IMPLEMENTASI MODEL *CONVOLUTIONAL RECURRENT NEURAL NETWORK* (CRNN) DENGAN *BIDIRECTIONAL LONG SHORT-TERM MEMORY* (BI-LSTM) UNTUK PENGENALAN EMOSI PADA SUARA

(Skripsi)

Oleh:

M. RADITYA ADHIRAJASA

NPM 2157051004

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRAK

IMPLEMENTASI MODEL *CONVOLUTIONAL RECURRENT NEURAL NETWORK* (CRNN) DENGAN *BIDIRECTIONAL LONG SHORT-TERM MEMORY* (BI-LSTM) UNTUK PENGENALAN EMOSI PADA SUARA

Oleh

M. RADITYA ADHIRAJASA

Pengenalan emosi pada suara merupakan tugas kompleks yang krusial untuk interaksi manusia-mesin. Penelitian ini mengimplementasikan dan mengevaluasi model *Convolutional Recurrent Neural Network* (CRNN) yang menggabungkan keunggulan *Convolutional Neural Network* (CNN) untuk ekstraksi fitur spasial dengan *Bidirectional Long Short-Term Memory* (BiLSTM) untuk analisis dependensi temporal. Model ini dilatih dan diuji menggunakan dataset CREMA-D, yang mencakup enam kelas emosi: marah, jijik, takut, bahagia, netral, dan sedih. Fitur diekstraksi dari sinyal audio menggunakan *log-Mel spectrogram* untuk merepresentasikan karakteristik suara yang relevan dengan persepsi manusia. Untuk meningkatkan variasi data dan generalisasi model, diterapkan teknik augmentasi data berupa *pitch shifting* dan *time stretching*. Hasil evaluasi menunjukkan bahwa model CRNN yang didukung dengan augmentasi data berhasil mencapai akurasi validasi sebesar 68.23%, sebuah peningkatan dari 66.08% yang dicapai oleh model tanpa augmentasi. Model ini menunjukkan performa yang menjanjikan, terutama dalam mengklasifikasikan emosi marah dan netral. Penelitian ini menyimpulkan bahwa arsitektur CRNN dengan BiLSTM efektif untuk tugas pengenalan emosi suara, dan augmentasi data berperan penting dalam meningkatkan kinerjanya.

Kata Kunci: Pengenalan Emosi Suara, CRNN, BiLSTM, *Log-Mel Spectrogram,* Augmentasi Data.

ABSTRACT

IMPLEMENTATION OF A CONVOLUTIONAL RECURRENT NEURAL NETWORK (CRNN) WITH BIDIRECTIONAL LONG SHORT-TERM MEMORY (BI-LSTM) FOR SPEECH EMOTION RECOGNITION

By

M. RADITYA ADHIRAJASA

Speech emotion recognition is a complex task crucial for human-machine interaction. This study implements and evaluates a Convolutional Recurrent Neural Network (CRNN) model that combines the advantages of a Convolutional Neural Network (CNN) for spatial feature extraction with Bidirectional Long Short-Term Memory (BiLSTM) for temporal dependency analysis. The model is trained and tested using the CREMA-D dataset, which includes six emotion classes: anger, disgust, fear, happiness, neutral, and sadness. Features are extracted from the audio signals using log-Mel spectrograms to represent sound characteristics relevant to human perception. To enhance data variation and model generalization, data augmentation techniques such as pitch shifting and time stretching were implemented. The evaluation results show that the CRNN model with data augmentation achieved a validation accuracy of 68.23%, an increase from the 66.08% achieved by the model without augmentation. The model shows promising performance, especially in classifying the anger and neutral emotions. This study concludes that the CRNN architecture with BiLSTM is effective for the task of speech emotion recognition, and that data augmentation plays a crucial role in enhancing its performance.

Keywords: *Speech Emotion Recognition, CRNN, BiLSTM, Log-Mel Spectrogram, Data Augmentation.*

IMPLEMENTASI MODEL *CONVOLUTIONAL RECURRENT NEURAL NETWORK* (CRNN) DENGAN *BIDIRECTIONAL LONG SHORT-TERM MEMORY* (BI-LSTM) UNTUK PENGENALAN EMOSI PADA SUARA

Oleh

M. RADITYA ADHIRAJASA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA KOMPUTER

Pada

Program Studi Ilmu Komputer
Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

Judul Skripsi	:	**IMPLEMENTASI MODEL CONVOLUTIONAL RECURRENT NEURAL NETWORK (CRNN) DENGAN BIDIRECTIONAL LONG SHORT-TERM MEMORY (BI-LSTM) UNTUK PENGENALAN EMOSI PADA SUARA**
Nama Mahasiswa	:	M. Raditya Adhirajasa
Nomor Pokok Mahasiswa	:	2157051004
Program Studi	:	S1 Ilmu Komputer
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Favorisen R. Lumbanraja, Ph.D.	**Yunda Heningtyas, M. Kom.**
NIP. 19830110 200812 1 002	NIP. 19890108 201903 2 014

2. Mengetahui

Ketua Jurusan Ilmu Komputer	Ketua Program Studi S1 Ilmu Komputer
Dwi Sakethi, S.Si., M.Kom.	**Tristiyanto, M.I.S., Ph.D.**
NIP. 19680611 199802 1 001	NIP. 19810414 200501 1 001

MENGESAHKAN

1. Tim Penguji

Ketua : **Favorisen R. Lumbanraja, Ph.D.**

Sekretaris : **Yunda Heningtyas, M. Kom.**

Penguji

Bukan Pembimbing : **Dr. rer. nat. Akmal Junaidi, S.Si., M.Sc.**

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.

NIP. 19711001 200501 1 002

Tanggal Lulus Ujian Skripsi: **20 Juni 2025**

PERNYATAAN

Saya bertanda tangan di bawah ini:

Nama : M. Raditya Adhirajasa

NPM : 2157051004

Dengan ini menyatakan bahwa skripsi saya yang berjudul "**Implementasi Model *Convolutional Recurrent Neural Network* (CRNN) Dengan *Bidirectional Long Short-Term Memory* (Bi-LSTM) Untuk Pengenalan Emosi Pada Suara**" merupakan hasil karya saya sendiri dan bukan karya orang lain. Seluruh tulisan yang tertuang di skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti skripsi saya merupakan hasil penjiplakan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang telah saya terima.

Bandar Lampung, 22 Agustus 2025

M. Raditya Adhirajasa

NPM. 2157051004

RIWAYAT HIDUP

Penulis bernama lengkap Muhammad Raditya Adhirajasa, dilahirkan pada 21 Maret 2003 di Kota Bandar Lampung, Provinsi Lampung, sebagai anak pertama daripasangan Bapak Akhmad Hafifi dan Ibu Suri Novita Yanti.

Penulis menempuh Pendidikan di Taman Kanak-Kanak (TK) Az-zahra pada tahun 2008 dan melanjutkan pendidikan di SD Negeri 2 Palapa Bandar Lampung pada Tahun 2009-2015. Kemudian penulis melanjutkan pendidikan menengah pertama di SMP Negeri 16 Bandar Lampung pada Tahun 2015-2018, dan pendidikan menengah atas di SMA Negeri 2 Bandar Lampung di Tahun 2018-2021.

Pada Tahun 2021 penulis terdaftar sebagai mahasiswa Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SMMPTN. Selama menjadi mahasiswa penulis melakukan beberapa kegiatan antara lain:

1. Menjadi Asisten Dosen Jurusan Ilmu Komputer Mata Kuliah Kecerdasan Buatan pada Tahun Ajaran 2023/2024.
2. Menjadi Asisten Dosen Jurusan Ilmu Komputer Mata Kuliah Pemrograman Deklaratif pada Tahun Ajaran 2023/2024.
3. Melaksanakan Kuliah Kerja Nyata di Desa Rejomulyo, Kecamatan Jati Agung, Kabupaten Lampung Selatan pada bulan Januari 2024 .
4. Melaksanakan Kerja Praktik di Badan Pusat Statistik Republik Indonesia pada bulan Agustus 2024.

SANWACANA

Segala puji dan syukur penulis panjatkan ke hadirat Allah Subhanahu Wa Ta'ala atas segala limpahan nikmat, rahmat, dan hidayah-Nya sehingga penulisan skripsi ini dapat terselesaikan dengan baik dan lancar. Shalawat serta salam semoga senantiasa tercurah kepada Nabi besar Muhammad Shallallahu 'Alaihi Wasallam, pembawa risalah kebenaran bagi umat manusia. Skripsi berjudul "**Implementasi Model *Convolutional Recurrent Neural Network* (CRNN) Dengan *Bidirectional Long Short-Term Memory* (Bi-LSTM) Untuk Pengenalan Emosi Pada Suara**" telah disusun dengan sebaik-baiknya dan sebagai salah satu syarat untuk mendapatkan gelar sarjana Ilmu Komputer di Universitas Lampung.

Penulis menyadari bahwa dalam proses penyusunan skripsi ini tidak terlepas dari dukungan, bimbingan, serta bantuan dari berbagai pihak. Oleh karena itu penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Dekan FMIPA Universitas Lampung.
2. Bapak Dwi Sakethi, S.Si., M. Kom. selaku Ketua Jurusan Ilmu Komputer FMIPA Universitas Lampung.
3. Ibu Yunda Heningtyas, M. Kom. selaku Sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung.
4. Bapak Tristiyanto, S. Kom., M.I.S., Ph.D. selaku Ketua Program Studi Ilmu Komputer FMIPA Universitas Lampung.
5. Ibu Ossy Dwi Endah Wulansari, S.Si, M.T. selaku Pembimbing Akademik penulis yang senantiasa memberikan dukungan dan arahan dalam pengembangan akademik selama masa studi.

6. Bapak Favorisen Rosyking Lumbanraja, Ph.D. selaku dosen Pembimbing Utama, yang telah dengan sabar dan penuh dedikasi membimbing penulis selama proses penyusunan skripsi ini. Terima kasih atas waktu, ilmu, arahan serta koreksi yang sangat berarti dalam menyempurnakan karya ini.
7. Ibu Yunda Heningtyas, M. Kom. selaku dosen Pembimbing Kedua, yang telah memberikan arahan, masukan dan motivasi yang sangat membantu dalam proses penyusunan skripsi ini. Terima kasih atas waktu, ilmu, arahan, serta koreksi yang sangat berarti dalam menyempurnakan karya ini.
8. Bapak Dr. rer. nat. Akmal Junaidi, S.Si., M.Sc. selaku dosen Pembahas, yang telah memberikan kritik, saran dan masukan konstruktif yang sangat berarti bagi penyempurnaan skripsi ini. Terima kasih atas perhatian dan kontribusi bapak dalam proses akademik penulis.
9. Bapak dan Ibu Dosen Jurusan Ilmu Komputer FMIPA Universitas Lampung, yang telah membagikan ilmu, motivasi serta pengalaman berharga selama penulis menempuh pendidikan di Jurusan Ilmu Komputer Universitas Lampung.
10. Seluruh staf dan karyawan Jurusan Ilmu Komputer, Ibu Ade Nora Maela, Bang Zainuddin dan Mas Syam yang telah membantu kelancaran berbagai urusan administrasi, laboratorium serta peminjaman ruangan selama masa perkuliahan penulis di Jurusan Ilmu Komputer.

Penulis menyadari bahwa skripsi ini masih memiliki berbagai kekurangan dan belum mencapai kesempurnaan. Namun demikian, penulis berharap karya ini dapat memberikan manfaat khususnya bagi civitas akademika Universitas Lampung, serta menjadi referensi dan kontribusi positif bagi mahasiswa Ilmu Komputer.

Bandar Lampung, 22 Agustus 2025

M. Raditya Adhirajasa
NPM. 2157051004

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR KODE PROGRAM

Kode Program Halaman

I. PENDAHULUAN

1.1. Latar Belakang

Kondisi emosi manusia memainkan peran penting dalam interaksi dan perilaku antar manusia. Emosi menunjukkan pengaruh yang cukup besar pada komunikasi verbal dan visual, seperti ekspresi wajah dan karakter suara. Suara adalah salah satu cara yang efektif untuk mengekspresikan emosi (Qamhan *et al*., 2020). Emosi merupakan sifat alami manusia. Untuk memahami emosi manusia, sistem kecerdasan buatan yang didasarkan pada pengenalan emosi semakin banyak bermunculan (Gokilavani, 2023).

Pengenalan emosi pada suara merupakan kemampuan komputer untuk mengenali dan memahami suara manusia kemudian mengubahnya menjadi teks atau perintah yang akan dieksekusi. Teknologi ini dapat digunakan di berbagai bidang seperti mengembangkan sistem interaksi manusia-komputer yang lebih manusiawi, pengenalan suara yang lebih baik, dan pada bidang psikologi untuk mempelajari hubungan antara suara dan emosi (Mukarram *et al*., 2024). Namun, terdapat beberapa tantangan dalam pengenalan emosi pada suara karena emosi sangat bergantung pada bahasa, gaya bicara, latar belakang budaya, konteks, dan kondisi lingkungan (Yadav *et al*., 2021).

