Sistem Deteksi Emosi Manusia Menggunakan Metode *Audio-Visual* Berbasis MFCC, *Random Forest*, dan CNN

Laporan Tugas Akhir

Oleh:
M. Afif Fuadie Yusuf (4212531004)

Program Studi Teknik Mekatronika
Jurusan Teknik Elektro
Politeknik Negeri Batam
2025

Pernyataan Keaslian Tugas Akhir

Saya yang bertandatangan dibawah ini menyatakan bahwa isi sebagian maupun keseluruhan Tugas Akhir saya yang berjudul : "Sistem Deteksi Emosi Manusia Menggunakan Metode *Audio-Visual* Berbasis MFCC, *Random Forest* dan *CNN*" adalah **hasil karya sendiri**, **diselesaikan tanpa menggunakan bahan-bahan yang tidak diizinkan**, dan **bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri.** Semua referensi yang dikutip atau dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan saya ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Batam, 17 Juli 2026

M. Afif Fuadie Yusuf

NIM: 4212531004

Lembar Pengesahan

Laporan Tugas Akhir disusun untuk digunakan sebagai rencana kerja pada pelaksanaan Tugas Akhir

Disusun oleh:

M. Afif Fuadie Yusuf (4212531004)

Tanggal Seminar: 07 Juli 2026

Disetujui oleh :

1. Diono, S.Tr.T., M.Sc

NIK: 120243

1. Feralia Fitri,S.T, M.T

NIK: 125364

2. Sulistiono, S.T., M.T

NIK: 125367

Sistem Deteksi Emosi Manusia Menggunakan Metode *Audio - Visual* Berbasis MFCC, *Random Forest*, dan *CNN*

Abstrak

Perkembangan teknologi kecerdasan buatan memberikan peluang bagi sistem komputer untuk mengenali dan memahami emosi manusia secara otomatis. Namun, sebagian besar sistem deteksi emosi hanya memanfaatkan satu modalitas, seperti suara atau citra wajah, sehingga performanya masih dipengaruhi oleh kondisi lingkungan, seperti kebisingan dan pencahayaan. Penelitian ini bertujuan mengembangkan sistem Audio-Visual Emotion Recognition (AVER) yang mampu mendeteksi emosi secara real-time menggunakan kombinasi informasi suara dan citra wajah. Pada sistem audio digunakan metode *Mel-Frequency Cepstral Coefficients (MFCC)* sebagai ekstraksi fitur dan algoritma Random Forest sebagai metode klasifikasi, sedangkan pada sistem visual digunakan *Convolutional Neural Network (CNN)* untuk mengenali ekspresi wajah. Hasil prediksi kedua modalitas kemudian digabungkan menggunakan metode *Weighted Late Fusion* dengan pembobotan 75% pada sistem visual dan 25% pada sistem audio. Penelitian difokuskan pada empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad*.

Hasil pelatihan menunjukkan bahwa model audio memperoleh akurasi sebesar 75,56%, sedangkan model visual memperoleh akurasi sebesar 87,48%. Selanjutnya, sistem diuji secara real-time terhadap 10 responden, dimana setiap responden melakukan 60 kali pengujian sehingga diperoleh 600 data pengujian pada setiap mode. Berdasarkan hasil pengujian, Mode Visual menunjukkan tingkat keberhasilan tertinggi pada emosi *Happy* sebesar 90,00% dan *Neutral* sebesar 86,00%. Pada Mode Audio, tingkat keberhasilan tertinggi diperoleh pada emosi *Sad* sebesar 88,67%. Sementara itu, Mode *Fusion* menunjukkan tingkat keberhasilan tertinggi pada emosi *Happy* sebesar 80,67% dan *Neutral* sebesar 68,00%. Hasil tersebut menunjukkan bahwa sistem visual memiliki performa yang lebih baik dibandingkan sistem audio, sedangkan metode *Weighted Late Fusion* mampu menghasilkan deteksi emosi yang lebih stabil melalui penggabungan informasi audio dan visual.

Berdasarkan hasil penelitian yang telah dilakukan, sistem deteksi emosi berbasis audio-visual berhasil diimplementasikan dalam bentuk aplikasi GUI berbasis Python yang mampu melakukan deteksi emosi secara real-time menggunakan webcam dan mikrofon. Penelitian ini diharapkan dapat menjadi salah satu referensi dalam pengembangan sistem *Human Computer Interaction (HCI).*

Kata kunci: *Audio -Visual*, *CNN*, deteksi emosi ,ekspresi wajah, MFCC, *Random Forest*,

Human Emotion Detection System Using Audio -Visual Methods Based on MFCC, Random Forest, and CNN

