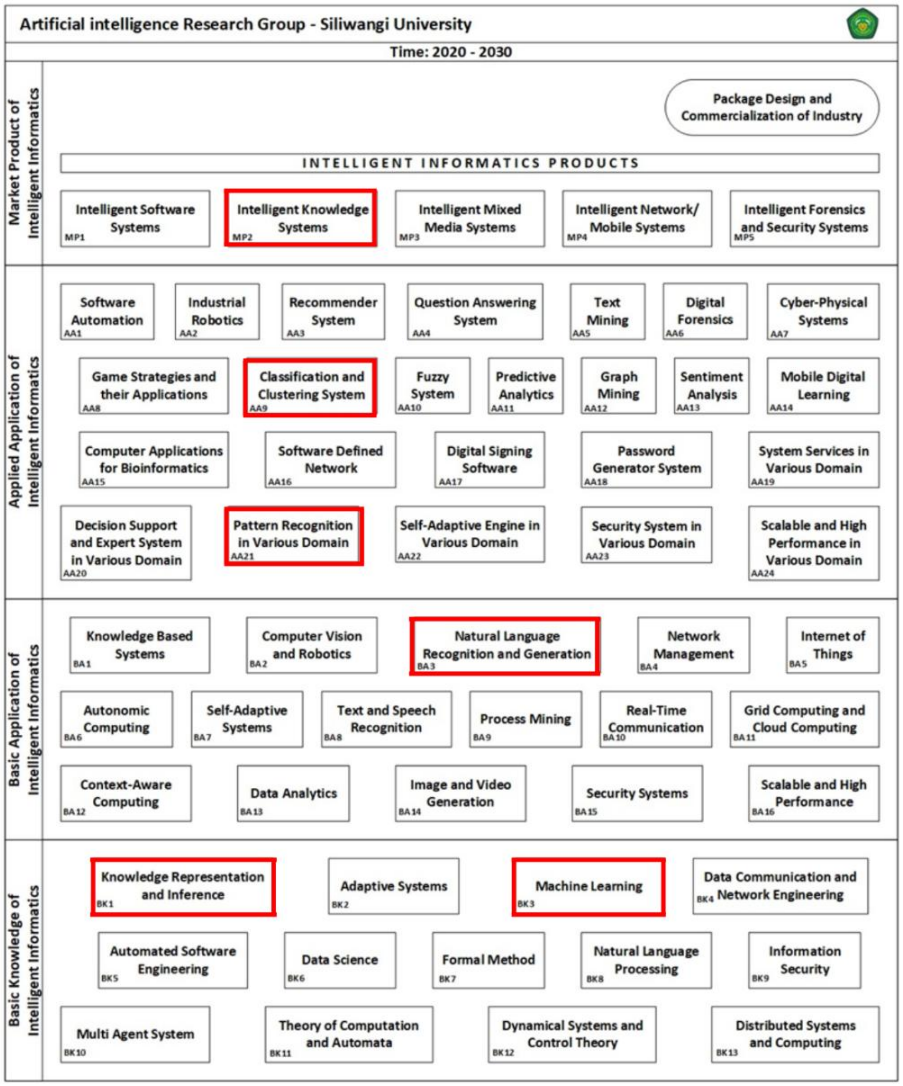
BAB III

METODOLOGI PENELITIAN

3.1 Peta Jalan *(Roadmap)* Penelitian

Penelitian ini mengacu pada **Gambar 3.1** yang merupakan peta jalan penelitian Kelompok Keahlian Informatika dan Sistem Inteligen (ISI), khususnya pada bidang Jaringan, Keamanan, dan Digital Forensik (JKF) di Universitas Siliwangi Tahun 2020-2030.



Gambar 3.1. Peta Jalan Penelitian (AIS Universitas Siliwangi, 2024)

Peta jalan penelitian tersebut menggambarkan arah pengembangan penelitian kecerdasan buatan secara berkelanjutan mulai dari tahapan konsep dasar hingga pengembangan *market product* dalam rentang waktu 2020–2030. Posisi penelitian yang dilakukan berada pada pengembangan sistem kecerdasan buatan berbasis *Speech Emotion Recognition* (SER) untuk mendukung pengolahan dan analisis emosi pada data suara berbahasa Indonesia. Penelitian ini termasuk dalam bidang *Basic Knowledge of Intelligent Informatics*, khususnya pada *Machine Learning* dan *Deep Learning,* karena menggunakan pendekatan pembelajaran mendalam melalui arsitektur *hybrid* CNN-BiLSTM berbasis *Temporal Attention* dalam proses klasifikasi emosi ujaran. Penelitian ini juga menerapkan pendekatan *early fusion* fitur akustik berupa MFCC, *Hilbert Multispectrum,* dan *Cochleagram* untuk menghasilkan representasi emosi yang lebih komprehensif dari sinyal suara. Penelitian ini berfokus pada proses ekstraksi, representasi, dan pembelajaran pola emosional dari data ucapan menggunakan metode *deep learning*.

Penelitian ini termasuk ke dalam bidang *Basic Application of Intelligent Informatics*, khususnya *Speech Processing,* yang berfokus pada pengolahan sinyal suara untuk mengenali emosi berdasarkan karakteristik akustik seperti frekuensi, energi, harmonik, dan pola temporal suara. Penelitian ini menggunakan data ujaran berbahasa Indonesia pada dataset *IndoWaveSentiment* sehingga relevan dalam pengembangan teknologi pemrosesan suara berbasis bahasa lokal. Penelitian ini tidak melakukan analisis semantik maupun linguistik teks, tetapi tetap berkaitan dengan pemrosesan bahasa alami berbasis suara karena

memanfaatkan aspek paralinguistik berupa emosi dalam ujaran sebagai objek analisis.