Seiring dengan perkembangan kecerdasan buatan, model *deep learning* telah terbukti unggul dalam pengenalan emosi pada suara. Salah satu arsitektur yang populer adalah *Convolutional Recurrent Neural Network* (CRNN), yang menggabungkan keunggulan *Convolutional Neural Network* (CNN) dan *Recurrent*

Neural Network (RNN). Arsitektur CNN adalah kumpulan jaringan saraf yang tersusun dengan berbagai macam lapisan dalam urutan tertentu dan setiap lapisan memberikan kontribusi tertentu (Qamhan *et al*., 2020). CNN secara otomatis mengekstrak fitur akustik dari spektrogram ucapan (Mukarram *et al*., 2024). Khususnya *Long Short-Term Memory* (LSTM) yang merupakan pengembangan dari algoritma RNN, digunakan untuk menangkap ketergantungan temporal pada data berurutan. Kombinasi ini memungkinkan CRNN untuk secara efektif mengenali pola-pola kompleks pada sinyal suara yang mengandung informasi emosional. Kombinasi antara CNN dengan *Bidirectional Long Short-Term Memory* (BiLSTM) yang merupakan pengembangan dari LSTM, memperkuat model CRNN dengan memungkinkan analisis data suara dalam dua arah untuk mengatasi kesulitan dalam melatih model RNN seperti *vanishing gradient* dan *exploding gradient* (Pham *et al.*, 2023). BiLSTM menerapkan dua LSTM terpisah untuk membaca informasi dari dua arah, satu untuk arah maju dan satu untuk arah mundur (Nugroho *et al*., 2021).

Penelitian Zielonka *et al.* (2022) meningkatkan kinerja pengenalan emosi pada suara melalui penggunaan *Convolutional Neural Networks* (CNN), dengan fokus pada evaluasi metode ekstraksi fitur berupa *spectrogram* dan *mel-spectrogram*. Penelitian ini mengeksplorasi perbedaan dalam akurasi antara kedua metode tersebut saat digunakan untuk melatih CNN dalam mengenali emosi dari ucapan. Penelitian ini menggunakan beberapa dataset untuk menguji kemampuan model pengenalan emosi dalam berbagai lingkungan dan aktor.

Penelitian lainnya, seperti yang dilakukan Yadav *et al*. (2021), menguji model CRNN dengan ektraksi fitur *log-mel spectograms* pada dataset *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) yang merupakan dataset berbahasa inggris dan *Berlin Database of Emotional Speech* (EMO-DB) yang merupakan dataset berbahasa jerman. Penelitian tersebut menunjukkan peningkatan kinerja model setelah penerapan teknik augmentasi data seperti penambahan *noise* dan manipulasi *pitch*. Metode ekstraksi fitur *log-mel spectogram* yang diusulkan cocok untuk menangkap informasi emosional pada suara untuk

model CRNN, namun proses ekstraksi *log-mel spectrogram* langkah komputasi yang rumit, sehingga membuat prosesnya menjadi kompleks.

Penelitian ini akan menggunakan model CRNN dengan menggabungkan CNN dan BiLSTM untuk mengklasifikasi emosi manusia berdasarkan suara. Dengan menggabungkan kedua arsitektur tersebut, model dapat mengekstrak fitur lokal dan memahami hubungan konteks dalam urutan waktu dari suara (Yadav *et al*., 2021). Selain itu, penelitian ini akan menerapkan teknik augmentasi data, seperti *pitch shifting* dan *time stretching* untuk meningkatkan variasi dalam pelatihan. Penggunaan *log-mel spectrogram* memungkinkan model menangkap fitur akustik yang lebih kaya dan representatif terhadap persepsi manusia. Kombinasi keduanya diharapkan dapat meningkatkan akurasi pengenalan emosi serta membuat model lebih adaptif dalam situasi dunia nyata.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang dikemukakan, rumusan masalah pada penelitian ini adalah implementasi model CRNN yang menggabungkan CNN dan BiLSTM untuk meningkatkan akurasi dalam pengenalan emosi pada suara dan melakukan evaluasi kinerja model dalam mengklasifikasi emosi pada suara.

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini adalah:

a. Dataset yang digunakan yaitu CREMA-D diperoleh dari kaggle dan terbatas pada audio berbahasa inggris.
b. Dataset memiliki 6 jenis emosi yaitu, marah, bahagia, sedih, takut, netral, dan jijik.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

a. Mengimplementasi model CRNN dengan BiLSTM dalam pengenalan emosi pada suara menggunakan dataset CREMA-D.
b. Mengevaluasi dan menganalisis efektivitas model CRNN dengan BiLSTM dalam mengidentifikasi emosi pada suara, untuk menentukan seberapa baik model ini dapat meningkatkan akurasi pengenalan emosi pada suara.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

a. Menambah pengetahuan tentang cara kerja model CRNN dengan BiLSTM dalam mengidentifikasi emosi manusia pada suara.
b. Memberikan kontribusi pada pengembangan teknologi pengenalan emosi berbasis suara, yang dapat diaplikasikan pada berbagai bidang seperti kesehatan, pendidikan, dan interaksi manusia-mesin.

II. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian yang pernah dilakukan dan berkaitan dengan implementasi model *convolutional recurrent neural network* (CRNN) dengan *bidirectional long short-term memory* (BiLSTM) untuk pengenalan emosi pada suara dapat dilihat pada Tabel 1.

Tabel 1. Penelitian Terdahulu Terkait Pengenalan Emosi pada Suara

No	Penulis	Data	Model	Hasil
1	*Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets* (Zielonka *et al*., 2022)	**CREMA-D** Jumlah: 7442 Marah: 1271 Jijik: 1271 Takut: 1271 Senang: 1271 Netral: 1087 Sedih: 1271 **IEMOCAP** Jumlah: 5531 Marah: 1103 Senang: 1636 Netral: 1708 Sedih: 1084	Model: *Convolutional Neural Network* (CNN) Ekstraksi Fitur: *Spectrogram, Mel-Spectogram* Evaluasi: *Confusion Matrix*	Akurasi IEMOCAP: 69.13% Akurasi CREMA-D: 53.66%
2	*Speech Emotion Recognition using Convolutional Recurrent Neural Network* (Yadav *et al.*, 2021)	**RAVDESS** Jumlah: 1440 Marah: 192 Tenang: 192 Jijik: 192 Takut: 192	Model: *Convolutional Recurrent Neural Network* (CRNN)	Akurasi RAVDESS: 85.76 %

Tabel 1. Lanjutan

No	Penulis	Data	Model	Hasil
		Senang: 192 Netral: 96 Sedih: 192 Kaget: 192 **Berlin EMO-DB** Jumlah: 535 Marah: 127 Bosan: 81 Jijik: 46 Takut: 69 Senang: 71 Netral: 79 Sedih: 62	Ekstraksi Fitur: *Log-Mel Spectograms* Evaluasi: *Confusion Matrix*	Akurasi Berlin EMO-DB: 91.59%
3	*Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition* (Pham *et al.*, 2023)	**Berlin EMO-DB** Jumlah: 535 Marah: 127 Bosan: 81 Jijik: 46 Takut: 69 Senang: 71 Netral: 79 Sedih: 62 **CREMA-D** Jumlah: 5229 Marah: 893 Jijik: 893 Takut: 893 Senang: 893 Netral: 764 Sedih: 893	Model: *Modified Attention-Based Dilated Convolutional Recurrent Neural Network* (mADCRNN) Ekstraksi Fitur: 3D *Log-Mel Scaled Spectograms* Evaluasi: *Confusion Matrix*	Akurasi Berlin EMO-DB: 88,03% Akurasi CREMA-D: 66,56%

Penelitian yang dilakukan Zielonka *et al*. (2022) menunjukkan bahwa *mel-spectrogram lebih* unggul dibandingkan dengan *spectrogram* dalam pengenalan emosi dari ucapan, dengan akurasi yang lebih baik di berbagai dataset dan konfigurasi. Penelitian ini menekankan pentingnya pembagian data yang tepat antara pelatihan dan pengujian. Pembagian yang salah menghasilkan akurasi yang jauh lebih tinggi karena model mengenali fitur individu dari aktor tersebut, bukan

kemampuan untuk menggeneralisasi ke aktor baru. Pembagian data yang benar menghasilkan hasil yang lebih realistis.

Penelitian terkait implementasi CRNN untuk pengenalan emosi pada suara yang dilakukan Yadav *et al.* (2021) menggabungkan CNN dan BiLSTM. Dataset yang digunakan yaitu Berlin EMO-DB yang diambil dari 10 aktor di Berlin, Jerman (Burkhardt *et al.,* 2005) dan RAVDESS. Penelitian ini memanfaatkan ekstraksi fitur *log-mel spectrogram* yang merupakan representasi spektral dua dimensi dari sinyal suara untuk mengidentifikasi emosi manusia pada suara. Penelitian ini menerapkan augmentasi data, seperti *noise injection*, *time stretching*, dan *pitch shifting*, yang terbukti dapat meningkatkan akurasi model. Hasil penelitian juga menunjukkan bahwa *Log-mel spectrogram*, yang digunakan sebagai ekstraksi fitur memberikan hasil yang baik, meskipun proses ekstraksinya membutuhkan daya komputasi tinggi.

Penelitian selanjutnya terkait implementasi CRNN untuk pengenalan emosi pada suara yang dilakukan Pham *et al.* (2023) mengembangkan metode mADCRNN dan gabungan teknik augmentasi. Dataset yang digunakan yaitu CREMA-D dan Berlin EMO-DB. Penelitian ini memperkenalkan metode *hybrid data augmentation* (HDA) yang menggabungkan augmentasi tradisional seperti *time shifting* dan *pitch shifting* dengan teknik augmentasi berbasis *Generative Adversarial Networks with Waveform* (WaveGAN) untuk menghasilkan sampel data tambahan yang bervariasi dalam pengenalan emosi suara. Pendekatan ini bertujuan untuk meningkatkan kinerja model, terutama ketika menghadapi dataset yang terbatas dan tidak seimbang. Ekstraksi fitur dilakukan menggunakan *3D Log Mel Spectrogram* yang efektif menangkap emosi pada sinyal suara yang telah diolah. Hasil dari penelitian ini membuktikan WaveGAN menjadi teknik augmentasi yang efektif untuk mengatasi keterbatasan data dan ekstraksi fitur *3D Log Mel Spectrogram* turut berkontribusi pada peningkatan kinerja model.

2.2. Emosi

Emosi manusia adalah respons psikologis yang kompleks yang dipengaruhi oleh berbagai faktor, seperti budaya, sosial, dan individu. Beberapa emosi yang umum dikenali adalah bahagia, sedih, marah, takut, jijik, dan netral. Meskipun banyak emosi yang bisa dikenali, penelitian ini memilih hanya enam emosi utama ini karena mereka dianggap sebagai emosi universal yang dapat dikenali secara luas di berbagai budaya. Enam emosi ini, yaitu bahagia, sedih, marah, takut, jijik, dan netral, memiliki ekspresi yang jelas dan mudah dikenali, sedangkan emosi seperti kaget dianggap ambigu karena sering kali tumpang tindih dengan emosi lain dan lebih sulit diekspresikan secara konsisten (Cao *et al*., 2014). Pendeteksian emosi menjadi semakin penting, terutama dalam aplikasi di bidang pemasaran, psikologi, interaksi manusia-mesin, dan pengobatan masalah kesehatan mental seperti depresi dan kecemasan. Keadaan emosional manusia mempengaruhi interaksi antar sesama serta komunikasi verbal dan non-verbal, seperti ekspresi wajah, karakter suara, dan isi pembicaraan. Suara, khususnya, merupakan salah satu cara efektif untuk mengekspresikan emosi (Qamhan *et al*., 2020).

Emosi dapat dibedakan berdasarkan frekuensi. Menurut Sauter *et al.* (2010) setiap emosi memiliki karakteristik frekuensi, yaitu:

a. Marah, memiliki nada yang bervariasi dari rendah hingga tinggi, memiliki intensitas yang tinggi, dan energi yang tinggi (500 – 2000 Hz).
b. Jijik, memiliki nada yang nada yang rendah, energi rendah (500 Hz), dan durasi waktu yang sedikit.
c. Takut, memiliki nada yang tinggi, sedikit variasi, energy rendah, dan kecepatan bicara lebih cepat dengan adanya banyak jeda dalam berbicara.
d. Sedih, memiliki nada yang tinggi, intensitas rendah tetapi lebih banyak energi (2000 Hz), dan durasi waktu yang panjang dengan lebih banyak jeda.