Abstract

The advancement of artificial intelligence technology has enabled computer systems to automatically recognize and understand human emotions. However, most existing emotion recognition systems rely on a single modality, such as speech or facial expressions, making their performance susceptible to environmental factors, including background noise and lighting conditions. This research aims to develop an ***Audio-Visual Emotion Recognition (AVER)*** *system capable of detecting human emotions in* ***real-time*** *by combining speech and facial expression information. In the audio subsystem,* ***Mel-Frequency Cepstral Coefficients (MFCC)*** *are employed for feature extraction, while the* ***Random Forest*** *algorithm is used for emotion classification. In the visual subsystem, a* ***Convolutional Neural Network (CNN)*** *is utilized to recognize facial expressions. The prediction results from both modalities are integrated using the* ***Weighted Late Fusion*** *method, with weighting factors of* ***75% for the visual modality*** *and* ***25% for the audio modality****. The proposed system focuses on four emotional categories:* ***Happy, Neutral, Angry, and Sad****.*

The training results indicate that the audio model achieved an accuracy of ***75.56%****, while the visual model achieved an accuracy of* ***87.48%****. The system was subsequently evaluated in* ***real-time*** *using* ***10 participants****, where each participant performed* ***60 emotion recognition trials****, resulting in* ***600 testing samples*** *for each operating mode. The experimental results show that the* ***Visual Mode*** *achieved the highest success rates for* ***Happy (90.00%)*** *and* ***Neutral (86.00%)*** *emotions. In the* ***Audio Mode****, the highest success rate was obtained for the* ***Sad*** *emotion at* ***88.67%****. Meanwhile, the* ***Fusion Mode*** *achieved the highest success rates for* ***Happy (80.67%)*** *and* ***Neutral (68.00%)****. These findings indicate that the visual subsystem outperformed the audio subsystem, while the* ***Weighted Late Fusion*** *approach provided a more stable emotion recognition performance by combining information from both audio and visual modalities.*

Overall, the proposed audio-visual emotion recognition system was successfully implemented as a ***Python-based graphical user interface (GUI)*** *capable of performing* ***real-time emotion detection*** *using a webcam and microphone. The proposed system is expected to contribute to the development of* ***Human-Computer Interaction (HCI)*** *applications and other intelligent systems requiring automatic human emotion recognition.*

Keywords: Audio -Visual, CNN, emotion recognition, facial expression, MFCC, Random Forest.

Kata Pengantar

Puji syukur kehadirat Allah SWT atas segala rahmat, karunia, dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Sistem Deteksi Emosi Manusia Menggunakan Metode Audio-Visual Berbasis MFCC, Random Forest, CNN" sebagai salah satu syarat untuk memperoleh gelar Sarjana Terapan pada Program Studi Teknik Elektronika, Politeknik Negeri Batam.

Penyusunan skripsi ini tidak terlepas dari bantuan, dukungan, bimbingan, serta doa dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Allah SWT atas segala rahmat, kesehatan, kemudahan, dan kesempatan sehingga penulis dapat menyelesaikan skripsi ini.
2. Kedua orang tua dan seluruh keluarga yang senantiasa memberikan doa, kasih sayang, dukungan, motivasi, serta pengorbanan yang tiada henti kepada penulis.
3. Ibu Feralia Fitri, S.T., M.T. selaku dosen pembimbing yang telah meluangkan waktu, tenaga, dan pikiran untuk memberikan arahan, bimbingan, masukan, serta motivasi selama proses penyusunan skripsi ini.
4. Bapak Djono, S.T., M.Sc. dan Bapak Sulistiono, S.T., M.T. selaku dosen penguji yang telah memberikan kritik, saran, dan masukan yang sangat membangun sehingga skripsi ini menjadi lebih baik.
5. Seluruh dosen dan staf Program Studi Teknik Elektronika Politeknik Negeri Batam yang telah memberikan ilmu pengetahuan, pengalaman, dan pelayanan akademik selama penulis menempuh pendidikan.
6. Seluruh responden yang telah bersedia meluangkan waktu untuk mengikuti proses pengujian sistem sehingga penelitian ini dapat terlaksana dengan baik.
7. Teman-teman seperjuangan serta semua pihak yang telah memberikan bantuan, dukungan, motivasi, dan doa selama proses penyusunan skripsi ini.

Penulis menyadari bahwa skripsi ini masih memiliki kekurangan dan jauh dari kata sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun demi penyempurnaan penelitian ini di masa yang akan datang.

Batam, 17 Juli 2026

Hormat Penulis,
M. Afif Fuadie Yusuf

Daftar Isi

Daftar Gambar

Daftar Tabel

Bab 1. Pendahuluan

1.1. Latar Belakang

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence/AI*) telah memberikan dampak besar pada berbagai bidang, termasuk interaksi manusia dan mesin. Salah satu penerapan AI yang semakin berkembang adalah *emotion recognition*, yaitu kemampuan sistem untuk mengenali dan menginterpretasikan emosi manusia melalui data yang ditangkap oleh sensor. Penggunaan teknologi ini menjadi penting karena emosi merupakan aspek yang sangat berpengaruh dalam komunikasi, perilaku, dan pengambilan keputusan manusia.