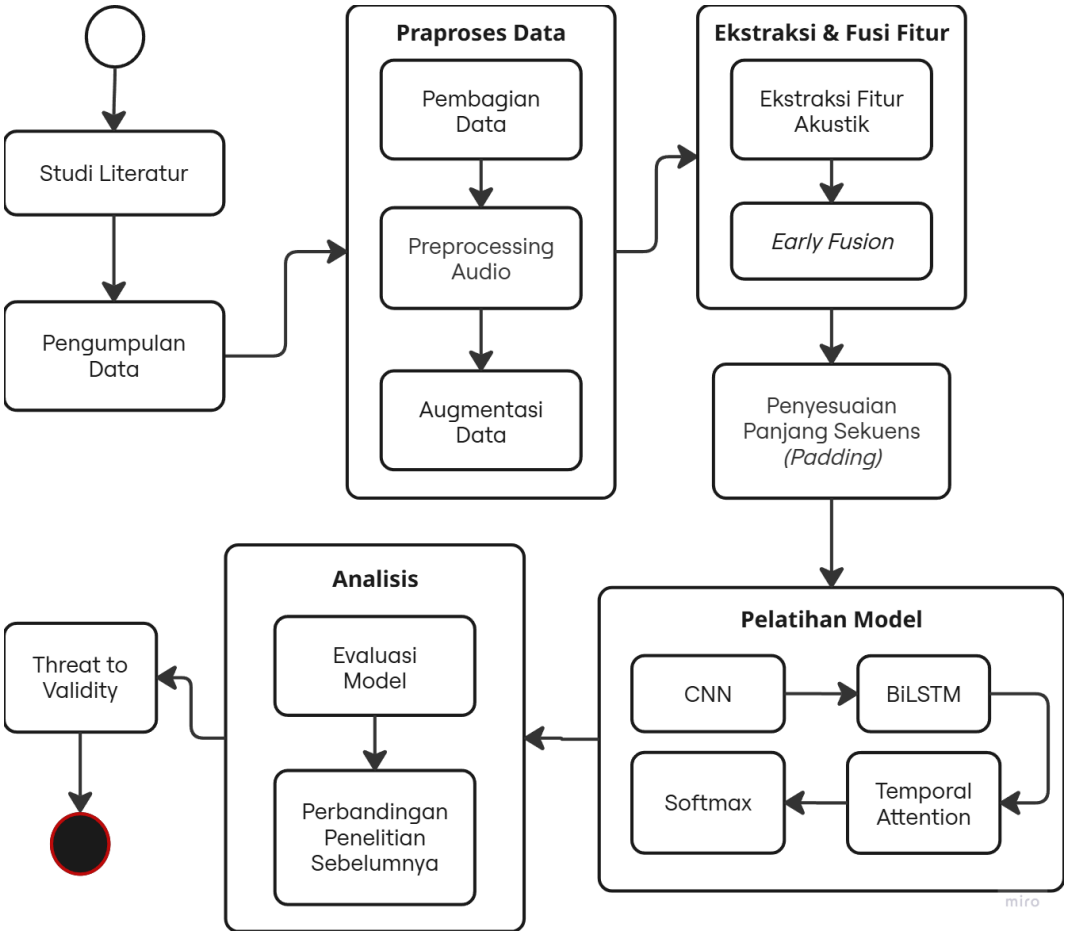
Penelitian ini termasuk ke dalam kategori *Applied Application of Intelligent Informatics*, khususnya *Classification System* dan *Pattern Recognition*. Kategori tersebut tercermin melalui proses klasifikasi beberapa kategori emosi berdasarkan pola akustik sinyal suara menggunakan pendekatan *deep learning*. Penelitian ini juga melibatkan proses pengenalan pola dari kombinasi fitur MFCC, *Hilbert Multispectrum,* dan *Cochleagram* yang digabungkan menggunakan pendekatan *early fusion*. Representasi fitur tersebut kemudian diproses menggunakan arsitektur CNN-BiLSTM berbasis *Temporal Attention* untuk menghasilkan prediksi kelas emosi pada sinyal ujaran.

Penelitian ini termasuk ke dalam *Market Product of Intelligent Informatics*, khususnya *Intelligent Systems* karena sistem yang dikembangkan mampu mengenali dan mengekstraksi informasi emosional dari data suara secara otomatis. Kemampuan tersebut memungkinkan sistem untuk diterapkan pada berbagai pengembangan teknologi berbasis suara, seperti interaksi manusia dan komputer, layanan pelanggan cerdas, asisten virtual, serta sistem analisis emosi berbasis kecerdasan buatan.

3.2 Tahapan Penelitian

Penelitian ini bertujuan untuk membangun model *Speech Emotion Recognition* berbahasa Indonesia yang mampu melakukan klasifikasi emosi berdasarkan sinyal suara ke dalam lima kategori, yaitu netral, senang, terkejut, jijik, dan kecewa. Model yang diusulkan mengintegrasikan *early feature fusion*