2.3. Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) adalah proses untuk memprediksi secara otomatis keadaan emosi seseorang berdasarkan sampel suara mereka. Sampel suara umumnya memiliki berbagai informasi seperti identitas pembicara, bahasa, emosi, konteks, lingkungan, jenis kelamin, dan usia, yang semuanya saling terkait satu sama lain. Pikiran manusia secara alami mampu memilah informasi tersebut, namun tidak berlaku bagi mesin. Mesin memerlukan pelatihan khusus untuk dapat mengekstraksi petunjuk yang berkaitan dengan informasi tertentu. Di antara berbagai jenis informasi tersebut, mengekstraksi petunjuk spesifik emosi untuk SER masih dianggap sebagai tugas yang menantang. Tantangan ini terutama setiap individu memiliki perbedaan cara pengekspresian emosi. Perbedaan ini muncul dari faktor-faktor seperti budaya dan latar belakang, etnisitas, suasana hati, jenis kelamin, cara berbicara, dan sebagainya. Agar SER otomatis berhasil, mesin harus mampu mengekstraksi petunjuk emosi dengan mempertimbangkan semua faktor tersebut (Premjeet *et al*., 2022).

Penerapan *Speech Emotion Recognition* (SER) dapat diaplikasikan dalam studi psikologis untuk mengetahui perilaku manusia di bidang kesehatan, pendidikan, keamanan, dan hiburan (Yadav *et al.,* 2021). Sebagai contoh, asisten virtual yang dapat merespon emosi pengguna dengan lebih efektif, sistem pemantauan yang dapat mengidentifikasi tingkat stres pengemudi, atau teknologi untuk penyandang disabilitas yang dapat bereaksi dengan kebutuhan dan emosi mereka secara lebih sensitif. Dalam industri kesehatan, pengenalan emosi dalam suara dapat membantu dalam pemantauan serta pengobatan masalah emosional, sekaligus mendukung penelitian terkait kesehatan mental (Mukarram *et al*., 2024).

2.4. Gelombang Audio

Gelombang audio adalah sinyal akustik yang merambat melalui medium seperti udara dan dapat dipahami sebagai variasi tekanan atau perubahan gaya per satuan luas dalam ruang. Komponen utama dari gelombang audio mencakup amplitudo,

frekuensi, dan waktu, yang semuanya saling berkaitan. Amplitudo mengacu pada besar kecilnya gelombang audio dan berhubungan langsung dengan volume atau kekuatan suara. Semakin tinggi amplitudo, semakin keras suara yang terdengar, sementara amplitudo yang rendah menghasilkan suara yang lebih lembut. Frekuensi merujuk pada jumlah siklus per detik dari gelombang suara, yang diukur dalam satuan *hertz* (Hz). Frekuensi yang lebih tinggi berhubungan dengan nada tinggi seperti suara *treble*, sementara frekuensi yang lebih rendah berkaitan dengan nada rendah seperti suara *bass* (Hawley, 2013).

2.5. Fourier Transform

Fourier Transform adalah sebuah teknik matematis yang digunakan untuk menganalisis sinyal dengan cara mengubahnya dari domain waktu ke domain frekuensi. Tujuan dari transformasi ini adalah untuk merepresentasikan sinyal dalam bentuk jumlah dari frekuensi, yang memungkinkan untuk lebih memahami bagaimana sinyal tersebut tersebar dalam spektrum frekuensinya. Fungsi dasar dari *Fourier Transform* adalah mengonversi fungsi waktu menjadi fungsi frekuensi. Hal ini sangat berguna untuk mempelajari komponen frekuensi dalam sinyal yang berubah seiring waktu. *Fourier Transform* memungkinkan melihat bagaimana sinyal tertentu terdiri dari berbagai frekuensi, yang tidak dapat dilihat langsung dalam domain waktu. *Fourier transform* memberikan cara yang sangat efektif untuk memecah sinyal ke dalam komponennya yang paling mendasar, yaitu frekuensi. Ini sangat penting dalam analisis sinyal statis yang konstan (Cattermole, 1965).

2.5.1. *Short-Time Fourier Transform (STFT)*

Fourier Transform memungkinkan untuk menguraikan sinyal kompleks menjadi komponen frekuensi, memberikan gambaran yang jelas tentang bagaimana energi dalam sinyal terdistribusi dalam berbagai frekuensi. Namun, sinyal ucapan bersifat tidak statis, yang berarti frekuensi kontennya berubah seiring waktu. Oleh karena itu, STFT digunakan untuk memecah sinyal menjadi segmen-segmen kecil yang

saling tumpang tindih, kemudian menerapkan *Fourier Transform* pada setiap segmen untuk mendapatkan gambaran perubahan frekuensi yang dinamis sepanjang waktu. STFT adalah metode yang sering digunakan dalam pengolahan suara untuk menganalisis bagaimana konten frekuensi dari sinyal berubah seiring waktu. Hal ini memungkinkan untuk menganalisis sinyal suara yang bersifat dinamis dan tidak statis, yang sangat relevan untuk tugas-tugas seperti pengenalan emosi (Vazhenina & Markov, 2020). STFT menghasilkan spektrum waktu-frekuensi yang memberi informasi tentang kedua domain tersebut, yang kemudian dapat digunakan untuk ekstraksi fitur, misalnya menggunakan *Log Mel Spectrogram*, untuk mengenali emosi berdasarkan variasi frekuensi.

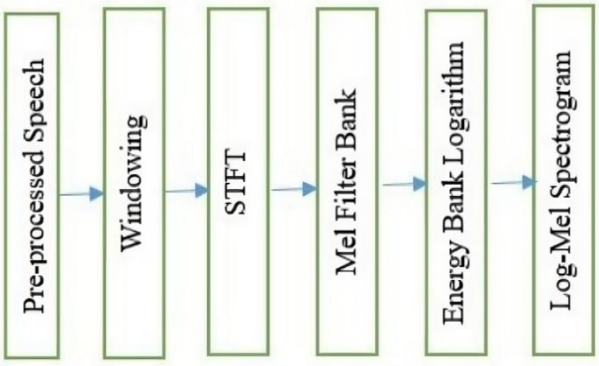
2.6. Feature Extraction

Feature Extraction atau ekstraksi fitur adalah proses mengidentifikasi dan mengambil informasi relevan dari data suara yang dapat digunakan untuk menganalisis atau mendeteksi emosi. Ekstraksi fitur pada suara melibatkan pengolahan sinyal audio untuk mengidentifikasi karakteristik penting, seperti frekuensi, intensitas, durasi, dan perubahan *pitch*. Fitur-fitur ini digunakan untuk membedakan emosi yang terkandung dalam suara. Ekstraksi fitur dalam pengenalan emosi pada suara sangat penting karena membantu meningkatkan akurasi pengenalan dan kinerja sinyal (Mukarram *et al*., 2024). Salah satu teknik yang umum digunakan ialah *log mel spectogram* yang dapat menangkap informasi penting dari karakteristik suara.

2.6.1. *Log Mel Spectogram*

Log-mel spectrogram adalah representasi visual dari spektrum frekuensi dalam sinyal suara yang bervariasi seiring waktu. Teknik ini sangat berguna dalam tugas pemrosesan audio seperti pengenalan emosi pada suara karena dapat menangkap fitur-fitur yang relevan dengan persepsi manusia terhadap sinyal suara. *Log-mel spectrogram* berasal dari skala mel, yaitu skala nada yang dipersepsikan oleh manusia dengan jarak yang sama. Hal ini membuat *log-mel spectrogram* lebih

sesuai dengan persepsi pendengaran manusia jika dibandingkan dengan skala frekuensi linier (Yadav *et al*., 2021). Proses komputasi *log-mel spectrogram* memiliki beberapa tahapan seperti yang ditampilkan pada Gambar 1.



Gambar 1. Proses Komputasi *Log-Mel Spectogram* (Yadav *et al.*, 2021).

Tahapan komputasi *log-mel spectogram* menurut Yadav *et al.* (2021), yaitu:

a. *Windowing*, membagi sinyal audio menjadi *frame* pendek agar dapat dianalisis secara lokal. Fungsi *window* yang sering digunakan adalah *Hanning Window* karena dapat mengurangi kebocoran spektral dengan membuat transisi yang halus di antara *frame*.

b. *Short-Time Fourier Transform* (STFT), mengubah sinyal dari domain waktu ke domain frekuensi pada setiap *frame*. STFT menghasilkan spektrum daya yang menunjukkan distribusi energi pada setiap frekuensi dalam *frame* tersebut. Persamaan STFT dapat dilihat pada Persamaan 1.

$$Y(k,\omega) = \sum_{n=-\infty}^{\infty} y[n]w[n-k]e^{-j\omega n} \quad \text{.....(1)}$$

c. *filterbank Mel*, memproyeksikan spektrum daya menjadi skala mel dan mendistribusikan energi pada frekuensi sesuai dengan skala mel. Setiap *band* pada skala Mel menangkap energi frekuensi dalam rentang tertentu yang relevan untuk persepsi manusia. Persamaan skala mel dapat dilihat pada Persamaan 2.

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \quad \text{.....(2)}$$

d. *Energy Bank Logarithm,* konversi ke skala logaritmik yang mencerminkan sensitivitas manusia terhadap perubahan intensitas suara dengan sifat logaritmik. Persamaan skala logaritmik dapat dilihat pada Persamaan 3.

$$d = 10 \log_{10}(\frac{m}{r}) \quad (3)$$

2.7. Augmentasi Data

Proses augmentasi pada data dilakukan untuk meningkatkan jumlah sampel pelatihan sesuai dengan kebutuhan di mana ukuran data asli yang tersedia tidak cukup. Proses ini menghasilkan sampel data pelatihan tambahan dengan mengubah bentuk data asli dalam kumpulan data pelatihan. Label dari data yang ditambahkan tidak dapat diubah, yang merupakan syarat untuk augmentasi data. Untuk augmentasi data audio, teknik yang sering digunakan adalah *Noise Injection*, *Time Shifting, Pitch Shifting,* dan *Time Stretching* (Yadav *et al*., 2021). Teknik augmentasi data menurut Makhmudov *et al*. (2024), yaitu:

a. *Time Strecthing,* Teknik augmentasi khusus ini berfokus pada manipulasi panjang atau durasi klip audio dengan cara meregangkan *atau* mengompresi audio tanpa memengaruhi *pitch.*

b. *Pitch Shifting,* Teknik ini berfokus pada modifikasi *pitch* dari audio sambil mempertahankan panjang atau durasinya tetap konstan. Dengan menyesuaikan *pitch* ke atas atau ke bawah, kita memperkenalkan berbagai variasi yang secara efektif meniru situasi di mana frekuensi vokal mungkin berbeda, seperti perbedaan dalam usia, suasana hati, atau karakteristik vokal individu. Variasi ini penting karena manusia secara alami memiliki *pitch* suara yang beragam, mulai dari suara *bass* yang dalam hingga nada tinggi.

2.8. Normalisasi Data

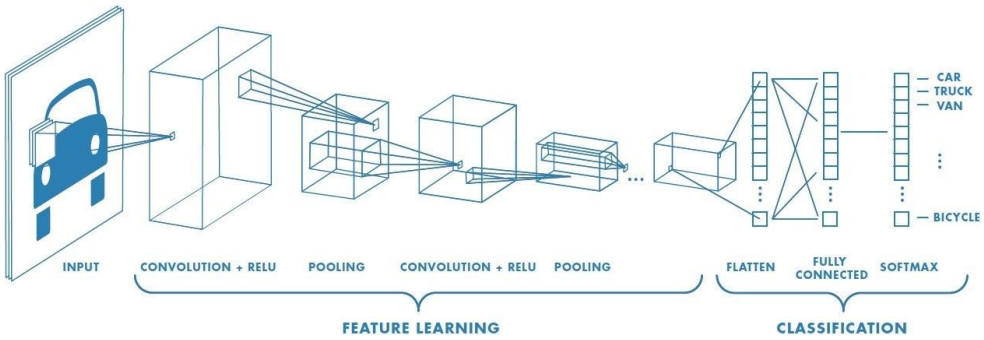
Normalisasi data merupakan proses penting dalam *pre-processing* yang bertujuan untuk menyamakan kontribusi masing-masing fitur dalam dataset, terutama ketika

fitur memiliki satuan atau rentang nilai yang berbeda. Salah satu metode normalisasi yang umum digunakan adalah *Z-Score Normalization*, juga dikenal sebagai standardisasi. Metode ini bekerja dengan mentransformasi setiap nilai data x menjadi nilai standar z. Pada rumus, μ adalah rata-rata dari fitur, dan σ adalah standar deviasi dari fitur tersebut. Proses ini menghasilkan distribusi data baru dengan mean 0 dan standar deviasi 1, tanpa mengubah bentuk distribusi asli. Metode ini mempertahankan bentuk distribusi data dan membuatnya tahan terhadap outlier (Singh & Singh, 2019). Menurut Singh & Singh (2019), Persamaan *Z-Score Normalization* dapat dilihat pada Persamaan 4.