Pada umumnya, sistem deteksi emosi hanya memanfaatkan satu jenis data, baik dari suara atau ekspresi wajah. Deteksi emosi berbasis suara bergantung pada karakteristik akustik seperti intonasi, kecepatan bicara, dan perubahan nada. Sementara itu, deteksi emosi berbasis kamera mengandalkan ekspresi wajah seperti gerakan alis, bentuk mulut, dan perubahan pada mata. Namun, penggunaan satu jenis data sering kali memiliki keterbatasan. Misalnya, ekspresi wajah sulit dikenali pada kondisi pencahayaan kurang, sedangkan intonasi suara dapat terganggu oleh kebisingan lingkungan.

Untuk mengatasi keterbatasan tersebut, pendekatan *Audio -Visual Emotion Recognition (AVER)* mulai dikembangkan. Sistem ini menggabungkan dua sumber informasi sekaligus, yaitu suara dan citra wajah. Dengan mengombinasikan kedua data tersebut, akurasi sistem dalam mengenali emosi dapat meningkat secara signifikan karena informasi yang diperoleh lebih kaya dan saling melengkapi. Pendekatan multimodal ini juga lebih mendekati cara manusia memahami emosi lawan bicara, yakni melalui pengamatan *Visual* dan pendengaran secara bersamaan.

Implementasi sistem analisis emosi berbasis suara dan kamera berpotensi memberikan berbagai manfaat, seperti mendukung layanan kesehatan mental, membantu interaksi robot dan manusia (*human–robot interaction*), meningkatkan kualitas layanan pelanggan, hingga digunakan pada sistem pembelajaran adaptif. Sejalan dengan bidang Mekatronika, penerapan teknologi ini dapat diintegrasikan pada sistem cerdas, robot, atau perangkat monitoring berbasis sensor.

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk merancang dan menganalisis sistem deteksi emosi manusia menggunakan kombinasi suara dan kamera sehingga dapat menghasilkan model yang lebih akurat dan handal dibandingkan pendekatan tunggal.

1.2. Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana melakukan ekstraksi fitur suara menggunakan *MFCC* agar dapat merepresentasikan karakteristik akustik emosi manusia?
2. Bagaimana proses klasifikasi emosi berdasarkan suara menggunakan algoritma *Random Forest?*
3. Bagaimana melakukan ekstraksi fitur *Visual* dari ekspresi wajah menggunakan kamera dengan metode berbasis *CNN*.
4. Bagaimana menggabungkan hasil prediksi dari model *Audio* (*Random Forest)* dan model *Visual* (*CNN*) menggunakan pendekatan *late Fusion* untuk menghasilkan sistem deteksi emosi yang lebih akurat?
5. Bagaimana performa sistem dalam mengenali empat emosi utama—senang, sedih, marah, dan netral—baik secara *Audio* , *Visual*, maupun gabungan?

1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengembangkan sistem ekstraksi fitur *Audio* menggunakan MFCC untuk mendeteksi pola emosi pada suara.
2. Menerapkan algoritma *Random Forest* sebagai model klasifikasi emosi berbasis suara.
3. Mengembangkan sistem deteksi emosi berbasis kamera menggunakan model *CNN*.
4. Menggabungkan hasil prediksi *Audio* dan *Visual* melalui metode *late Fusion* untuk meningkatkan akurasi deteksi emosi.
5. Mengevaluasi dan membandingkan performa model *Audio* , model *Visual*, dan model gabungan pada empat kategori emosi.

1.4. Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Menjadi contoh implementasi sistem deteksi emosi berbasis *Audio-Visual* menggunakan metode MFCC, *Random Forest*, dan *CNN*.
2. Memberikan referensi dalam pengembangan teknologi deteksi emosi berbasis kecerdasan buatan secara realtime.
3. Menghasilkan aplikasi deteksi emosi berbasis desktop yang dapat digunakan melalui webcam dan microphone menggunakan antarmuka GUI.
4. Memberikan contoh penerapan penggabungan input *Audio* dan *Visual* pada sistem kecerdasan buatan.
5. Dapat digunakan sebagai media pendukung untuk membantu monitoring kondisi emosi seseorang melalui suara dan ekspresi wajah secara realtime.
6. Dapat digunakan sebagai media pembelajaran dan pengembangan penelitian selanjutnya di bidang Artificial Intelligence dan Computer Vision.

1.5. Batasan Masalah

Batasan dalam penelitian ini meliputi:

1. Penelitian ini hanya menggunakan empat kategori emosi: senang, sedih, marah, dan netral.
2. Data *Audio* diolah menggunakan fitur *MFCC*, lalu diklasifikasikan menggunakan algoritma *Random Forest*.
3. Data *Visual* yang digunakan berupa citra wajah dari kamera, yang diproses menggunakan model *Visual* berbasis *CNN***.**
4. Penggabungan informasi *Audio* dan *Visual* dilakukan menggunakan pendekatan *late Fusion*, yaitu penggabungan hasil prediksi akhir.
5. Sistem hanya mendeteksi satu wajah pada satu waktu (tidak multi-face recognition).
6. Pengujian dilakukan dalam kondisi pencahayaan dan kebisingan yang relatif terkontrol untuk menjaga konsistensi performa model.
7. Menggunakan perangkat berupa laptop data diambil dari kamera laptop dan *Audio* laptop.

