correlation Coefficient* (CCC) sebagai metrik evaluasi utama pada korpus musik yang berjalan secara kontinu.

CNN dan CRNN telah membuktikan bahwa model tersebut menunjukan kinerja yang menjanjikan dalam melakukan analisis MER, namun masih terbatas kajian penelitian yang menggunakan ekstraksi fitur *mel-spectrogram* dan *chroma* berbasis regresi *valence* dan *arousal* dalam melakukan perbandingan analisis terhadap kedua model tersebut dengan metrik evaluasi yang seragam dan pada korpus musik yang sama. Dengan demikian, masih dibutuhkan penelitian yang mengkaji secara komprehensif perbedaan kemampuan CNN dan CRNN Bi-GRU dalam menangkap informasi spasial dan temporal untuk pemodelan emosi musik secara kontinu.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini dapat dijabarkan sebagai berikut:

1. Bagaimana implementasi arsitektur CNN dan CRNN (menggunakan *Bi-Directional GRU*) dalam memodelkan regresi nilai *valence* dan *arousal* dengan ekstraksi fitur *mel-spectrogram* dan *chroma*?

2. Bagaimana perbandingan kinerja antara model CNN yang bersifat statis dengan model CRNN yang memiliki pemodelan temporal dalam memprediksi dinamika emosi musik?

3. Seberapa baik tingkat akurasi dan kesesuaian arsitektur yang diusulkan ketika dievaluasi menggunakan metrik regresi kontinu?

1.3 Batasan Permasalahan

Untuk menjaga agar ruang lingkup penelitian tetap fokus dan terarah, ditetapkan beberapa batasan masalah sebagai berikut:

1. Data audio yang digunakan terbatas pada dataset DEAM, yang merepresentasikan emosi musik dalam durasi potongan 45 detik per lagu secara sekuensial.

2. Fitur audio yang diekstraksi dan dijadikan masukan (*input*) bagi model dibatasi pada *mel-spectrogram* (untuk memodelkan energi/ritme) dan *chroma* (untuk memodelkan harmoni/nada), tanpa melibatkan ekstraksi fitur tekstual seperti lirik lagu.

3. Arsitektur *Deep Learning* yang dievaluasi difokuskan pada perbandingan antara *Convolutional Neural Network* (CNN) standar dan *Convolutional Recurrent Neural Network* (CRNN) dengan varian *Bi-Directional Gated Recurrent Unit* (Bi-GRU).

4. Nilai anotasi emosi (*valence* dan *arousal*) yang terdapat pada dataset DEAM diasumsikan sebagai standar mutlak (*ground truth*), sehingga bias subjektivitas dari *volunteer* penganotasi data berada di luar kendali dan fokus penelitian ini.

1.4 Tujuan Penelitian

Sejalan dengan rumusan masalah di atas, tujuan yang hendak dicapai dari penelitian ini adalah:

1. Mengimplementasikan arsitektur CNN dan CRNN (Bi-GRU) berbasis ekstraksi fitur *mel-spectrogram* dan *chroma* untuk memprediksi emosi musik menggunakan intensitas *valence* dan *arousal*.

2. Menganalisis dan membandingkan efektivitas algoritma CNN dengan CRNN dalam menangkap informasi spasial dan temporal dari dinamika emosi musik.
3. Mengevaluasi tingkat akurasi prediksi model regresi yang dibangun menggunakan metrik regressi seperti MAE, MSE, RMSE, R^2, dan CCC.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari pelaksanaan penelitian ini adalah sebagai berikut:

1. Memberikan kontribusi pada perkembangan literatur *Music Emotion Recognition* (MER) dengan menyediakan analisis komprehensif mengenai efektivitas kombinasi fitur *mel-spectrogram* dan *chroma* pada model regresi CRNN asimetris.
2. Model yang dihasilkan dapat diimplementasikan sebagai *feature extractor* untuk melakukan anotasi emosi secara otomatis, yang sangat krusial dalam mendukung pengembangan sistem rekomendasi musik berbasis afektif/emosional di dunia industri.
3. Menjadi referensi dan landasan komparatif bagi penelitian lanjutan di bidang pengolahan sinyal audio dan *Deep Learning* sekuensial.

1.6 Sistematika Penulisan

Berikut merupakan sistematika penulisan dari pengerjaan penelitian ini:

- Bab 1 PENDAHULUAN
  Latar belakang berisi penjelasan tentang emosi musik dan pengaruh terhadap emosi manusia, pengenalan *Music Emotion Recognition* (MER), pendekatan yang ada dalam MER (*categorical* dan *dimensional*, serta perbandingan dari algoritma CNN dan CRNN dalam memprediksi emosi musik. Rumusan masalah, batasan, tujuan dan manfaat penelitian juga di jelaskan dalam bab ini.
- Bab 2 LANDASAN TEORI
  Landasan teori membahas dasar teori pada MER, dimensi *valence-arousal*, representasi *mel-spectrogram*, *chroma*, arsitektur CNN dan CRNN, dan metrik evaluasi regresi.

- Bab 3 METODOLOGI PENELITIAN

  Metode penilitian menjelaskan tentang metode yang dilakukan untuk analisis emosi musik dengan menggunakan perbandingan CNN dan CRNN, proses ekstrasi fitur *mel-spectrogra*m dan *chroma*, dataset DEAM yang digunakan, arsitektur model CNN dan CRNN Bi-GRU yang dipakai, proses pelatihan dan validasi model.

- Bab 4 HASIL DAN DISKUSI

  Hasil dan diskusi dijabarkan dalam bab 4. Melampirkan hasil dari proses training daripada model CNN dan CRNN dalam memprediksi emosi musik. Dalam bab ini hasil dituliskan secara rinci antara CNN dan CRNN, termasuk evaluasi dari hasil training.

- Bab 5 KESIMPULAN DAN SARAN

  Kesimpulan didapatkan dari hasil dan analisis penelitian yang sudah dilakukan antara CNN dan CRNN dalam memprediksi emosi musik, serta penemuan terhadap variable yang mempengaruhi evaluasi model. Saran digunakan untuk membantu perkembangan lanjutan terhadap penelitian yang sudah dilakukan.