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

2.9. Convolutional Neural Network (CNN)

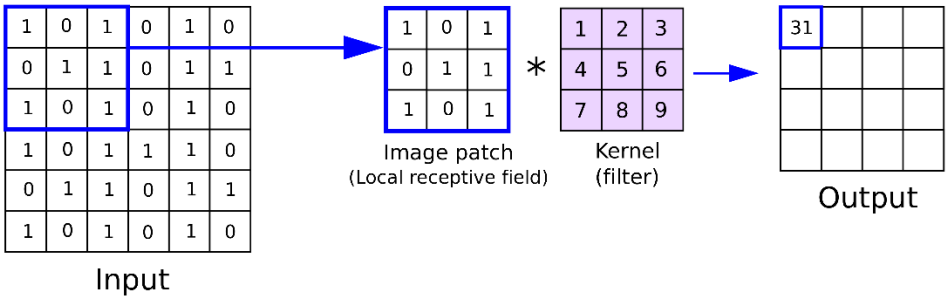
Convolutional Neural Network (CNN) adalah jenis jaringan saraf tiruan yang dirancang untuk memproses data gambar dua dimensi. CNN terdiri dari beberapa lapisan seperti *convolutional layer* untuk ekstraksi fitur, *pooling layer* untuk pengurangan dimensi, dan *fully-connected layer* untuk klasifikasi (Karrach & Pivarciova, 2023). CNN menunjukkan kinerja unggul dalam beragam tugas visual seperti klasifikasi dan segmentasi gambar, pengambilan serta deteksi objek, pengkategorian gambar, pengenalan wajah, estimasi posisi, pengenalan tanda lalu lintas, dan pemrosesan suara. Sebagai jenis jaringan saraf, CNN juga banyak digunakan dalam pemrosesan data gambar digital (Ayeni, 2022). Gambar 2 menunjukkan arsitektur *Convolutional Neural Network*.



Gambar 2. Arsitektur *Convolutional Neural Network* (Purwono *et al.*, 2023).

2.9.2. *Convolutional Layer*

Convolutional layers adalah komponen dasar CNN, yang banyak digunakan di berbagai bidang seperti pengenalan gambar dan suara. Lapisan ini bertanggung jawab untuk mengekstraksi fitur secara otomatis dengan melakukan operasi seperti *cross-correlation* antara data *input* dan serangkaian kernel yang dapat dipelajari. Kernel, yang awalnya diinisialisasi secara acak, dioptimalkan selama pelatihan untuk meminimalkan *loss function*, sehingga memungkinkan jaringan untuk mempelajari fitur-fitur yang relevan dari data (Lee, 2023). Gambar 3 menunjukkan proses *Convolutional Layer.*

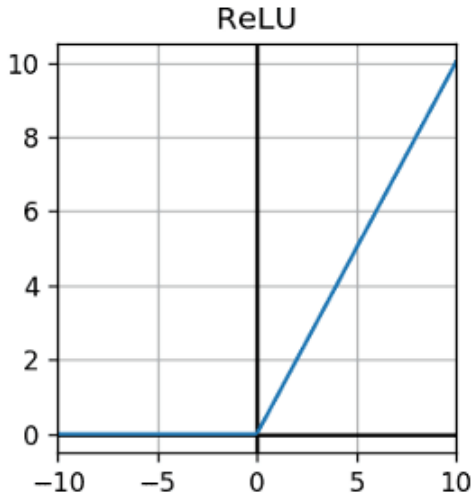


Gambar 3. Ilustrasi *Convolutional Layer* (Satrio & Wibowo, 2021).

Volume output dalam *convolutional layer* ditentukan oleh tiga *hyperparameter* utama, yaitu *depth*, *stride*, dan *padding*. *Depth* menunjukkan jumlah filter yang digunakan dalam operasi konvolusi, di mana masing-masing filter belajar mengenali fitur berbeda seperti tepi, warna, atau tekstur. Stride adalah jumlah langkah pergeseran filter di setiap gerakan, *stride* 1 menggeser filter satu piksel, sedangkan *stride* 2 menggeser dua piksel, yang menghasilkan volume output yang lebih kecil secara spasial. *Padding* digunakan untuk mengontrol ukuran *output* dengan menambahkan nol di sekitar tepi *input*, untuk mencegah kehilangan informasi (Bezdan & Bacanin, 2019).

2.9.3. *Activation Function*

Activation function merupakan fungsi CNN yang mengonversi sinyal linear dari layer sebelumnya menjadi non-linear. *Activation function* yang umum digunakan adalah *Rectified Linear Unit* (ReLU). ReLU merupakan *activation function* yang digunakan pada neuron seperti fungsi aktivasi lainnya, setiap *node* menggunakan *activation function rectifier* yang dikenal dengan *node* ReLU. Alasan utama ReLU digunakan karena lebih efisien dalam melakukan komputasi dibandingkan dengan *activation function* konvensional seperti *sigmoid* dan *hyperbolic tangent*, tanpa membuat perbedaan yang signifikan terhadap generalisasi pada akurasi (Satrio & Wibowo, 2021). Gambar 4 menunjukkan ilustrasi *Rectified Linear Unit*.

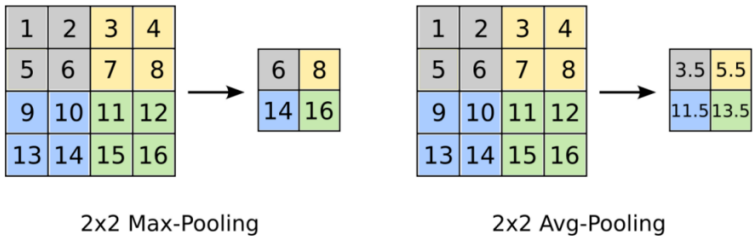


Gambar 4. Ilustrasi *Rectified Linear Unit* (Lee, 2023).

2.9.4. *Pooling Layer*

CNN sering menggunakan operasi *pooling layer* setelah *convolution layer*, yang berfungsi untuk mengurangi dimensi, juga dikenal sebagai *subsampling* atau *downsampling.* Dua jenis *pooling layer* yang sering digunakan adalah *max pooling* dan *average pooling*, masing-masing mengambil nilai maksimum dan rata-rata. *Max pooling* lebih sering digunakan dibandingkan *average pooling* (Bezdan & Bacanin, 2019). *Max Pooling* lebih efisien dan efektif dalam mengurangi dimensi,

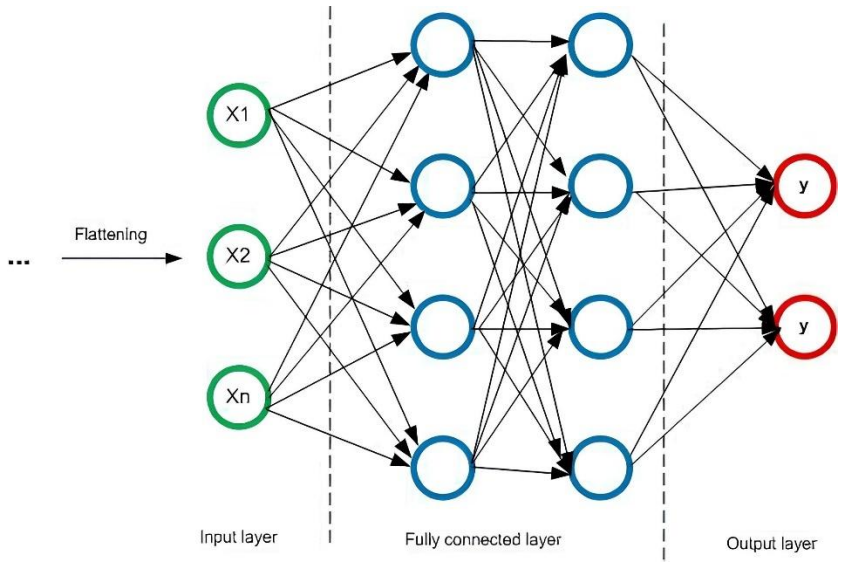
namun *average pooling* mampu mempertahankan informasi leih banyak (Lee, 2023). Gambar 5 menunjukkan proses *Max Pooling* dan *Average Pooling*.



Gambar 5. *Max Pooling* dan *Average Pooling* (Karrach & Pivarciova, 2023).

2.9.5. *Fully Connected Layer*

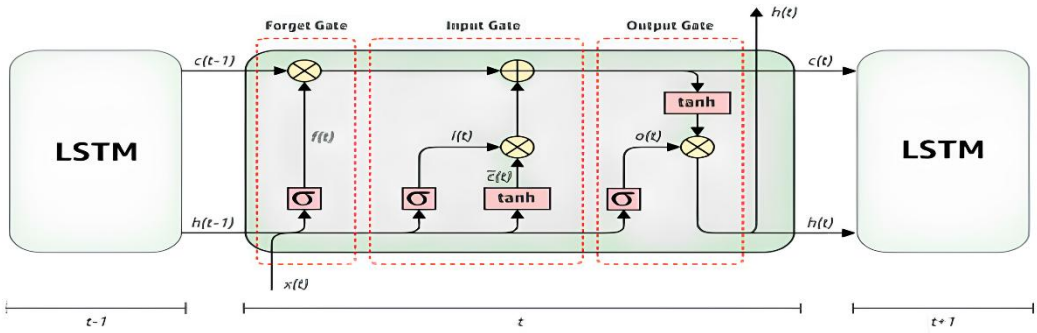
Fully connected layer berhubungan dengan lapisan sebelumnya, *convolution layer* dan *pooling layer*, setiap neuron pada *fully connected layer* terhubung ke setiap neuron di lapisan sebelumnya. *Fully connected layer* juga merupakan lapisan terakhir dalam arsitektur CNN, sehingga outputnya akan menjadi hasil akhir. *Fully connected layer* terakhir dalam arsitektur memiliki jumlah *neuron output* yang sama dengan jumlah kelas yang akan diklasifikasi (Bezdan & Bacanin, 2019). Gambar 6 menunjukkan proses *Fully connected layer.*



Gambar 6. Ilustrasi *Fully connected layer* (Kurnia & Wibowo, 2021).

2.10. Long Short-Term Memory (LSTM)

Arsitektur RNN memiliki kelemahan terutama pada masalah *vanishing gradient* yang sering terjadi saat menangani urutan data yang panjang. Untuk mengatasi masalah tersebut arsitektur RNN dikembangkan, salah satunya, yaitu *Long Short-Term Memory* (LSTM). LSTM menggunakan *memory cells* dan tiga gerbang utama, yaitu *input gate* berfungsi dalam memutuskan informasi baru yang masuk kedalam *cell state, forget gate* berfungsi untuk menghapus informasi dari *cell state,* dan *output gate* yang berfungsi dalam memilah informasi yang berguna dari arus *cell state* dan menampilkannya sebagai output (Puteri, 2023). LSTM memungkinkan penyimpanan informasi penting untuk jangka waktu yang lebih lama. LSTM dirancang khusus untuk menangani data berurutan dengan lebih efektif yang sangat cocok untuk tugas-tugas seperti analisis sentimen, penerjemahan bahasa, dan pengenalan pola dalam urutan data yang panjang. Gambar 7 menunjukkan arsitektur *Long Short-Term Memory.*



Gambar 7. Arsitektur *Long Short-Term Memory* (Nugroho *et al*., 2021).

LSTM memiliki beberapa komponen. Kompenen LSTM menurut Nugroho *et al*. (2021), yaitu:

a. *Forget Gate*

Forget gate merupakan gerbang pertama pada LSTM untuk menentukan informasi mana yang akan dipertahankan atau dibuang dari *cell state*. Gerbang ini menerima input h_{t-1} dan X_t untuk menghasilkan nilai 0 atau 1 pada C_{t-1}.

Ketika *forget gate* bernilai 1, maka *cell state* akan menyimpan informasi, sedangkan jika bernilai 0 maka informasi akan dibuang dari *cell state.* Persamaan *Forget Gate* dapat dilihat pada Persamaan 5.

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \quad (5)$$

b. *Input Gate*

Input gate merupakan gerbang kedua pada LSTM untuk menentukan informasi apa yang akan disimpan pada *cell state*. Gerbang ini terdiri dari lapisan *sigmoid* dan lapisan *tanh*. Lapisan *sigmoid* memutuskan nilai mana yang akan diperbarui. Lapisan *tanh* membuat nilai baru C_t untuk ditambahkan ke *cell state*. *Output* kedua lapisan ini digabungkan untuk memperbarui informasi *cell state*. Persamaan lapisan *sigmoid* dapat dilihat pada Persamaan 6 dan persamaan lapisan *tanh* dapat dilihat pada Persamaan 7.