Bab 2. Tinjauan Pustaka

2.1 Dasar Teori Emosi Manusia

Emosi merupakan respons psikologis dan fisiologis yang muncul sebagai reaksi terhadap stimulus dari lingkungan maupun kondisi internal individu. Emosi memengaruhi cara seseorang berpikir, mengambil keputusan, berkomunikasi, serta berinteraksi dengan lingkungan sekitarnya. Dalam bidang *Artificial Intelligence*, emosi menjadi salah satu aspek penting yang dipelajari untuk meningkatkan kemampuan sistem dalam memahami perilaku manusia melalui pendekatan *Human–Computer Interaction* (*HCI*) [1].

Secara umum, emosi dapat dikenali melalui beberapa modalitas, seperti ekspresi wajah, suara, gerakan tubuh, sinyal fisiologis, maupun kombinasi dari beberapa modalitas tersebut. Perkembangan teknologi *Emotion Recognition* menunjukkan bahwa penggunaan satu modalitas sering kali menghasilkan performa yang kurang optimal karena dipengaruhi oleh berbagai faktor, seperti pencahayaan pada citra wajah, kebisingan pada sinyal suara, maupun perbedaan karakteristik setiap individu. Oleh karena itu, penelitian terkini mulai mengembangkan pendekatan *multimodal* dengan menggabungkan informasi dari beberapa sumber untuk meningkatkan akurasi dan kestabilan proses pengenalan emosi [2].

Dalam penelitian ini, emosi yang digunakan terdiri dari empat kategori, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad*. Keempat kategori tersebut dipilih karena merupakan kategori emosi yang tersedia pada dataset *RAVDESS* untuk data suara dan *FER2013* untuk data citra wajah. Penggunaan kategori emosi yang sama pada kedua dataset bertujuan untuk mempermudah proses integrasi pada sistem *Audio-Visual Emotion Recognition* (*AVER*), sehingga hasil klasifikasi dari sistem

audio dan sistem visual dapat digabungkan menggunakan metode *Weighted Late Fusion* [3].

Perkembangan penelitian mengenai pengenalan emosi menunjukkan peningkatan yang sangat pesat dalam beberapa tahun terakhir. Berbagai metode berbasis *Machine Learning* dan *Deep Learning*, seperti *Random Forest*, *Support Vector Machine* (*SVM*), *Convolutional Neural Network* (*CNN*), *Long Short-Term Memory* (*LSTM*), serta pendekatan *multimodal* telah banyak digunakan untuk meningkatkan kemampuan sistem dalam mengenali emosi manusia. Berdasarkan berbagai penelitian, metode *CNN* menunjukkan performa yang sangat baik dalam mengenali ekspresi wajah, sedangkan *Random Forest* masih menjadi salah satu algoritma yang efektif untuk klasifikasi emosi berbasis fitur suara, khususnya ketika menggunakan fitur *Mel-Frequency Cepstral Coefficients* (*MFCC*) [4], [5].

Seiring berkembangnya teknologi *Human–Computer Interaction*, sistem deteksi emosi tidak lagi hanya difokuskan pada peningkatan akurasi model, tetapi juga pada kemampuan sistem untuk bekerja secara *real-time*, memiliki tingkat *robustness* yang tinggi terhadap perubahan lingkungan, serta mampu memberikan hasil deteksi yang stabil pada berbagai kondisi pengguna. Oleh karena itu, penelitian ini mengembangkan sistem deteksi emosi berbasis audio-visual yang menggabungkan metode *MFCC*, *Random Forest*, dan *CNN* agar mampu mengenali emosi manusia secara *real-time* melalui mikrofon dan kamera dengan performa yang lebih stabil dibandingkan penggunaan satu modalitas saja [5],[6].

2.2.1 *Pitch*

Pitch merupakan karakteristik suara yang menunjukkan tinggi atau rendahnya nada yang dihasilkan oleh getaran pita suara. *Pitch* berhubungan langsung dengan frekuensi fundamental (*Fundamental Frequency*/*F0*), yaitu jumlah getaran pita suara setiap detik yang dinyatakan dalam satuan *Hertz* (Hz). Semakin tinggi frekuensi fundamental, maka semakin tinggi pula *pitch* yang dihasilkan.

Sebaliknya, frekuensi fundamental yang rendah akan menghasilkan suara dengan *pitch* yang lebih rendah [7], [8].