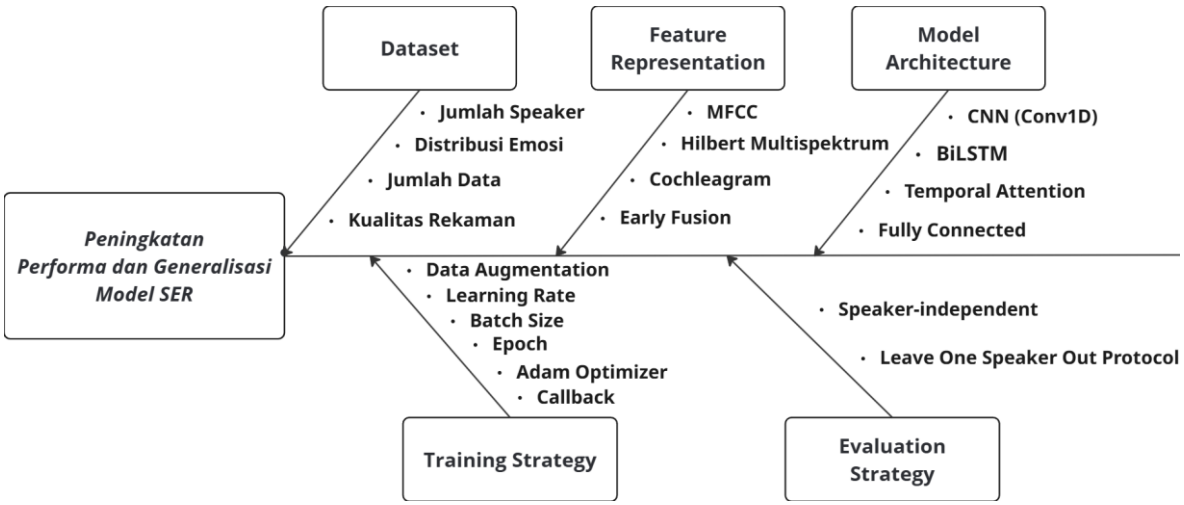
yang menggabungkan fitur MFCC, *Hilbert Multispectrum,* dan Cochleagram dengan arsitektur CNN-BiLSTM berbasis *Temporal Attention*. Tahapan penelitian mengadaptasi *deep learning workflow* dari penelitian (Septiyanto dkk., 2025) dan disesuaikan dengan kebutuhan pengembangan model SER yang berfokus pada peningkatan representasi fitur akustik serta pemodelan karakteristik temporal emosi pada data suara bahasa Indonesia. **Gambar 3.2** menunjukkan rangkaian tahapan penelitian yang dilakukan.



Gambar 3.2. Tahapan Penelitian

Komponen-komponen penting yang memengaruhi peningkatan performa dan kemampuan generalisasi model pada penelitian ini ditunjukkan pada **Gambar**

3.3 dalam bentuk *fishbone* diagram. *Fishbone (Ishikawa diagram)* digunakan untuk mengidentifikasi dan mengelompokkan faktor-faktor yang berpotensi memengaruhi suatu hasil secara sistematis berdasarkan hubungan sebab-akibat (Kumah dkk., 2024). *Fishbone* pada penelitian ini berbeda dengan tahapan penelitian karena memetakan lima kelompok faktor utama, yaitu Dataset (jumlah *speaker,* distribusi emosi, jumlah data, dan kualitas rekaman) (Bustamin dkk., 2024), *Feature Representation* (MFCC, *Hilbert Multispectrum, Cochleagram,* dan *early fusion*) (Shixin dkk., 2024), *Model Architecture* (CNN Conv1D, BiLSTM, *Temporal Attention,* dan *Fully Connected*) (Peng dkk., 2023), *Training Strategy* (*data augmentation, learning rate, batch size, epoch, Adam Optimizer*, dan *callback*) (Septiyanto dkk., 2025), serta *Evaluation Strategy* yang menerapkan skema *speaker-independent* melalui *Leave-One-Speaker-Out* (LOSO) *Cross Validation* untuk mengevaluasi kemampuan generalisasi model terhadap pembicara yang belum pernah digunakan selama pelatihan (C. Lu dkk., 2024).



Gambar 3.3. *Fishbone* Faktor Pengaruh Performa dan Generalisasi SER

3.2.1 Studi Literatur

Studi literatur dan riset pendahuluan merupakan tahapan paling awal dari penelitian ini yang dilakukan dengan tujuan mencari informasi dan referensi ilmiah mengenai *speech emotion recognition* seperti definisi, penerapan SER dalam berbagai bidang, dan metode *deep learning* yang sering digunakan dalam sistem SER. Kajian teori juga dilakukan untuk memahami *domain* penelitian dan solusi yang pernah dilakukan dari penelitian-penelitian sebelumnya. Tahapan ini menghasilkan landasan konseptual yang digunakan sebagai dasar dalam penyusunan metodologi penelitian.

3.2.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *IndoWaveSentiment*, yaitu dataset audio berbahasa Indonesia yang bersifat *open-access* dan dipublikasikan pada 24 Juni 2024. Dataset ini dirancang untuk klasifikasi emosi berbasis suara dengan lima kategori emosi, yaitu netral, senang, terkejut, jijik, dan kecewa. Seluruh data disimpan dalam format .wav dan terdiri dari 300 rekaman audio yang dihasilkan oleh 10 aktor (5 pria dan 5 wanita). Setiap aktor merekam lima emosi dengan dua tingkat intensitas (normal dan strong) serta tiga kali pengulangan, sehingga distribusi data antar pembicara bersifat seimbang. Seluruh aktor mengucapkan kalimat yang sama, yaitu “Kualitas HP ini cukup bagus”, sehingga variasi utama pada data berasal dari ekspresi emosional dan karakteristik akustik pembicara, bukan dari perbedaan teks.

Penamaan berkas audio seperti yang ditunjukkan pada **Tabel 3.1** menggunakan format “Responden–Emosi–Intensitas–Pengulangan.wav” yang

mempermudah proses identifikasi label secara sistematis. Kode responden menunjukkan identitas aktor, kode emosi merepresentasikan lima kategori emosi, kode intensitas menunjukkan level normal atau *strong*, dan kode pengulangan menunjukkan jumlah pengulangan (Bustamin dkk., 2024).