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \quad (6)$$

$$C_t = tanh(W_c \cdot [h_{t-1}, X_t] + b_c) \quad (7)$$

Setalah memperbarui nilai *cell state* lama C_{t-1} menjadi C_t dengan mengalikan *cell state* lama dengan f_t untuk menghapus nilai pada *forget gate* sebelumnya. Selanjutnya, ditambahkan dengan $i_t C_t$ sebagai nilai baru dan digunakan untuk memperbarui nilai *cell state.* Persamaan untuk memperbarui nilai *cell state* dapat dilihat pada Persamaan 8.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \quad (8)$$

c. *Output Gate*

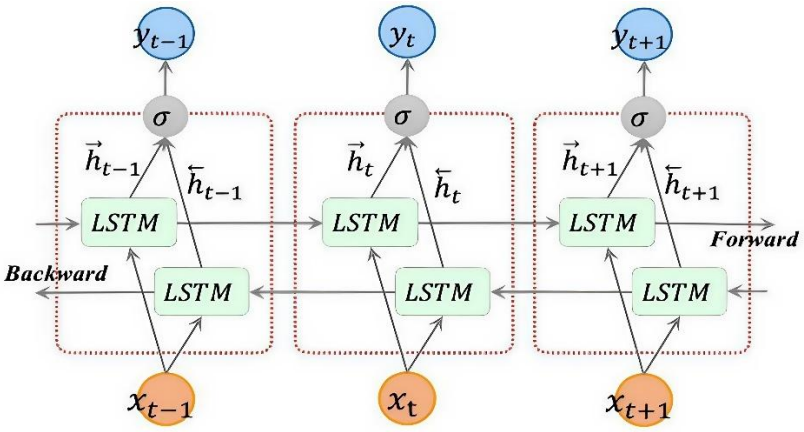
Output gate merupakan gerbang terakhir pada LSTM untuk menentukan output dari *cell state*. Pertama, lapisan *sigmoid* menentukan bagian dari *cell state* mana yang menjadi *output*. Persamaan tersebut dapat dilihat pada persamaan 9. Selanjutnya, *output* tersebut dimasukkan kedalam lapisan *tanh* dan dikalikan dengan lapisan *sigmoid* agar *output* sesuai dengan yang diputuskan sebelumnya. Persamaan tersebut dapat dilihat pada persamaan 10.

$$O_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \quad (9)$$

$$h_t = O_t \cdot tanh(C_t) \quad (10)$$

2.10.1. *Bidirectional Long Short-Term Memory* (BI-LSTM)

BiLSTM adalah varian dari LSTM yang memiliki dua jaringan LSTM dimana jaringan LSTM pertama berfungsi dalam memproses urutan masukan data ke arah depan (*forward*) dan jaringan LSTM kedua berfungsi dalam memproses urutan data dari arah sebaliknya (*backward*). Kemudian output dari jaringan LSTM *forward* dan *backward* digabungkan pada setiap urutan waktu (Puteri, 2023). BiLSTM mampu mengambil informasi dengan membaca konteks melalui dua arah sekaligus. Namun, BiLSTM membutuhkan dataset yang cukup besar untuk menghindari model yang over-fitting. Selain itu, biaya dan waktu komputasi yang dibutuhkan juga tinggi (Nugroho *et al*., 2021). Gambar 8 menunjukkan arsitektur *Bidirectional Long Short-Term Memory*.



Gambar 8. Arsitektur *Bidirectional Long Short-Term Memory* (Puteri, 2023).

BiLSTM memiliki beberapa komponen. Kompenen BiLSTM menurut Puteri (2023), yaitu:

a. *Forward*

Bagian ini bekerja seperti LSTM biasa dengan memproses informasi secara berurutan dari langkah waktu pertama hingga terakhir. Dalam proses ini informasi sebelumnya digunakan untuk mempengaruhi keputusan dan output

di langkah waktu berikutnya. Proses *forward* LSTM dapat dilihat pada Persamaan 11.

$$\overrightarrow{h_t} = LSTM(x_t, h_{t-1}) \quad (11)$$

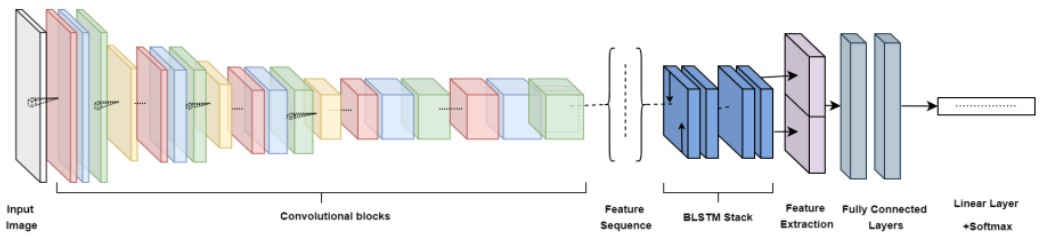
b. *Backward*

Bagian ini memproses urutan data dalam arah terbalik dari langkah waktu terakhir hingga pertama. Proses ini memungkinkan model untuk menangkap konteks yang lebih luas seperti informasi yang relevan setelah langkah waktu tertentu yang mungkin tidak ditangkap oleh *forward.* Proses *backward* LSTM dapat dilihat pada Persamaan 12.

$$\overleftarrow{h_t} = LSTM(x_t, h_{t+1}) \quad (12)$$

2.11. Convolutional Recurrent Neural Network (CRNN)

Model CRNN merupakan kombinasi dari *Convolutional Neural Network* dan *Recurrent Neural Network* (Yadav *et al*., 2021). Model ini menggunakan komponen CNN yang memproses fitur masukan, seperti *log-Mel spectrogram* dengan memperlakukannya sebagai gambar untuk mengekstrak pola spasial. Lapisan BiLSTM digunakan untuk memahami hubungan antar waktu dalam data, sehingga model dapat mengenali urutan isyarat emosi dalam suara. BiLSTM sangat efektif dalam menangani data ucapan karena dapat memproses informasi dari dua arah sehingga mampu memahami konteks dari suara sebelumnya maupun yang akan datang (Meyer *et al*., 2021). Gambar 9 menunjukkan arsitektur *Convolutional Recurrent Neural Network.*



Gambar 9. Arsitektur CRNN (Markou *et al.*, 2021).

CRNN memiliki beberapa komponen. Komponen CRNN menurut Markou *et al.* (2021), yaitu:

a. *Convolutional stage*

   Pada tahap ini, model bertugas mengekstraksi fitur spasial input. input diproses melalui beberapa blok konvolusi yang dirancang untuk menangkap pola visual. Tiap blok konvolusi mencakup lapisan konvolusi dengan kernel untuk mendeteksi fitur lokal, lalu *batch normalization* yang menstabilkan distribusi data antar lapisan, serta fungsi aktivasi yang memastikan gradien tetap ada meskipun neuron tidak aktif. Beberapa blok juga menggunakan *max pooling* untuk mengurangi dimensi, sehingga meningkatkan efisiensi komputasi.

b. *Recurrent stage*

   Pada tahap ini, model menangkap informasi temporal dari fitur yang dihasilkan oleh tahap konvolusi. Tahap ini menggunakan *Bidirectional Long Short-Term Memory* (BiLSTM), yang memproses data sekuensial dalam dua arah. BiLSTM mampu menangkap dependensi temporal dari masa lalu maupun masa depan, sehingga menghasilkan representasi fitur yang lebih kaya dan informatif. Hasil *output* dari dua arah ini digabungkan untuk membentuk representasi tunggal yang lebih lengkap. Tahap ini sangat penting dalam tugas sekuensial atau memiliki hubungan yang berurutan. BiLSTM dapat memodelkan hubungan tersebut dengan efektif.

2.12. *Confusion Matrix*

Confusion Matrix adalah pengukuran performa untuk masalah klasifikasi *machine learning* dimana keluaran dapat berupa dua kelas atau lebih. *Confusion Matrix* berupa tabel yang berisi informasi nilai aktual dan prediksi yang dibuat oleh sistem klasifikasi. (Hutagaol & Mauritsius, 2020). Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada *confusion matrix,* yaitu *True Positif, True Negatif, False Positif,* dan *False Negatif. Confusion Matrix* untuk *multiclass* dapat dilihat pada Tabel 2.

Tabel 2. *Multiclass Confusion Matrix* (Markoulidakis *et al.*, 2021)

		Prediksi			
		C_1	C_2	…	C_n
Aktual	C_1	$C_{1.1}$	FP	…	$C_{1.n}$
	C_2	FN	TP	…	FN
	…	…	…	…	…
	C_n	$C_{n.1}$	FP	…	$C_{n.n}$

Keterangan:

True Positive (TP) = Data aktual positif yang diprediksi positif

False Positive (FP) = Data aktual negatif yang diprediksi positif

True Negative (TN) = Data aktual negatif yang diprediksi negatif

False Negative (FN) = Data aktual positif yang diprediksi negatif

Untuk mengukur performa *confussion matrix,* Hutagaol & Mauritsius (2020) menggunakan perhitungan akurasi*,* presisi*, recall,* dan *F1-score* berikut:

a. Akurasi

Akurasi adalah metrik evaluasi yang mengukur seberapa baik model membuat prediksi yang benar dari total prediksi yang dilakukan. Dalam konteks klasifikasi, akurasi memberikan gambaran mengenai seberapa sering model memprediksi kelas yang benar, baik itu kelas positif maupun negatif. Akurasi merupakan hasil perhitungan ketepatan suatu model dalam mengklasifikasikan data untuk diprediksi dengan benar. Akurasi juga dapat menggambarkan kedekatan nilai prediksi dengan nilai aktual, sehingga menjadi indikator yang penting dalam menilai performa keseluruhan model klasifikasi. Persamaan akurasi dapat dilihat pada Persamaan 13.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

b. Presisi

Presisi adalah metrik evaluasi yang mengukur seberapa baik model membuat prediksi yang benar untuk kelas positif dari total prediksi positif yang dilakukan. Presisi membantu menghitung seberapa sering model memprediksi kelas positif dengan benar, di antara semua prediksi positif yang dibuat oleh model. Persamaan presisi dapat dilihat pada Persamaan 14.

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (14)$$

c. *Recall*

Recall adalah metrik evaluasi yang menggambarkan seberapa baik suatu model dalam mengidentifikasi kelas positif dengan benar. Persamaan *recall* dapat dilihat pada Persamaan 15.

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

d. F1-Score

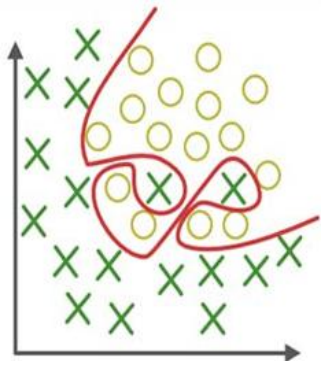
F1-Score merupakan metrik evaluasi yang mencerminkan keseimbangan antara Presisi dan *Recall*. Nilai *F1-Score* akan memberikan informasi tentang seberapa baik model kita dalam menggabungkan kemampuan Presisi dan *Recall*, sehingga kita bisa memahami seberapa efektif model kita dalam melakukan klasifikasi. Persamaan *F1-Score* dapat dilihat pada Persamaan 16.

$$F1\text{-}Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (16)$$

2.13. *Overfitting*

Overfitting adalah masalah yang sering terjadi dalam model *machine learning* ketika model terlalu kompleks, sehingga tidak hanya menangkap pola mendasar dalam data, tetapi juga *noise* atau fluktuasi acak yang ada dalam data. Hal ini menyebabkan model dapat menghasilkan prediksi yang sangat akurat pada data latih, tetapi gagal dalam generalisasi ketika diuji dengan data yang belum pernah dilihat sebelumnya. *Overfitting* umumnya terjadi ketika model memiliki terlalu banyak parameter dibandingkan dengan jumlah data yang tersedia, yang

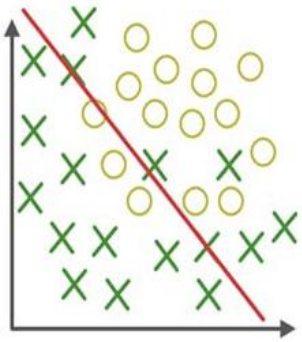
menyebabkan model menghafal data latih dan bukannya mempelajari pola yang lebih umum (Pothuganti, 2018). Gambar 10 menunjukkan visualisasi *overfitting*.



Gambar 10. Visualisasi *overfitting* (Kolluri *et al.*, 2020).

2.14. *Underfitting*

Underfitting adalah kebalikan dari *overfitting*, di mana model terlalu sederhana untuk menangkap pola yang ada dalam data. Hal ini terjadi ketika model tidak memiliki cukup parameter atau fitur untuk mencerminkan kompleksitas data yang ada atau ketika model tidak dilatih cukup lama untuk mempelajari pola tersebut. Dalam kasus *underfitting*, model tidak dapat memberikan prediksi yang baik, baik pada data latih maupun data uji. Untuk mengatasi *underfitting*, diperlukan peningkatan kompleksitas model, seperti penggunaan model yang lebih canggih atau menambah fitur yang relevan dalam proses pelatihan (Pothuganti, 2018). Gambar 11 menunjukkan visualisasi *underfitting*.



Gambar 11. Visualisasi *underfitting* (Kolluri *et al.*, 2020).