Dalam proses komunikasi, perubahan *pitch* sering kali mencerminkan kondisi emosional seseorang. Individu yang mengalami emosi *happy* atau *angry* cenderung memiliki frekuensi fundamental yang lebih tinggi sehingga suara terdengar lebih nyaring dan ekspresif. Sebaliknya, pada emosi *sad*, frekuensi fundamental umumnya menurun sehingga suara terdengar lebih rendah, pelan, dan lambat. Sementara itu, emosi *neutral* memiliki variasi *pitch* yang relatif stabil dibandingkan kategori emosi lainnya [9], [10].

Dalam penelitian *Speech Emotion Recognition* (*SER*), *pitch* merupakan salah satu parameter prosodik yang paling sering digunakan karena mampu memberikan informasi mengenai perubahan kondisi emosional seseorang. Namun demikian, penggunaan *pitch* sebagai satu-satunya fitur belum mampu menghasilkan klasifikasi emosi yang optimal karena nilainya dipengaruhi oleh berbagai faktor, seperti jenis kelamin, usia, karakteristik biologis, serta kebiasaan berbicara setiap individu. Oleh karena itu, penelitian terkini lebih banyak mengombinasikan *pitch* dengan karakteristik akustik lainnya, seperti intensitas, durasi, *speech rate*, serta fitur spektral yang diekstraksi menggunakan metode *Mel-Frequency Cepstral Coefficients* (*MFCC*) sehingga diperoleh representasi suara yang lebih komprehensif [8], .

Secara fisiologis, hubungan antara frekuensi fundamental dengan karakteristik pita suara dapat dinyatakan menggunakan persamaan berikut.

$$f_0 = \frac{1}{2L}\sqrt{\frac{T}{\mu}}$$

Keterangan :

f_0 = frekuensi dasar L = Panjang pita suara

T = Tegangan pita suara

μ = Massa per satuan panjang pita suara

Persamaan tersebut menunjukkan bahwa frekuensi fundamental dipengaruhi oleh panjang, massa, dan tegangan pita suara. Semakin besar tegangan pita suara, maka frekuensi fundamental akan meningkat sehingga menghasilkan *pitch* yang lebih tinggi. Sebaliknya, pita suara yang lebih panjang atau memiliki massa per satuan panjang yang lebih besar akan menghasilkan frekuensi fundamental yang lebih rendah [11].

Pada penelitian ini, nilai *pitch* tidak digunakan secara langsung sebagai fitur masukan ke dalam model klasifikasi. Informasi mengenai perubahan *pitch* direpresentasikan melalui proses ekstraksi fitur menggunakan metode *Mel-Frequency Cepstral Coefficients* (*MFCC*). Metode *MFCC* mampu menangkap karakteristik spektral suara yang dipengaruhi oleh perubahan frekuensi fundamental sehingga informasi mengenai *pitch* tetap menjadi bagian dari fitur yang dipelajari oleh algoritma *Random Forest* dalam proses klasifikasi empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad* [8], [12].

2.2.2 Intensitas Suara

Intensitas suara merupakan karakteristik akustik yang menunjukkan besar kecilnya energi atau kekuatan sinyal suara yang dihasilkan ketika seseorang berbicara. Intensitas berkaitan dengan tekanan udara yang dihasilkan selama proses fonasi dan umumnya dinyatakan dalam satuan *decibel* (dB). Semakin besar energi suara yang dihasilkan, maka semakin tinggi pula nilai intensitas suara, sedangkan energi yang lebih rendah akan menghasilkan intensitas suara yang lebih kecil [7], [8].

Perubahan intensitas suara dipengaruhi oleh kondisi emosional seseorang. Individu yang sedang mengalami emosi *angry* umumnya berbicara dengan tekanan suara yang lebih kuat sehingga menghasilkan intensitas yang tinggi. Sebaliknya, emosi *sad* cenderung menghasilkan intensitas suara yang lebih rendah karena tekanan udara yang dikeluarkan lebih kecil. Pada emosi *happy*, intensitas suara berada pada tingkat sedang hingga tinggi dengan variasi yang lebih dinamis,

sedangkan emosi *neutral* umumnya memiliki intensitas yang relatif stabil selama proses berbicara [9], [10].

Dalam penelitian *Speech Emotion Recognition* (*SER*), intensitas suara merupakan salah satu parameter prosodik yang sering digunakan bersama karakteristik suara lainnya, seperti *pitch*, durasi, dan *speech rate*. Meskipun intensitas mampu memberikan informasi mengenai perubahan energi suara akibat emosi, penggunaannya sebagai fitur tunggal belum mampu menghasilkan klasifikasi yang optimal karena dipengaruhi oleh berbagai faktor, seperti jarak mikrofon, sensitivitas perangkat perekam, tingkat kebisingan lingkungan, serta volume suara masing-masing individu. Oleh karena itu, penelitian terkini lebih banyak mengombinasikan intensitas dengan fitur-fitur akustik lainnya agar diperoleh representasi suara yang lebih akurat [8], [12].