Tabel 3.1. Aturan Penamaan *File* Dataset *IndoWaveSentiment*

Format Penomoran	Penanda	Deskripsi
01 – 10	Aktor (Responden)	Angka ganjil menandakan aktor pria dan angka genap menandakan aktor wanita.
01 – 05	Emosi	(01) netral, (02) senang, (03) terkejut, (04) jijik, (05) kecewa.
01 – 02	Intensitas	(01) *normal*, (02) *strong*
01 – 03	Pengulangan	(01) satu kali pengulangan, (02) dua kali pengulangan, (03) tiga kali pengulangan.

3.2.3 Pembagian Data Menggunakan *Leave-One-Speaker-Out* (LOSO)

Penelitian ini menggunakan metode validasi *Leave-One-Speaker-Out* (LOSO) untuk membagi data latih dan data uji. LOSO merupakan salah satu strategi *cross-validation* yang banyak diterapkan pada penelitian *Speech Emotion Recognition* (SER) karena digunakan untuk mengevaluasi kemampuan generalisasi model dalam mengenali emosi dari pembicara yang tidak pernah dilibatkan selama proses pelatihan model (Sajjad & Kwon, 2020).

Metode LOSO menggunakan data dari satu *speaker* sebagai data uji *(test set),* sedangkan data dari *speaker* lainnya digunakan sebagai data latih *(training set)*. Proses tersebut dilakukan secara berulang hingga seluruh *speaker* memperoleh kesempatan menjadi data uji sebanyak satu kali. Dataset *IndoWaveSentiment* yang digunakan dalam penelitian ini terdiri atas 10 *speaker* (Bustamin dkk., 2024) sehingga skema LOSO menghasilkan sepuluh *fold* validasi.

Setiap *fold* LOSO meliputi berbagai tahapan seperti praproses data, augmentasi, ekstraksi fitur, fusi fitur, pelatihan model, dan evaluasi dilakukan secara independen. Validitas eksperimen dijaga dengan menerapkan augmentasi hanya pada data latih sehingga tidak terjadi *data leakage* antar data latih dan data uji (Septiyanto dkk., 2025).

Tabel 3.2. Mekanisme *Leave-One-Speaker-Out* (LOSO)

Fold	Data Uji (Speaker)	Data Latih (Speaker)
1	Speaker 1	Speaker 2, 3, 4, 5, 6, 7, 8, 9, 10
2	Speaker 2	Speaker 1, 3, 4, 5, 6, 7, 8, 9, 10
3	Speaker 3	Speaker 1, 2, 4, 5, 6, 7, 8, 9, 10
4	Speaker 4	Speaker 1, 2, 3, 5, 6, 7, 8, 9, 10
5	Speaker 5	Speaker 1, 2, 3, 4, 6, 7, 8, 9, 10
6	Speaker 6	Speaker 1, 2, 3, 4, 5, 7, 8, 9, 10
7	Speaker 7	Speaker 1, 2, 3, 4, 5, 6, 8, 9, 10
8	Speaker 8	Speaker 1, 2, 3, 4, 5, 6, 7, 9, 10
9	Speaker 9	Speaker 1, 2, 3, 4, 5, 6, 7, 8, 10
10	Speaker 10	Speaker 1, 2, 3, 4, 5, 6, 7, 8, 9

3.2.4 *Preprocessing Audio*

Tahap *preprocessing audio* dalam penelitian ini bertujuan untuk menstandarkan kualitas sinyal suara sebelum dilakukan ekstraksi fitur, sehingga informasi yang diperoleh lebih representatif. Proses dimulai dengan memuat audio menggunakan *sampling rate* sebesar 16 kHz untuk menjaga konsistensi antar data. Normalisasi amplitudo kemudian dilakukan agar rentang sinyal berada pada skala yang seragam dan tidak dipengaruhi oleh perbedaan volume rekaman. Proses *silence trimming* kemudian diterapkan untuk menghilangkan bagian hening di awal dan akhir sinyal berdasarkan ambang tertentu, sehingga hanya bagian ucapan yang relevan yang diproses lebih lanjut. Penelitian ini juga menerapkan *pre-emphasis filter* untuk memperkuat komponen frekuensi tinggi pada sinyal suara

sehingga karakteristik akustik yang berkaitan dengan emosi dapat direpresentasikan dengan lebih baik. Rangkaian *preprocessing* ini bertujuan untuk mengurangi *noise* yang tidak informatif serta meningkatkan kualitas sinyal sebagai input pada tahap ekstraksi fitur akustik (Cherukuru & Mustafa, 2024).

3.2.5 Augmentasi Data

Tahap berikutnya adalah augmentasi data pada sinyal audio. Augmentasi diterapkan untuk meningkatkan variasi akustik data latih dan membantu model menjadi lebih *robust* terhadap variasi suara yang mungkin muncul pada kondisi nyata. Penelitian ini menggunakan lima teknik augmentasi, yaitu *Additive Gaussian Noise, Speed Perturbation, Time Stretch, Time Shift,* dan *Pitch Shift* (Septiyanto dkk., 2025).

Parameter yang digunakan pada masing-masing teknik augmentasi disajikan pada **Tabel 3.3.** *Additive Gaussian Noise* diterapkan dengan menambahkan *noise* acak sebesar 0,005 × hp.random.randn(len(y)) untuk mensimulasikan gangguan lingkungan. *Time stretch* menggunakan nilai *rate* yang dipilih secara acak pada rentang 0,8-1,2 sehingga durasi audio dapat diperlambat hingga 0,8 kali atau dipercepat hingga 1,2 kali dari kecepatan aslinya tanpa mengubah tinggi nada. *Pitch Shift* mengubah tinggi nada audio secara acak pada rentang -3 hingga +3 *semitone*. *Time shift* menggeser posisi sinyal audio menggunakan operasi *np.roll* dengan besar pergeseran acak sebesar -20% hingga +20% dari panjang sinyal. *Speech perturbation* mengubah kecepatan pengucapan dengan faktor acak pada rentang 0,9-1,1 untuk menghasilkan variasi tempo bicara.