2.15. *Python*

Python merupakan bahasa pemrograman utama di bidang Kecerdasan Buatan yang berkembang pesat. Kombinasi aksesibilitas, ekosistem yang kuat, dan kemampuan yang serbaguna telah menjadikannya pilihan utama bagi peneliti, pengembang, dan organisasi yang mendorong batasan kecerdasan mesin. Sebagai bahasa pemrograman tingkat tinggi dan interpretatif, *Python* dirancang dengan menekankan pada keterbacaan kode melalui penggunaan sintaks yang sederhana dan bersih, sehingga memudahkan proses pengembangan dan pemeliharaan program. Kekuatan ekosistemnya didukung oleh pustaka standar yang sangat luas serta ketersediaan ribuan pustaka pihak ketiga yang dikelola oleh komunitas, khususnya untuk bidang komputasi ilmiah dan pembelajaran mesin (Mantrala, 2025).

2.16. *Library*

Library dalam *Python* adalah kumpulan modul atau paket yang dapat digunakan kembali, yang menyediakan berbagai perangkat, fungsi, dan kelas untuk tugas-tugas spesifik. *Library* membantu pemrogram agar tidak perlu menulis kode dari awal, sehingga secara signifikan menyederhanakan dan mempercepat pengembangan perangkat lunak di berbagai bidang. Pendekatan ini meningkatkan produktivitas dan kualitas kode secara keseluruhan, terutama dalam menyederhanakan proses-proses penting seperti analisis dan visualisasi data, pengambilan data tidak terstruktur dari web, pemrosesan gambar, serta pembangunan model *machine learning* dan pengolahan informasi tekstual (Gholizadeh, 2022). Berikut contoh *library*, yaitu:

a. *Library* Tensorflow

   TensorFlow dirancang untuk perhitungan numerik berkinerja tinggi, yang terutama digunakan dalam aplikasi *machine learning* dan *deep learning*. TensorFlow memiliki arsitektur fleksibel yang memungkinkan penerapan yang mudah di berbagai platform (Pavithra & Prakash, 2019).

b. *Library* Pandas

Pandas digunakan untuk merepresentasikan dan memanipulasi data terstruktur dalam memori. Pandas menyediakan *DataFrame*, yang memfasilitasi analisis dan manajemen data yang efisien. (Reiss *et al.*, 2021).

c. *Library* Numpy

Numpy digunakan untuk komputasi numerik dan ilmiah. *Library* ini menyediakan cara yang kuat dan efisien untuk memanipulasi *array* dan matriks multidimensi yang besar (Diana & Mathivanan, 2023).

d. *Library* Librosa

Librosa digunakan untuk analisis musik dan suara, menyediakan alat penting untuk bekerja dengan data audio. *Library* ini memfasilitasi ekstraksi fitur dan memungkinkan visualisasi sinyal audio (Patil & Zade, 2023).

e. *Library* Matplotlib

Matplotlib digunakan untuk visualisasi dan pembuatan grafik. Pustaka ini menyediakan rangkaian fungsi yang fleksibel dan komprehensif untuk menghasilkan berbagai macam plot, diagram, dan gambar, baik statis, animasi, maupun interaktif (Diana & Mathivanan, 2023).

f. *Library* Scikit-learn

Scikit-learn menyediakan berbagai fungsionalitas untuk tugas *machine learning*, analisis data, dan pemodelan. *Library* ini menawarkan berbagai alat untuk *preprocessing*, ekstraksi fitur, seleksi fitur, dan penskalaan fitur (Diana & Mathivanan, 2023).

2.17. *Kaggle*

Kaggle adalah sebuah platform online yang berfungsi sebagai komunitas dan tempat kompetisi bagi para praktisi ilmu data. *Kaggle* menyediakan lingkungan kernel, yaitu sebuah ruang kerja interaktif berbasis *browser* tempat pengguna dapat menulis dan menjalankan kode secara langsung tanpa memerlukan instalasi yang rumit di komputer lokal. Fitur ini juga dirancang agar memungkinkan para pengguna untuk saling berbagi kode dan metode. (Hayashi *et al.*, 2021).

III. METODOLOGI PENELITIAN

3.1. Waktu dan Tempat

3.1.1. Tempat

Penelitian dilaksanakan di Laboratorium Rekayasa Perangkat Lunak, Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung yang beralamatkan di Jalan Prof. Dr. Ir. Sumantri Brojonegoro No.1, Gedong Meneng, Kota Bandar Lampung, Provinsi Lampung.

3.1.2. Waktu

Penelitian ini dilaksanakan pada pada bulan Oktober 2024 sampai bulan Juni 2025. Untuk penjelasan kegiatannya terdapat pada Tabel 3.

3.2. Data dan Alat

3.2.1. Data

Dataset yang digunakan, yaitu CREMA-D. Dataset ini merupakan audio dalam bahasa inggris. CREMA-D terdiri dari 7.442 klip audio dengan format .wav dari 91 aktor dan durasi audio sekitar 3 detik. Para aktor mengucapkan salah satu dari 12 kalimat. Kalimat-kalimat tersebut disajikan dengan menggunakan salah satu dari enam emosi berbeda (Marah, Jijik, Takut, Bahagia, Netral, dan Sedih) dan empat tingkat intensitas yang berbeda (Rendah, Sedang, Tinggi, dan *unspecified*).

Tabel 3. Alur Pelaksanaan Penelitian

Kegiatan	2024																2025															
	Okt				**Nov**				**Des**				**Jan**				**Feb**				**Mar**				**Apr**				**Mei**				**Jun**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Studi Literatur																																				
Pengumpulan Data																																				
Preprocessing																																				
Resampling Audio																																				
Trimming																																				
Noise Reduction																																				
Padding																																				
Augmentasi Data																																				
Feature Extraction																																				
Data Preparation																																				
Pembuatan Model																																				
Evaluasi Model																																				
Penyusunan Skripsi																																				

Klip berasal dari 48 aktor pria dan 43 aktor wanita. Usia aktor berkisar antara 20 hingga 74 tahun dengan usia rata-rata 43 tahun. Aktor memiliki latar belakang ras dan etnis yang beragam (Afrika-Amerika, Asia, Kaukasia, Hispanik, *Unspecified*), mereka sebagian besar adalah orang Kaukasia, tetapi beberapa orang Afrika-Amerika, Hispanik, dan Asia juga berpartisipasi. Distribusi usia aktor dapat dilihat pada Tabel 4 dan distribusi ras serta etnis aktor dapat dilihat pada Tabel 5.

Tabel 4. Distribusi Usia Aktor (Cao *et al*., 2014)

Umur	**Jumlah**
20-29 Tahun	34
30-39 Tahun	23
40-49 Tahun	16
50-59 Tahun	12
60-69 Tahun	5
Diatas 70 Tahun	1

Tabel 5. Distribusi Ras dan Etnis aktor (Cao *et al*., 2014)

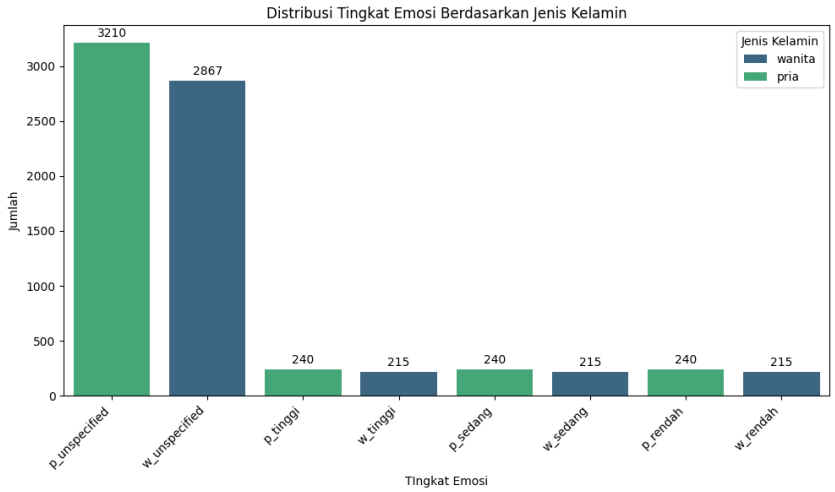
		Etnis		
		Not Hispanic	*Hispanic*	Total
Ras	*Caucasian*	53	8	61
	African American	21	1	22
	Asian	7	0	7
	Unspecified	0	1	1

Aktor mengucapkan kalimat dengan tingkat dan jenis emosi yang beragam. Berikut 12 kalimat yang diucapkan oleh para aktor:

a. *It's eleven o'clock.*
b. *That is exactly what happened.*
c. *I'm on my way to the meeting.*
d. *I wonder what this is about.*
e. *The airplane is almost full.*
f. *Maybe tomorrow it will be cold.*
g. *I would like a new alarm clock*
h. *I think I have a doctor's appointment.*
i. *Don't forget a jacket.*

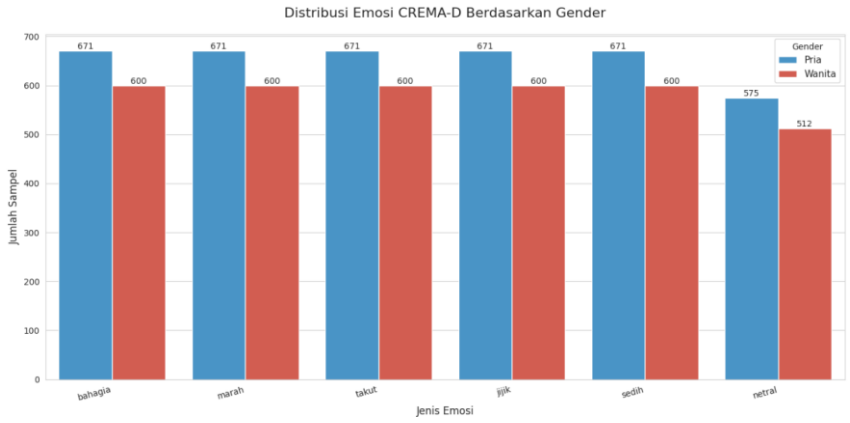
j. *I think I've seen this before.*
k. *The surface is slick.*
l. *We'll stop in a couple of minutes.*

Para aktor diarahkan untuk mengekspresikan kalimat pertama dalam tiga tingkat intensitas: rendah, sedang, dan tinggi. Untuk 11 kalimat lainnya, tingkat intensitasnya tidak ditentukan. Dataset memiliki jumlah data yang signifikan untuk tingkat intensitas *unspecified.* Untuk tingkat emosi yang lebih terdefinisi, seperti tinggi, sedang, dan rendah, distribusinya cenderung seimbang antara pria dan wanita. Distribusi tingkat intensitas berdasarkan jenis kelamin dapat dilihat pada Gambar 12.



Gambar 12. Distribusi Tingkat Emosi Berdasarkan Gender Dataset CREMA-D.

Distribusi jumlah emosi pada dataset cenderung seimbang antara pria dan wanita. Untuk emosi bahagia, marah, takut, jijik, dan sedih perbedaan jumlah sampelnya relatif kecil dan konsisten di kelima kelas emosi tersebut. Sementara itu, untuk emosi netral, jumlah sampel pria sedikit lebih sedikit. Keseimbangan distribusi sampel memastikan model dapat belajar dan menggeneralisasi karakteristik suara dari kedua gender secara adil, sehingga tidak ada bias yang signifikan terhadap gender tertentu saat mengklasifikasikan emosi. Distribusi jumlah kelas berdasarkan gender dapat dilihat pada Gambar 13.



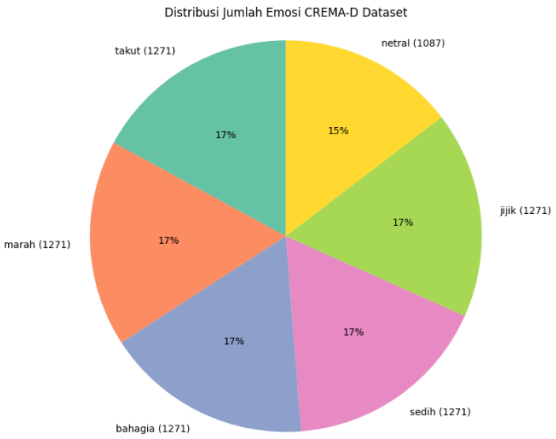
Gambar 13. Distribusi Jenis Emosi Berdasarkan Gender Dataset CREMA-D.

Proses pengolahan data diawali dengan pembuatan *dataframe* yang berfungsi untuk menyimpan label emosi dan *path* audio dari dataset. Bentuk dari data yang telah didapatkan dapat dilihat pada Gambar 14.

	Emosi	Path
0	takut	/content/cremad/AudioWAV/1084_TAI_FEA_XX.wav
1	netral	/content/cremad/AudioWAV/1027_DFA_NEU_XX.wav
2	marah	/content/cremad/AudioWAV/1058_IWL_ANG_XX.wav
3	sedih	/content/cremad/AudioWAV/1063_IWL_SAD_XX.wav
4	bahagia	/content/cremad/AudioWAV/1046_IEO_HAP_LO.wav
5	jijik	/content/cremad/AudioWAV/1035_TIE_DIS_XX.wav

Gambar 14. Dataset CREMA-D.