Metode *Mel-Frequency Cepstral Coefficients* (*MFCC*) mampu merepresentasikan karakteristik spektral suara secara lebih komprehensif. Informasi mengenai perubahan intensitas suara secara tidak langsung direpresentasikan melalui koefisien *MFCC* bersama karakteristik akustik lainnya, sehingga model klasifikasi mampu mengenali pola emosi secara lebih efektif dibandingkan apabila hanya menggunakan parameter intensitas secara terpisah [8], [12]. Secara umum, intensitas suara dapat dihitung menggunakan persamaan logaritmik berikut.

$$I_{dB} = 10\log\left(\frac{I}{I_0}\right)$$

Ketereangan :

I = intensitas suara (W/m^2)

I_o = intensitas referensi (1×10^{-12} W/m^2),

Persamaan tersebut menunjukkan bahwa tingkat intensitas suara dinyatakan sebagai perbandingan antara intensitas suara yang diukur terhadap intensitas referensi menggunakan skala logaritmik. Semakin besar nilai intensitas suara, maka semakin besar

pula energi akustik yang diterima oleh mikrofon. Sebaliknya, intensitas yang rendah menunjukkan energi suara yang lebih kecil sehingga suara terdengar lebih lemah [11].

Pada penelitian ini, intensitas suara tidak digunakan sebagai parameter masukan secara langsung. Karakteristik tersebut telah direpresentasikan melalui proses ekstraksi fitur menggunakan metode *MFCC*, kemudian hasil ekstraksi diklasifikasikan menggunakan algoritma *Random Forest* untuk mengenali empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad*. Pendekatan ini dipilih karena *MFCC* mampu mengintegrasikan berbagai karakteristik akustik, termasuk perubahan energi suara, sehingga menghasilkan representasi fitur yang lebih efektif dalam proses klasifikasi emosi secara *real-time* [8], [12].

2.2.3 Durasi

Durasi merupakan karakteristik akustik yang menunjukkan lamanya waktu yang diperlukan untuk mengucapkan suatu kata, kalimat, atau keseluruhan sinyal suara. Durasi umumnya diukur dalam satuan detik (*second*) atau milidetik (*millisecond*) dan dipengaruhi oleh kecepatan berbicara, pola artikulasi, serta kondisi emosional seseorang. Perubahan durasi dalam proses berbicara dapat memberikan informasi penting mengenai keadaan psikologis dan emosi yang sedang dialami oleh individu [7], [8].

Pada proses komunikasi, setiap emosi memiliki pola durasi yang berbeda. Emosi *sad* umumnya ditandai dengan tempo bicara yang lebih lambat sehingga menghasilkan durasi pengucapan yang lebih panjang. Sebaliknya, emosi *happy* dan *angry* cenderung memiliki tempo bicara yang lebih cepat dengan durasi pengucapan yang lebih singkat. Sementara itu, emosi *neutral* biasanya menunjukkan durasi yang relatif stabil tanpa adanya perubahan tempo yang signifikan [9], [10].

Dalam penelitian *Speech Emotion Recognition* (*SER*), durasi merupakan salah satu parameter akustik yang digunakan untuk membantu membedakan

karakteristik antar emosi. Namun, penggunaan durasi sebagai fitur tunggal belum mampu menghasilkan performa klasifikasi yang optimal karena dipengaruhi oleh kebiasaan berbicara setiap individu, panjang kalimat yang diucapkan, bahasa yang digunakan, serta kondisi proses perekaman suara. Oleh karena itu, penelitian terkini lebih banyak mengombinasikan durasi dengan karakteristik akustik lainnya, seperti *pitch*, intensitas suara, *speech rate*, serta fitur spektral yang diekstraksi menggunakan metode *Mel-Frequency Cepstral Coefficients* (*MFCC*) sehingga diperoleh representasi suara yang lebih komprehensif [8], [12].

Pada penelitian ini, karakteristik durasi direpresentasikan secara tidak langsung melalui proses ekstraksi fitur menggunakan metode *MFCC*. Sinyal suara hasil perekaman dibagi ke dalam beberapa *frame* sebelum dilakukan ekstraksi fitur, sehingga perubahan tempo berbicara akibat perbedaan emosi dapat dipelajari oleh algoritma *Random Forest*. Dengan demikian, informasi mengenai durasi tetap berkontribusi dalam proses klasifikasi empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad*, meskipun tidak digunakan sebagai parameter pengukuran secara langsung [8], [12].

2.2.4 Kecepatan Suara

Kecepatan suara atau *speech rate* merupakan karakteristik akustik yang menunjukkan seberapa cepat seseorang mengucapkan kata atau kalimat dalam proses berbicara. *Speech rate* umumnya dinyatakan sebagai jumlah kata per menit (*Words Per Minute*/*WPM*) atau jumlah suku kata per detik (*Syllables Per Second*/*SPS*). Kecepatan berbicara dipengaruhi oleh berbagai faktor, seperti kebiasaan individu, kondisi psikologis, lingkungan, serta emosi yang sedang dirasakan [7], [8].