Tabel 3.3. Parameter Augmentasi Audio (Septiyanto dkk., 2025)

Teknik Augmentasi	Parameter	Keterangan
Additive Gaussian Noise	0.005 * *np.random.randn*(len(y))	Menambahkan *noise* acak *(Gaussian noise)* ke dalam sinyal audio untuk mensimulasikan gangguan lingkungan
Time Stretch	*rate* = *np.random.uniform*(0.8, 1.2)	Audio dapat diperlambat hingga 0.8× atau dipercepat hingga 1.2× dari kecepatan asli untuk mensimulasikan variasi tempo bicara.
Pitch Shift	steps = *np.random.uniform*(-3, 3)	Mengubah *pitch audio* pada rentang -3 hingga +3 *semitone.*
Time Shift	*shift* = *np.random.uniform*(-0.2, 0.2) * len(y)	Menggeser posisi sinyal audio secara acak menggunakan operasi *np.roll*
Speed Pertubation	*speed* = *np.random.uniform*(0.9, 1.1)	Audio diperlambat hingga 0.9× atau dipercepat hingga 1.1× untuk memberikan variasi kecepatan pengucapan.

Augmentasi hanya diterapkan pada data latih pada setiap *fold* LOSO untuk menghindari *data leakage,* sedangkan data uji tetap menggunakan audio asli. Setiap sampel audio pada data latih mengalami dua kali proses augmentasi, di mana pada setiap proses satu teknik augmentasi dipilih secara acak dari lima teknik yang tersedia dan diterapkan langsung pada sinyal audio asli. Setiap sampel audio menghasilkan tiga data pelatihan, yaitu satu data asli dan dua data hasil augmentasi. Jumlah data latih pada setiap *fold* LOSO meningkat dari 270 audio menjadi 810 audio sebelum digunakan pada proses pelatihan model.

3.2.6. Ekstraksi dan Representasi Fitur Akustik

Tahap ekstraksi fitur akustik bertujuan untuk merepresentasikan sinyal audio hasil *preprocessing* ke dalam bentuk numerik yang mampu menangkap karakteristik emosional suara secara efektif. Penelitian ini menggunakan tiga jenis fitur, yaitu *Mel-Frequency Cepstral Coefficients* (MFCC), *Hilbert Multispectrum*, dan *Cochleagram* yang diekstraksi secara paralel dari setiap sampel audio sebelum dikombinasikan menggunakan pendekatan *early fusion*. MFCC

digunakan untuk merepresentasikan karakteristik spektral berbasis persepsi pendengaran manusia dan banyak digunakan pada penelitian SER (Septiyanto dkk., 2025). *Hilbert Multispectrum* digunakan untuk merepresentasikan perubahan energi dan amplitudo sinyal melalui transformasi *Hilbert* sehingga mampu menangkap dinamika temporal dan informasi modulasi emosi pada suara. *Cochleagram* digunakan untuk merepresentasikan karakteristik pendengaran biologis yang menyerupai mekanisme koklea manusia sehingga dapat menggambarkan pola harmonik dan distribusi energi suara secara lebih detail. Ketiga fitur tersebut dipilih karena memiliki karakteristik representasi yang saling melengkapi, sehingga diharapkan mampu meningkatkan kualitas representasi emosi dibandingkan penggunaan fitur tunggal (Gumelar dkk., 2020; Peng dkk., 2023).

Konfigurasi parameter ekstraksi fitur pada penelitian ini mengacu pada penelitian Septiyanto dkk. (2025) agar proses ekstraksi fitur konsisten dengan penelitian yang menjadi acuan pengembangan model. Penggunaan 40 koefisien MFCC (n_mfcc = 40) diadopsi dari penelitian tersebut karena mampu mempertahankan informasi spektral yang lebih kaya dibandingkan konfigurasi MFCC konvensional dalam tugas *Speech Emotion Recognition*. Parameter lainnya, seperti *sampling rate* 16 kHz, koefisien *pre-emphasis* sebesar 0,97, serta konfigurasi STFT dengan n_fft = 1024 dan hop_length = 512, juga mengikuti konfigurasi yang digunakan pada penelitian acuan sehingga representasi fitur yang dihasilkan tetap konsisten sebelum dilakukan proses *early fusion*. Parameter ekstraksi fitur yang digunakan pada penelitian ini disajikan pada **Tabel 3.4**.