Dataset menunjukkan bahwa emosi jijik, takut, marah, bahagia, dan sedih memiliki jumlah sampel yang seimbang, masing-masing sebanyak 1.271 data, yang setara dengan 17% dari total data. Sementara itu, emosi netral memiliki jumlah sampel yang lebih sedikit, yaitu 1.087 data. Distribusi kelas pada dataset dapat dilihat pada Gambar 15.



Gambar 15. Distribusi kelas dataset CREMA-D.

3.2.2. Alat

3.2.2.1. Perangkat Keras

Perangkat keras yang digunakan berupa laptop dengan spesifikasi:

a. *Processor* : 13th Gen Core(TM) i7-13620H 2.40 GHz
b. *RAM* : 16 GB
c. *GPU* : GeForce RTX 4050
d. *Storage* : SSD 512 GB

3.2.2.2. Perangkat Lunak

Perangkat lunak yang digunakan dalam penelitian ini, yaitu:

g. Sistem Operasi Windows 11 64-bit

   Sistem operasi berfungsi sebagai platform utama yang mengelola perangkat keras dan perangkat lunak. Selain itu sistem operasi juga berperan dalam penyusunan laporan penelitian.

h. *Kaggle*

   Pada penelitian ini*, Kaggle* sebagai platform berbasis *cloud* digunakan untuk mengolah, menganalisis dan mempersiapkan data audio, yang kemudian akan digunakan untuk melatih dan mengevaluasi model CRNN.

i. Python 3.10.12

Pada penelitian ini, python digunakan untuk membangun algoritma dan memanfaatkan berbagai *library* yang telah tersedia.

j. *Library* Tensorflow 2.17.0

Pada penelitian ini, tensorflow digunakan untuk membangun dan melatih model CRNN yang menggabungkan CNN dan BiLSTM.

k. *Library* Pandas 2.2.2

Pada penelitian ini, pandas digunakan untuk pengelolaan data audio, pemrosesan dataset, dan menyimpan data hasil *preprocessing* sebelum dimuat ke model CRNN.

l. *Library* Numpy 1.26.4

Pada penelitian ini, numpy digunakan untuk menangani sinyal audio dalam bentuk array, melakukan perhitungan matriks, dan menghitung *log-Mel Spectrogram* yang digunakan sebagai input model.

m. *Library* Librosa 0.10.2.post1

Pada penelitian ini, librosa digunakan untuk analisis suara yang membantu dalam membuat program python untuk menangani dokumen suara.

n. *Library* Matplotlib 3.8.0

Pada penelitian ini, matplotlib digunakan menampilkan visualisasi *log-Mel Spectrogram*, distribusi data audio, dan hasil evaluasi model seperti kurva akurasi dan *loss*.

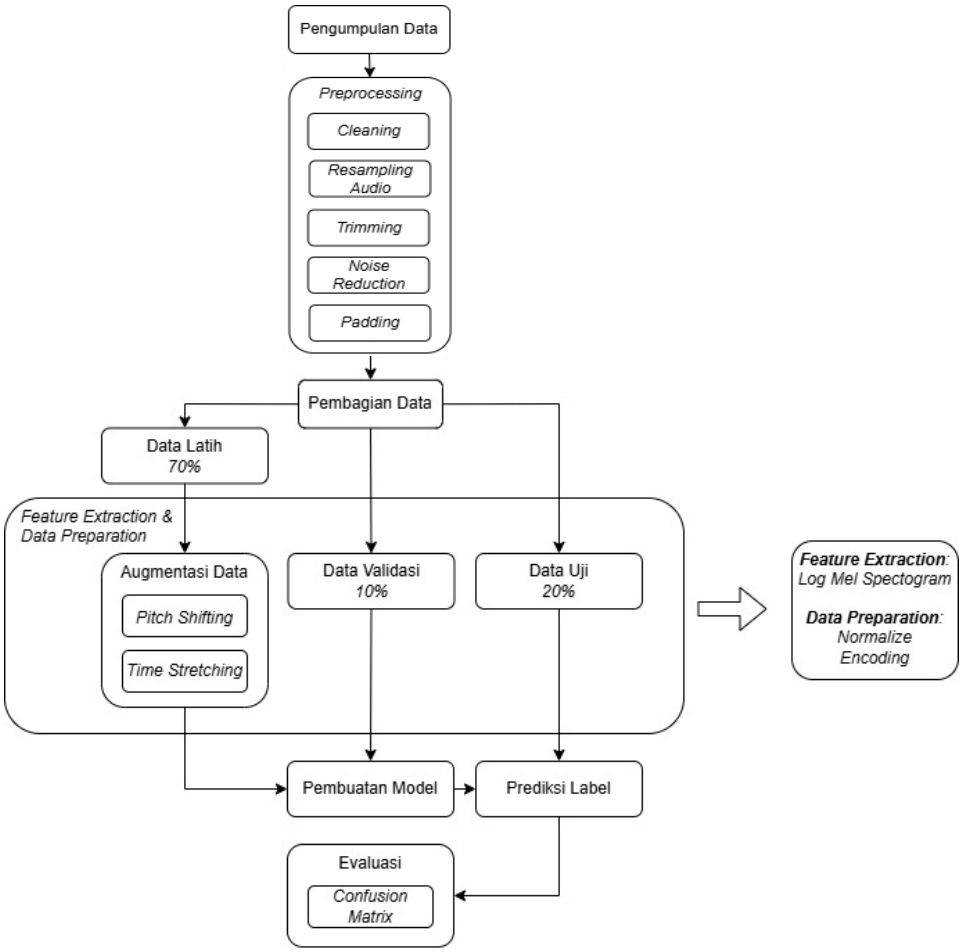
o. *Library* Scikit-learn 1.5.2

Pada penelitian ini, scikit-learn digunakan untuk normalisasi fitur, *encoding* label emosi, pembagian dataset, serta evaluasi model.

3.3. Tahap Penelitian

Tahapan yang akan dilakukan pada penelitian ini dimulai dari pengumpulan data, *preprocessing data*, *feature extraction, data preparation* pembuatan model, dan

evaluasi. Alur dari tahapan yang akan dilakukan pada penelitian ini diilustrasikan pada Gambar 16.



Gambar 16. Tahap Penelitian Implementasi Model CRNN Dengan BiLSTM Untuk Pengenalan Emosi Pada Suara.

3.3.1. Pengumpulan Data

Tahap pertama dalam penelitian ini adalah pengumpulan data, yang dilakukan dengan mengambil dataset dari Kaggle. Kaggle sendiri adalah platform yang menyediakan berbagai dataset dan dapat diakses secara publik. Dataset yang digunakan berjumlah 7442 data dalam bentuk audio dengan format .wav dan memiliki 6 kelas yang berbeda (Marah, Jijik, Takut, Bahagia, Netral, dan Sedih).

3.3.2. *Preprocessing*

Proses *preprocessing* sangat penting untuk memastikan bahwa data yang masuk ke dalam model sudah berada dalam format dan skala yang sesuai, sehingga model dapat belajar dengan lebih efektif dan akurat. Proses ini dilakukan menggunakan operasi numerik dan fungsi dari *library* Librosa. Pada tahap ini, beberapa langkah penting dilakukan untuk memastikan data siap digunakan dalam model, yaitu:

a. *Cleaning Data*

   Cleaning dilakukan untuk membersihkan data agar siap dilatih oleh model. Data diperiksa terlebih dahulu secara manual, jika terdapat data yang senyap atau tidak memiliki suara maka data tersebut akan dibuang.

b. *Resampling Audio*

   Resampling dilakukan untuk menyesuaikan frekuensi sampel suara agar seragam di seluruh dataset. Pada audio dilakukan *resample* ke 44.1 kHz, karena frekuensi ini cukup untuk mempertahankan informasi suara yang relevan tanpa membebani komputasi. Proses ini menggunakan fungsi *resample*() dari Librosa yang memungkinkan perubahan *sampling rate* dengan efisien.

c. *Trimming*

   Trimming dilakukan untuk memotong bagian-bagian suara yang tidak relevan seperti keheningan di awal atau akhir audio. Teknik ini membantu mengurangi *noise*, sehingga fokus analisis hanya pada bagian suara yang penting. Proses ini menggunakan fungsi *trim*() dari Librosa, yang secara otomatis mendeteksi dan menghapus bagian senyap dari audio.

d. *Noise Reduction*

   Noise reduction bertujuan untuk mengurangi gangguan suara latar seperti dengungan, suara angin, atau bising lingkungan. Dengan menggunakan *library* Librosa, suara utama dapat dipisahkan dari *noise*.

e. *Padding*

   Padding digunakan untuk menambahkan durasi audio yang terlalu pendek agar seragam dengan data lainnya. *Padding* sangat penting untuk memastikan

semua data memiliki dimensi *input* yang sama, terutama jika menggunakan algoritma berbasis *deep learning* yang membutuhkan ukuran *input* tetap. *Padding* dilakukan menggunakan fungsi *fix_length*() dari Librosa untuk memastikan semua data memiliki panjang yang sama.

3.3.3. Pembagian Data

Data akan dibagi menjadi data latih, data uji, dan data validasi menggunakan *train_test_split* dari *library* scikit-learn.

a. Data Latih

   Data latih berfungsi sebagai sumber pembelajaran utama bagi model. Pada data ini, model mempelajari pola dan hubungan yang ada sehingga dapat mengenali karakteristik tertentu dalam data tersebut. Dalam penelitian ini, 70% dari total data digunakan untuk data latih, yang memberikan model kesempatan untuk menyerap informasi yang cukup dan meningkatkan akurasi dalam mengenali pola yang dibutuhkan.

b. Data Uji

   Data uji digunakan untuk mengevaluasi performa model pada data yang belum pernah dilihat sebelumnya. Penelitian ini menggunakan 20% dari total data sebagai data uji. Data Uji dapat mengukur seberapa baik model mampu menggeneralisasi atau menerapkan pola yang telah dipelajari dari data latih ke data baru.

c. Data Validasi

   Data validasi digunakan sebagai perantara antara data latih dan data uji untuk menyetel model agar lebih optimal sebelum evaluasi akhir. Dalam penelitian ini, 10% dari data dialokasikan sebagai data validasi. Fungsi utama data validasi adalah untuk menilai performa model secara berkala selama proses pelatihan dan membantu dalam pengaturan *hyperparameter.*

3.3.4. Augmentasi Data

Augmentasi data diterapkan untuk meningkatkan jumlah dan variasi data pelatihan. Augmentasi membantu mengatasi *overfitting* pada model. Teknik yang digunakan adalah *pitch shifting*, yaitu mengubah frekuensi suara tanpa mengubah durasi. Proses ini menaikkan atau menurunkan nada suara, sehingga menghasilkan variasi data yang mencerminkan perbedaan karakteristik vokal. *Time Stretching* juga digunakan untuk mengubah durasi suara tanpa mengubah frekuensinya. Dengan memperpanjang atau memperpendek durasi audio atau suara. Proses augmentasi diterapkan menggunakan fungsi *pitch_shift*() dan *time_stretch*() pada library Librosa. Metode ini membantu model beradaptasi pada suara yang diucapkan lebih cepat atau lebih lambat yang sering terjadi dalam ekspresi emosi tertentu.

3.3.5. *Feature Extraction*

Proses ini secara efektif mentransformasi data audio satu dimensi menjadi sebuah citra dua dimensi yang menangkap informasi frekuensi, intensitas, dan temporal yang berkaitan dengan karakteristik emosi dalam suara. Alur pemrosesan data dimulai dengan sinyal audio yang telah melalui tahap *preprocessing*. Sinyal ini kemudian dianalisis menggunakan *Short-Time Fourier Transform* (STFT), yang memecah sinyal menjadi segmen-segmen waktu yang tumpang untuk dianalisis komponen frekuensinya. Hasil dari STFT adalah sebuah spektrogram linear yang menunjukkan bagaimana energi suara terdistribusi di seluruh spektrum frekuensi dari waktu ke waktu. Selanjutnya, sumbu frekuensi dari spektrogram linear ini dipetakan ke skala Mel, sebuah skala perseptual yang meniru cara telinga manusia mendengar suara. Hasilnya adalah sebuah *Mel spectrogram*. Terakhir, nilai amplitudo atau energi pada *Mel spectrogram* dikonversi menjadi skala logaritmik untuk menyesuaikan dengan dinamika energi suara yang sangat bervariasi. Transformasi akhir ini menghasilkan sebuah *Log-Mel Spectrogram*, yang merupakan representasi visual akhir dari suara dan berfungsi sebagai fitur input untuk model. Seluruh proses perhitungan dan konversi ini diimplementasikan untuk

memastikan fitur yang dihasilkan stabil dan relevan untuk proses pembelajaran mesin.