Perubahan *speech rate* merupakan salah satu indikator penting dalam analisis emosi berbasis suara. Individu yang sedang mengalami emosi *happy* atau *angry* cenderung berbicara dengan tempo yang lebih cepat sehingga jumlah kata yang diucapkan dalam satuan waktu menjadi lebih banyak. Sebaliknya, emosi *sad* biasanya ditandai dengan tempo bicara yang lebih lambat disertai jeda yang lebih panjang pada setiap kalimat. Sementara itu, emosi *neutral* memiliki kecepatan

berbicara yang relatif stabil sehingga perubahan tempo tidak terlalu signifikan [9], [10].

Dalam penelitian *Speech Emotion Recognition* (*SER*), *speech rate* merupakan salah satu karakteristik prosodik yang dapat membantu membedakan pola emosi antar individu. Akan tetapi, parameter ini tidak digunakan secara terpisah karena dipengaruhi oleh berbagai faktor, seperti panjang kalimat, gaya berbicara, bahasa yang digunakan, serta karakteristik setiap pembicara. Oleh karena itu, penelitian terkini lebih banyak memanfaatkan metode ekstraksi fitur, seperti *Mel-Frequency Cepstral Coefficients* (*MFCC*), yang mampu merepresentasikan berbagai karakteristik akustik secara bersamaan, termasuk perubahan tempo berbicara [8], [12].

Metode *MFCC* menganalisis sinyal suara dalam interval waktu yang sangat singkat (*frame*) sehingga perubahan kecepatan berbicara dapat direpresentasikan melalui perubahan pola spektral pada setiap *frame*. Pendekatan ini memungkinkan algoritma klasifikasi mempelajari hubungan antara tempo bicara dan karakteristik emosi tanpa harus menghitung nilai *speech rate* secara langsung. Dengan demikian, informasi temporal suara tetap dipertimbangkan dalam proses klasifikasi emosi [10], [12].

Pada penelitian ini, kecepatan berbicara tidak digunakan sebagai parameter masukan secara langsung, tetapi telah direpresentasikan melalui proses ekstraksi fitur menggunakan metode *MFCC*. Selanjutnya, fitur-fitur hasil ekstraksi diklasifikasikan menggunakan algoritma *Random Forest* untuk mengenali empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad*. Pendekatan ini menghasilkan representasi karakteristik suara yang lebih komprehensif dibandingkan penggunaan *speech rate* sebagai fitur tunggal, sehingga mampu meningkatkan performa sistem dalam proses deteksi emosi secara *real-time* [8], [12].

2.2.5 Amplitudo

Amplitudo merupakan karakteristik akustik yang menunjukkan besar kecilnya simpangan gelombang suara dari posisi keseimbangan sehingga menggambarkan kekuatan atau energi sinyal suara yang dihasilkan selama proses berbicara. Dalam produksi suara manusia, amplitudo dipengaruhi oleh tekanan udara yang berasal dari paru-paru serta getaran pita suara selama proses fonasi berlangsung. Semakin besar amplitudo, maka semakin tinggi energi suara yang dihasilkan, sedangkan amplitudo yang lebih kecil menunjukkan energi suara yang lebih rendah [7], [8].

Perubahan amplitudo berkaitan erat dengan kondisi emosional seseorang. Emosi *angry* umumnya menghasilkan amplitudo yang lebih tinggi karena individu berbicara dengan tekanan suara yang kuat dan energi vokal yang besar. Sebaliknya, emosi *sad* cenderung memiliki amplitudo yang lebih rendah akibat berkurangnya tekanan suara selama berbicara. Emosi *happy* biasanya menghasilkan amplitudo pada tingkat sedang hingga tinggi dengan variasi yang lebih dinamis, sedangkan emosi *neutral* menunjukkan amplitudo yang relatif stabil sepanjang proses komunikasi [9], [10].

Dalam bidang *Speech Emotion Recognition* (*SER*), amplitudo merupakan salah satu parameter prosodik yang digunakan untuk menggambarkan tingkat energi sinyal suara. Akan tetapi, amplitudo sangat dipengaruhi oleh berbagai faktor eksternal, seperti jarak pembicara terhadap mikrofon, sensitivitas perangkat perekam, tingkat kebisingan lingkungan, serta kualitas perangkat akuisisi suara. Oleh karena itu, amplitudo tidak digunakan sebagai fitur tunggal dalam proses klasifikasi emosi karena kurang mampu memberikan representasi yang konsisten pada berbagai kondisi perekaman [8], [12].

Penelitian terkini lebih banyak memanfaatkan metode *Mel-Frequency Cepstral Coefficients* (*MFCC*) untuk merepresentasikan karakteristik akustik suara secara menyeluruh. Pada proses ekstraksi fitur, informasi mengenai distribusi energi sinyal yang dipengaruhi oleh perubahan amplitudo turut direpresentasikan dalam koefisien *MFCC*. Dengan demikian, model klasifikasi dapat mempelajari pola perubahan energi suara secara lebih komprehensif tanpa harus menggunakan nilai amplitudo sebagai parameter masukan secara langsung [8][8], [12].