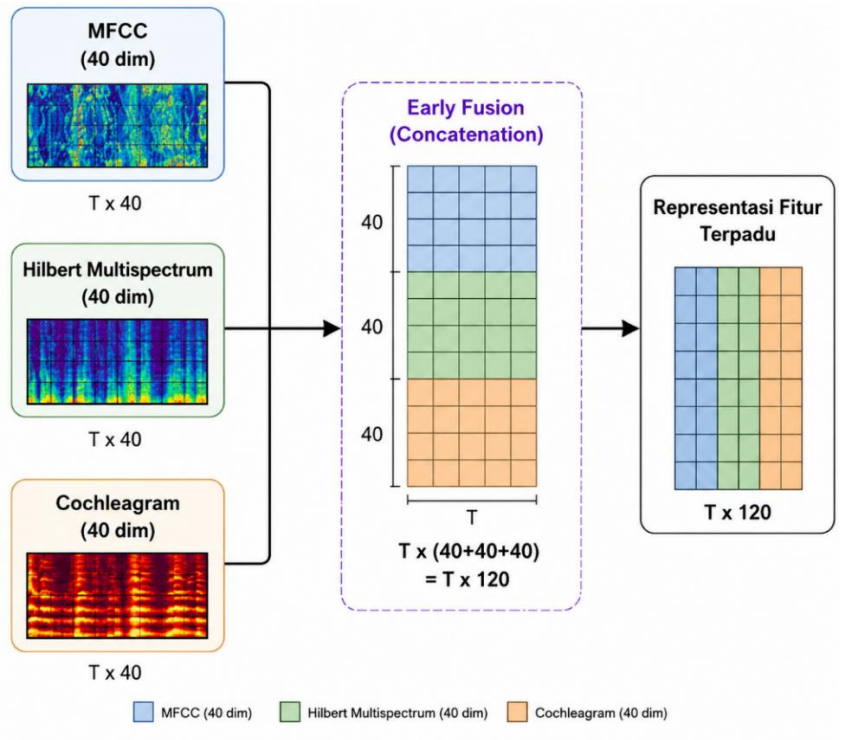
Tabel 3.4. Parameter Ekstraksi Fitur Akustik

Fitur Akustik	Parameter	Nilai
MFCC	n_mfcc	40
Hilbert Multispectrum	*n_features*	40
Cochleagram	n_bins	40
Audio Sampling Rate	sr	16000 Hz
Pre-emphasis	Koefisien	0.97
STFT	n_fft	1024
STFT	*hop_length*	512

3.2.7. Fusi Fitur Akustik

Penggunaan satu jenis fitur akustik dalam pengenalan emosi suara sering kali belum mampu merepresentasikan informasi emosional secara optimal karena setiap fitur seperti MFCC, *Hilbert Multispectrum,* dan *Cochleagram* memiliki karakteristik yang berbeda. Pendekatan fusi fitur digunakan untuk menggabungkan berbagai fitur guna menghasilkan representasi yang lebih komprehensif dan informatif (Zhang dkk., 2020).

Penelitian ini menggunakan jenis fusi fitur *early fusion.* Visualisasi *early fusion* yang diterapkan pada skema kombinasi tiga fitur (MFCC, *Hilbert Multispectrum,* dan *Cochleagram*) pada **Gambar 3.4** menunjukkan bahwa ketiga fitur digabungkan di awal dengan cara konkatenasi pada dimensi fitur. Hasil penggabungan ini kemudian menjadi input tunggal bagi setiap arsitektur model *deep learning* yang digunakan. *Early fusion* memiliki karakteristik menggabungkan informasi sejak awal sehingga representasi fitur menjadi lebih kaya (Sekkate dkk., 2019).



Gambar 3.4. Pendekatan *Early Fusion*

Setiap jenis fitur diekstraksi menjadi 40 dimensi sehingga seluruh fitur memiliki ukuran representasi yang seragam sebelum dilakukan proses *early fusion*. Penyamaan dimensi ini bertujuan agar setiap fitur memberikan kontribusi yang seimbang pada proses konkatenasi tanpa mendominasi fitur lainnya. Penggunaan 40 koefisien MFCC juga telah digunakan pada penelitian *Speech Emotion Recognition* pada penelitian sebelumnya karena mampu mempertahankan informasi spektral yang lebih kaya dibandingkan konfigurasi MFCC konvensional (Septiyanto dkk., 2025). Proses *early fusion* menghasilkan representasi akhir berdimensi 120 yang berasal dari penggabungan tiga vektor fitur berdimensi 40 (40 + 40 + 40).

3.2.8. Pelatihan Model

Proses pelatihan model pada penelitian ini menggunakan pendekatan

arsitektur *hybrid* CNN-BiLSTM dengan *Temporal Attention* seperti yang ditunjukkan pada **Gambar 3.5.**

Model ini dikembangkan berdasarkan pendekatan *baseline* BiLSTM pada penelitian (Septiyanto dkk., 2025) yang memanfaatkan jaringan BiLSTM untuk mempelajari dependensi temporal pada sinyal ucapan dalam *Speech Emotion Recognition.*

Arsitektur tersebut dimodifikasi dengan menambahkan blok CNN sebagai ekstraktor pola lokal serta mekanisme *Temporal Attention* untuk meningkatkan fokus model terhadap bagian sinyal suara yang mengandung informasi emosional paling dominan. Pendekatan *hybrid* CNN-BiLSTM dengan *attention* dipilih karena mampu mengombinasikan kemampuan ekstraksi fitur spasial dan temporal secara lebih efektif pada data audio emosional sebagaimana diterapkan pada penelitian (Zennou, 2026).

Model menerima masukan berupa sequence fitur hasil *early fusion* yang terdiri atas MFCC, *Hilbert Multispectrum*, dan *Cochleagram*. Blok CNN digunakan untuk mengekstraksi pola lokal dan karakteristik spektral dari sinyal audio. Parameter CNN-BiLSTM yang digunakan dalam pelatihan ini ditunjukkan pada **Tabel 3.5**.