3.3.6. *Data Preparation*

Tahapan ini bertujuan memastikan data berada dalam format dan skala yang sesuai agar model dapat belajar dengan optimal. Setelah fitur-fitur penting diekstraksi, terdapat dua jenis data yang harus dipersiapkan secara terpisah sebelum dapat dimasukkan ke dalam model, yaitu data fitur dan data label. Alur pertama adalah persiapan untuk data fitur. Nilai-nilai intensitas energi di dalam *Log-Mel Spectrogram* memiliki rentang yang sangat bervariasi. Untuk menyeragamkannya, setiap spectrogram dilewatkan melalui proses Normalisasi menggunakan metode *Z-Score Normalization*. Proses ini mengubah distribusi nilai pada setiap fitur sehingga memiliki rata-rata 0 dan standar deviasi 1. Tujuannya adalah untuk memastikan bahwa tidak ada satu fitur pun yang mendominasi proses pembelajaran secara tidak adil hanya karena memiliki rentang nilai numerik yang lebih besar. Secara paralel, data label yang masih berupa teks juga harus diubah menjadi format numerik yang dapat dipahami oleh model. Proses ini disebut *Encoding*. Label emosi yang bertipe kategorikal ini diubah menjadi representasi numerik menggunakan format vektor *one-hot encoding*, di mana setiap kategori emosi memiliki representasi vektor biner yang unik. Setelah kedua alur proses ini selesai, baik data fitur maupun data label kini berada dalam format yang sinkron dan siap untuk diproses oleh model. Seluruh proses normalisasi dan encoding ini diimplementasikan menggunakan fungsi-fungsi seperti *StandardScaler()* dan *LabelEncoder()*.

3.3.7. Pembuatan Model

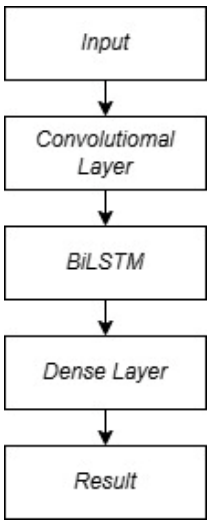
Model yang digunakan pada penelitian ini, yaitu CRNN yang menggabungkan CNN dengan BiLSTM. Arsitektur dan parameter model CRNN yang digunakan oleh Yadav *et al*. (2021) akan digunakan sebagai acuan pengembangan model pada penelitian ini. Model menggunakan lapisan *convolutional layer* diikuti oleh *max pooling* lalu normalisasi untuk mereduksi dimensi. Kemudian, data direshape dan

diproses oleh lapisan BiLSTM yang mengenali urutan temporal dalam audio. Lapisan akhir berupa *Dense* dengan aktivasi *softmax* untuk mengklasifikasikan 6 kelas emosi. Model ini dikompilasi menggunakan *optimizer Adam* dan *sparse categorical crossentropy* sebagai fungsi *loss* untuk klasifikasi multi-kelas. Nilai *hyperparameter* yang digunakan dapat dilihat pada Tabel 6.

Tabel 6. *Hyperparameter* Model CRNN

Hyperparameter	**Nilai**
Convolutional Layer	64, 128
BiLSTM Units	128
Dense	*Softmax*
Optimizer	*Adam*

Dua *callbacks* diterapkan, yaitu *EarlyStopping* untuk menghentikan pelatihan ketika *validation loss* tidak membaik dalam 15 epoch, dan *ReduceLROnPlateau* untuk menurunkan *learning rate* secara bertahap jika *validation loss* stagnan selama 5 epoch, sehingga model dapat beradaptasi selama pelatihan. Model CRNN yang menggabungkan CNN dan BiLSTM dapat dilihat pada Gambar 17.



Gambar 17. Model CRNN yang menggabungkan CNN dan BiLSTM.

3.3.8. Prediksi Label

Setelah model selesai dilatih, tahap prediksi label dilakukan dengan menggunakan Data Uji yang telah melalui seluruh proses persiapan yang sama dengan data latih. Setiap sampel dari Data Uji dimasukkan ke dalam model untuk menghasilkan vektor probabilitas dari *dense layer*, yang menunjukkan tingkat kepercayaan model untuk setiap kelas emosi. Label prediksi akhir ditentukan dengan memilih kelas yang memiliki probabilitas tertinggi dan kumpulan prediksi ini selanjutnya digunakan untuk tahap evaluasi model.

3.3.9. Evaluasi

Pengujian model dilakukan menggunakan *confusion matrix*, yang memberikan gambaran menyeluruh tentang performa klasifikasi model dengan menghitung akurasi, presisi*, recall*, dan *F1-score*. *Confusion matrix* menampilkan jumlah prediksi benar dan salah untuk setiap kelas, memudahkan identifikasi pola kesalahan dalam klasifikasi.

V. SIMPULAN DAN SARAN

5.1. Simpulan

Implementasi model CRNN dengan BiLSTM telah berhasil dilakukan pada tugas pengenalan emosi pada suara menggunakan dataset CREMA-D. Model ini dirancang dengan memanfaatkan arsitektur CNN untuk mengekstraksi fitur spasial dari log-Mel spectrogram, yang kemudian diikuti oleh lapisan BiLSTM untuk menangkap dinamika temporal dan arah sinyal suara, sehingga mampu membangun representasi emosi dari sinyal audio secara efektif. Berdasarkan hasil evaluasi dan analisis, model ini menunjukkan performa yang baik dalam mengklasifikasikan emosi, dengan akurasi validasi tertinggi mencapai 68.23% setelah diterapkan teknik augmentasi. Selain itu, model juga menunjukkan nilai *F1-score* yang tinggi pada emosi tertentu seperti marah dan netral, yang mengindikasikan kemampuannya dalam mengenali emosi-emosi dengan pola spektral yang khas.

5.2. Saran

Saran yang dapat diberikan untuk melanjutkan penelitian ini adalah sebagai berikut:

a. Eksplorasi metode augmentasi lanjutan, seperti *SpecAugment, voice conversion,* atau *GAN-based augmentation*, dapat menjadi solusi untuk menghasilkan variasi data latih yang lebih kaya dan mendekati data nyata.

b. Penambahan fitur lain selain *log-Mel spectrogram*, seperti MFCC, chroma, atau *prosodic features*, dapat membantu model dalam menangkap informasi emosional yang tidak terekam hanya melalui fitur *log-mel*.

DAFTAR PUSTAKA

Ayeni, J.A., 2022. Convolutional neural network (CNN): The architecture and applications. *Applied Journal of Physical Science*, 4(4), pp.42-50.

Bezdan, T. & Bacanin, N., 2019. Convolutional neural network layers and architectures. *Sinteza 2019 - International Scientific Conference on Information Technology and Data Related Research*, 4, pp.445-451.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., & Weiss, B., 2005. A database of German emotional speech. *Proceedings of Interspeech 2005*, pp.1-4.

Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., & Verma, R., 2014. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4), pp.377–390.

Cattermole, K.W., 1965. The Fourier Transform and its Applications. *Electronics and Power*, 11(10), pp. 357-359.

Diana, J.M.E., & Mathivanan, A., 2023. Exploring the Paradigm Shift: Harnessing Data Analytics for Real-World Applications. *International Journal of Science and Research*, 12(6), pp.1467-1480.

Gholizadeh, S., 2022. Top popular python libraries in research. *Journal of Robotics and Automation Research*, 3(2), pp.142-145.

Gokilavani, M., 2023. Ravdness, Crema-D, Tess based algorithm for emotional recognition using speech. *Proceedings of the Fourth International Conference on Smart Systems and Inventive Technology*, pp.1658-1664.

Hayashi, T., Shimizu, T., & Fukami, Y., 2021. Collaborative Problem Solving on a Data Platform Kaggle. *International Journal of Information Technology and Management Information Systems*, 16(2), pp.37-40.

Hawley, S.H., 2013. Fourier Transforms, Audio Engineering, and the Quantum Nature of Reality. Proceedings *AES 135th Convention*, pp.1-4.

Hutagaol, B.J. & Mauritsius, T., 2020. Risk level prediction of life insurance applicant using machine learning. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2), pp.2213-2220.

Karrach, L. & Pivarciova, E., 2023. Using a convolutional neural network for machine written character recognition. *TEM Journal*, 12(3), pp.1252-1259.

Kurnia, D. & Wibowo, A.T., 2021. Klasifikasi spesies tanaman kaktus grafting berdasarkan citra scion menggunakan metode convolutional neural network (CNN). *e-Proceeding of Engineering*, 8(4), pp.4171-4192.

Kolluri, J., Kumar, V., Phridviraj, M.S.B. & Razia, S., 2020. Reducing overfitting problem in machine learning using novel L1/4 regularization method. *Proceedings of the Fourth International Conference on Trends in Electronics and Informatics*, pp. 934-938.

Lee, Y., 2023. The CNN: The architecture behind artificial intelligence development. *Journal of Student Research*, 12(4).

Makhmudov, F., Kutlimuratov, A. & Cho, Y.-I., 2024. Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition. *Applied Sciences*, 14(23), pp.1-19.

Mantrala, S.S., 2025. Python's Pivotal Role in AI and Data Science. *International Journal of Information Technology and Management Information Systems*, 16(2), pp.1676-1686.

Markou, K., Tsochatzidis, L., Zagoris, K., Papazoglou, A., Karagiannis, X., Symeonidis, S. & Pratikakis, I., 2021. A Convolutional Recurrent Neural Network for the Handwritten Text Recognition of Historical Greek Manuscripts. *Lecture Notes in Computer Science*, pp.227-239.

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A. & Doulamis, N., 2021. Multiclass confusion matrix reduction method and its application on net promoter score classification problem. *Technologies*, 9(4), pp.1-22.

Meyer, P., Xu, Z., & Fingscheidt, T., 2021. Improving convolutional recurrent neural networks for speech emotion recognition. *Proceedings of the IEEE Spoken Language Technology Workshop*, pp.1-9.

Mukarram, K.A., Mukhlas, M.A. & Zahra, A., 2024. Enhancing speech emotion recognition with deep learning using multi-feature stacking and data augmentation. *Bulletin of Electrical Engineering and Informatics*, 13(3), pp.1920-1926.

Nugroho, K.S., Akbar, I., Suksmawati, A.N. & Istiadi, 2021. Deteksi depresi dan kecemasan pengguna Twitter menggunakan bidirectional LSTM. *The 4th Conference on Innovation and Application of Science and Technology (CIASTECH 2021)*, pp.287-296.

Patil, T.G., & Zade, A.V., 2023. To Design and Develop Advanced Speech Emotion Recognition Using MLP Classifier with Evolutionary LIBROSA Library. *International Journal for Multidisciplinary Research*, 5(3), pp.1-6.

Pavithra, A., & Prakash, S. M., 2019. TensorFlow in Deep Learning. *International Journal of Innovative Research in Technology*, 5(9), pp.142-147.

Pham, N.T., Dang, D.N.M., Nguyen, N.D., Nguyen, T.T., Nguyen, H., Manavalan, B., Lim, C.P. & Nguyen, S.D., 2023. Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. *Expert Systems With Applications*, 230, pp1-13.

Pothuganti, S., 2018. Review on over-fitting and under-fitting problems in Machine Learning and solutions. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 7(9), pp.3692-3695.

Premjeet, S., Md., Sahidullah. & Goutam, S., 2022. Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Communication*, pp.53-59.

Purwono, P., Ma'arif, A., Rahmaniar, W., Fathurrahman, H.I.K., Frisky, A.Z.K., Haq, Q.M. & Kusuma, A.Z., 2023. Understanding of convolutional neural network (CNN): A review. *International Journal of Robotics and Control Systems*, 2(4), pp.739-748.

Puteri, D.I., 2023. Implementasi long short term memory (LSTM) dan bidirectional long short term memory (BiLSTM) dalam prediksi harga saham syariah. *EULER: Jurnal Ilmiah Matematika, Sains dan Teknologi*, 11(1), pp.35-43.

Qamhan, M.A., Meftah, A.H., Selouani, S., Alotaibi, Y.A. & Seddiq, Y.M., 2020. Speech emotion recognition using convolutional recurrent neural networks and spectrograms. *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering*, pp.1-6.

Reiss, F., Cutler, B., & Eichenberger, Z., 2021. Natural Language Processing with Pandas DataFrames. *Proceedings of the 20th Python in Science Conference*, pp.49-57.

Satrio, M. & Wibowo, A.T., 2021. Klasifikasi tanaman aglaonema berdasarkan citra daun menggunakan metode convolutional neural network (CNN). *E-proceeding of engineering*, 8(5), pp.10621-10636.

Sauter, D.A., Eisner, F., Calder, A.J., & Scott, S.K., 2010. Perceptual cues in non-verbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, 63(11), pp. 2251–2272.

Singh, D. & Singh, B., 2019. Investigating the impact of data normalization on classification performance. *Applied Soft Computing Journal*, pp.1-23.

Vazhenina, D. & Markov, K., 2020. End-to-End Noisy Speech Recognition Using Fourier and Hilbert Spectrum Features. *Electronics*, 9(1157), pp.1-18.

Yadav, O.P., Bastola, L.P. & Sharma, J., 2021. Speech emotion recognition using convolutional recurrent neural network. *Proceedings of the 10th IOE Graduate Conference*, pp.1164-1172.

Zielonka, M., Piastowski, A., Czyzewski, A., Nadachowski, P., Operlejn, M. & Kaczor, K., 2022. Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets. *Electronics*, 11(3831), pp.1-12.