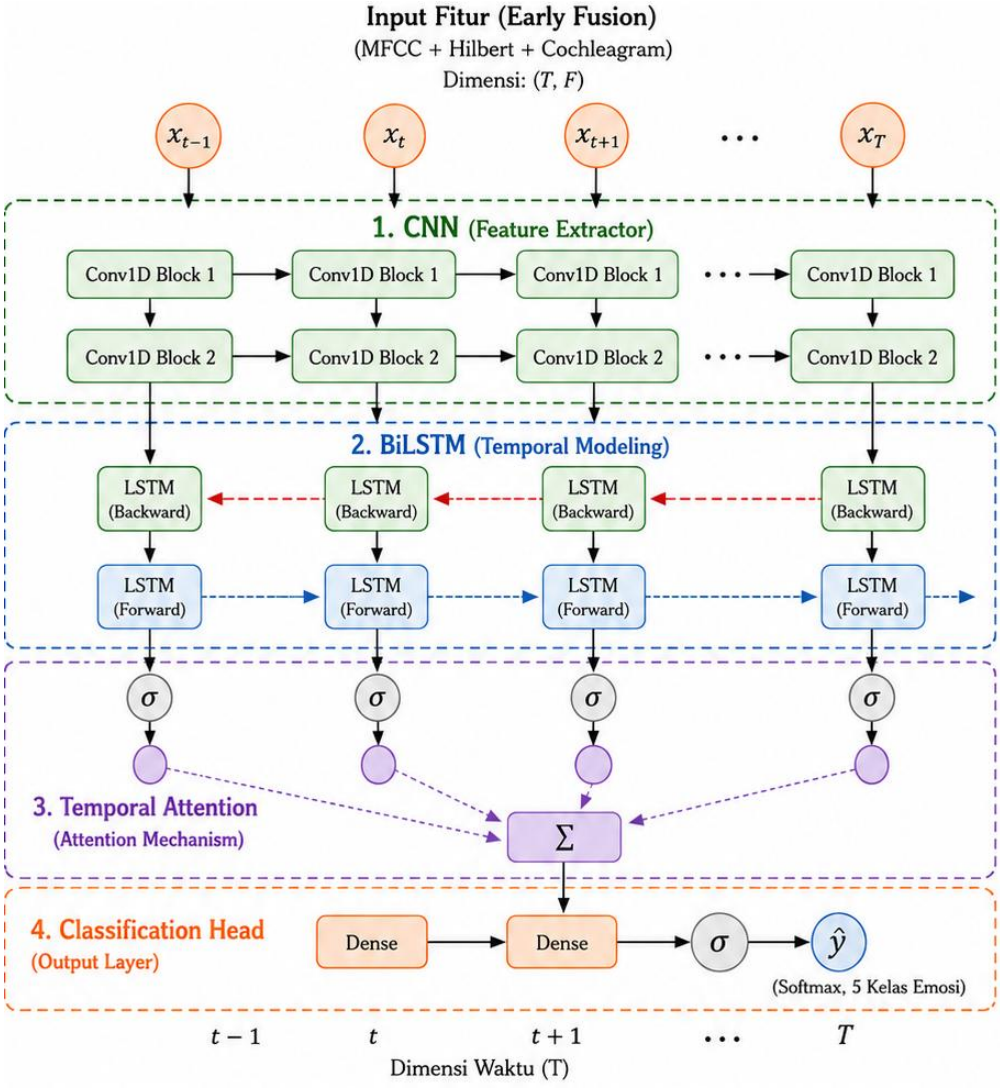
Tabel 3.5. Parameter Implementasi CNN-BiLSTM

Layer	***Output Shape***	**Konfigurasi**
Input Layer	(119, 120)	*Early Fusion*
Conv1D	(119, 128)	Kernel size = 5
Batch Normalization	(119, 128)	-
MaxPooling1D	(59, 128)	Pool size = 2
Dropout	(59, 128)	0.3
Conv1D	(59, 256)	Kernal size = 3
Batch Normalization	(59, 256)	-
MaxPooling1D	(29, 256)	Pool size = 2
Dropout	(29, 256)	0.3
BiLSTM (128)	(29, 256)	128 units
Batch Normalization	(29, 256)	-
Dropout	(29, 256)	0.3
BiLSTM (64)	(128)	64 units
Batch Normalization	(128)	-
Dropout	(128)	0.3

BiLSTM berfungsi mempelajari hubungan temporal dua arah pada urutan sinyal ucapan sehingga representasi emosi dapat dipahami secara lebih kontekstual. Mekanisme *Temporal Attention* diterapkan untuk memberikan bobot perhatian pada bagian temporal yang dianggap paling penting dalam proses pengenalan emosi. Representasi fitur hasil *attention* kemudian diteruskan menuju lapisan klasifikasi untuk menghasilkan prediksi lima kelas emosi yaitu *neutral, happy, surprise, disgust,* dan *disappointed.* Pendekatan ini diharapkan mampu meningkatkan kemampuan model dalam mengenali karakteristik emosional pada data suara berbahasa Indonesia secara lebih akurat dan adaptif. Parameter *Temporal Attention* yang digunakan pada penelitian ini ditunjukkan pada **Tabel 3.6** dan arsitektur pemodelan ditunjukkan pada **Gambar 3.5.**

Tabel 3.6. Parameter *Temporal Attention*

Layer	***Output Shape***	**Konfigurasi**
Dense Attention	(29, 1)	Dense(1), tanh
Flatten	(29)	*Flatten attention score*
Softmax Attention	(29)	*Attention weight*
Repeat Vector	(128, 29)	*Repeat attention weight*
Permute	(29, 128)	*Align attention dimension*
Multiply	(29, 128)	*Weighted feature representation*
Context Vector	(128)	*Sum over temporal dimension*



Gambar 3.5. Arsitektur CNN-BiLSTM dengan *Temporal Attention* (Dimodifikasi dari (Septiyanto dkk., 2025))

Proses pelatihan model dilakukan menggunakan *Adam optimizer* dengan *learning rate* sebesar 0.001. *Optimizer Adam* digunakan karena mampu melakukan penyesuaian bobot secara adaptif dan memiliki performa yang stabil pada proses pelatihan model *deep learning* berbasis *sequence* audio (Kingma & Ba, 2017). Fungsi *loss* yang digunakan adalah *Sparse Categorical Crossentropy* karena penelitian ini menggunakan klasifikasi multikelas dengan label emosi berbentuk integer hasil proses *label encoding*. Pelatihan model dilakukan menggunakan *batch size* sebesar 32 dan maksimum 150 *epoch*. Parameter pelatihan model yang diterapkan pada penelitian ini disajikan pada **Tabel 3.7.**

Tabel 3.7. Parameter Pelatihan Model
(Dimodifikasi dari (Septiyanto dkk., 2025))

No.	Parameter	Nilai/Konfigurasi
1.	*Optimizer*	Adam
2.	*Learning Rate*	0.001
3.	*Loss Function*	*Sparse Categorical Crossentropy*
4.	*Batch Size*	32
5.	*Epoch Maximum*	150
6.	*EarlyStopping*	Patience = 20
7.	*Restore Best Weights*	True
8.	*ReduceLROnPlateau*	Factor = 0.5, *Patience* = 5
9.	*Model Checkpoint*	*Save Best Model*
10.	*Validation Strategy*	*Leave-One-Speaker-Out* (LOSO)
11.	*Dropout* CNN & BiLSTM	0.3
12.	*Dropout Dense Layer*	0.5
13.	*Random Seed*	42
14.	*Metrics Evaluation*	*Accuracy, Precision, Recall, F1-Score*
15.	*Framework*	TensorFlow/Keras

Stabilitas proses pelatihan ditingkatkan melalui penerapan beberapa *callback* yaitu *EarlyStopping*, *ReduceLROnPlateau*, dan *ModelCheckpoint*. *EarlyStopping* digunakan untuk menghentikan proses pelatihan secara otomatis apabila performa model *tidak* mengalami peningkatan selama sejumlah *epoch* tertentu, sedangkan *ReduceLROnPlateau* digunakan untuk menurunkan learning

rate secara adaptif ketika proses pelatihan mengalami stagnasi. *ModelCheckpoint* digunakan untuk menyimpan bobot model terbaik yang diperoleh selama proses pelatihan. Seluruh proses pelatihan dan evaluasi model dilakukan pada setiap *fold* validasi *Leave-One-Speaker-Out* (LOSO), sehingga setiap speaker memperoleh kesempatan menjadi data uji sebanyak satu kali.

3.2.9. Evaluasi Model

Evaluasi model pada penelitian ini dilakukan untuk mengukur kemampuan model dalam mengklasifikasikan emosi suara berdasarkan fitur audio hasil *feature fusion*. Proses evaluasi dilakukan menggunakan skema *Leave-One-Speaker-Out* (LOSO) *Cross Validation* yang telah dijelaskan pada subab sebelumnya. **Tabel 3.8**, menyajikan metrik evaluasi yang digunakan meliputi *accuracy, precision, recall, F1-score, confusion matrix* yang umum digunakan pada penelitian *Speech Emotion Recognition* (SER) *modern* untuk mengevaluasi performa klasifikasi emosi secara menyeluruh (Wang, 2024).

Tabel 3.8. Metrik Evaluasi Model

No.	Metrik	Fungsi Evaluasi
1.	*Accuracy*	Mengukur proporsi prediksi benar terhadap seluruh data
2.	*Precision*	Mengukur ketepatan prediksi pada setiap kelas
3.	*Recall*	Mengukur kemampuan model mendeteksi kelas sebenarnya
4.	*F1-Score*	Mengukur keseimbangan antara *precision* dan *recall*
5.	*Confusion Matrix*	Menganalisis distribusi prediksi tiap kelas

Confusion matrix juga digunakan untuk menganalisis distribusi prediksi dan pola kesalahan klasifikasi antar kategori emosi. Analisis tambahan juga dilakukan berdasarkan performa setiap speaker dan setiap kategori emosi untuk mengidentifikasi tingkat generalisasi model serta emosi yang paling mudah maupun paling sulit dikenali. Nilai akhir performa sistem disajikan dalam bentuk rata-rata, standar deviasi, dan interval kepercayaan 95% dari seluruh fold LOSO

guna memberikan gambaran mengenai stabilitas model yang diusulkan.

3.2.10. *Threat to Validity*

Analisis validasi dilakukan terhadap hasil eksperimen yang telah dilakukan. Analisis ini bertujuan untuk memahami serta mengidentifikasi potensi bias atau keterbatasan yang dapat mempengaruhi hasil penelitian, mulai dari tahap pengumpulan data hingga proses pengembangan model. Analisis dilakukan menggunakan pendekatan *threats to validity* guna memastikan bahwa hasil yang diperoleh dapat diinterpretasikan secara tepat.

Threats to validity merupakan faktor-faktor yang dapat mempengaruhi keakuratan dan keandalan hasil penelitian. Ancaman terhadap validitas dalam konteks penelitian berbasis *machine learning* pada umumnya dikategorikan menjadi *internal validity, construct validity, conclusion validity,* dan *external validity* (Anglin, 2024; Barbierato & Gatti, 2024). *Internal validity* berkaitan dengan potensi bias selama proses eksperimen, seperti penerapan augmentasi data, pemilihan *hyperparameter*, dan proses pelatihan model. *Construct validity* berhubungan dengan kesesuaikan metode ekstraksi fitur, arsitektur model, serta metrik evaluasi yang digunakan untuk merepresentasikan kemampuan pengenalan emosi suara. *Conclusion validity* berkaitan dengan ketepatan interpretasi hasil evaluasi dan konsistensi performa model yang diperoleh dari seluruh *fold* LOSO. *External validity* mengacu pada kemampuan generalisasi model terhadap data dan karakteristik *speaker* yang belum pernah dilihat selama proses pelatihan, serta keterbatasan penggunaan satu dataset, yaitu *IndoWaveSentiment,* dalam penelitian ini.