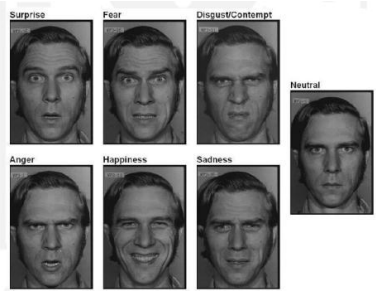
Pada penelitian ini, amplitudo tidak digunakan sebagai parameter masukan secara langsung, melainkan telah direpresentasikan melalui proses ekstraksi fitur menggunakan metode *MFCC*. Selanjutnya, fitur-fitur hasil ekstraksi diklasifikasikan menggunakan algoritma *Random Forest* untuk mengenali empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad*. Pendekatan ini menghasilkan representasi karakteristik suara yang lebih stabil dan lebih efektif dibandingkan penggunaan amplitudo sebagai fitur tunggal, sehingga mampu meningkatkan performa sistem dalam mendetksi emosi secara *real-time* [8], [12].

2.3 Deteksi Emosi Manusia Melalui Wajah

Deteksi emosi melalui wajah merupakan salah satu pendekatan utama dalam mengenali kondisi emosional seseorang berdasarkan perubahan ekspresi wajah. Wajah manusia mengandung berbagai informasi visual yang dapat mencerminkan keadaan emosional, seperti perubahan posisi alis, bentuk mata, gerakan pipi, serta bentuk mulut. Perubahan pada bagian-bagian wajah tersebut terjadi secara alami sebagai respons terhadap rangsangan emosional sehingga dapat digunakan sebagai indikator dalam proses pengenalan emosi [13], [14].

Perkembangan penelitian menunjukkan bahwa ekspresi wajah merupakan salah satu modalitas yang paling efektif dalam sistem *Emotion Recognition* karena mampu memberikan informasi yang bersifat nonverbal dan relatif mudah diamati. Berbagai penelitian mengelompokkan ekspresi wajah ke dalam beberapa kategori emosi dasar, seperti *happy*, *sad*, *angry*, *fear*, *surprise*, *disgust*, dan *neutral*. Kategori tersebut banyak digunakan sebagai dasar dalam pengembangan sistem *Facial Emotion Recognition* (*FER*) karena memiliki karakteristik ekspresi yang berbeda dan dapat dipelajari oleh model *Deep Learning* [13], [15].

Secara fisiologis, ekspresi wajah dikendalikan oleh sistem saraf yang mengatur kontraksi berbagai otot wajah. Ketika seseorang mengalami perubahan emosi, otak mengirimkan sinyal ke otot-otot wajah sehingga menghasilkan ekspresi tertentu. Perubahan posisi alis, mata, hidung, dan bibir menjadi ciri khas yang membedakan setiap kategori emosi. Informasi visual tersebut kemudian dimanfaatkan sebagai sumber utama dalam proses deteksi emosi berbasis citra wajah [14].



Gambar 2. 1 Contoh Citra Emosi

Dalam konteks teknologi, pengenalan emosi melalui wajah telah berkembang pesat seiring kemajuan *computer vision* dan *deep learning*. *Convolutional Neural Network* (*CNN*) merupakan salah satu metode yang paling banyak digunakan untuk mendeteksi ekspresi wajah karena kemampuannya mengekstraksi fitur-fitur visual secara otomatis. *CNN* mempelajari pola geometris wajah, seperti jarak antar titik wajah, perubahan intensitas piksel, tekstur wajah, serta perubahan morfologi yang berkaitan dengan kategori emosi tertentu. Berbagai penelitian menunjukkan bahwa metode *CNN* mampu menghasilkan tingkat akurasi yang lebih tinggi dibandingkan metode konvensional pada tugas *Facial Emotion Recognition* (*FER*), terutama ketika menggunakan dataset *FER2013* [13], [16].

Selain ekspresi wajah yang terlihat secara jelas (*macro-expression*), wajah juga menampilkan *micro-expression*, yaitu perubahan ekspresi yang berlangsung sangat singkat dan sering kali tidak disadari oleh individu. *Micro-expression* merupakan indikator yang muncul ketika seseorang berusaha menyembunyikan atau menahan emosinya. Dengan perkembangan teknologi *computer vision* dan *deep learning*, pola ekspresi yang sangat halus tersebut dapat dikenali secara lebih akurat menggunakan model *CNN*, sehingga meningkatkan kemampuan sistem dalam memahami kondisi emosional seseorang [14], [16].

Pada sistem pendeteksian emosi berbasis wajah, proses umum yang dilakukan meliputi:

1. *Deteksi wajah*, menggunakan algoritma seperti *Haar Cascade* atau *Multi-task Cascaded Convolutional Neural Network* (*MTCNN*) untuk menemukan lokasi wajah pada citra.
2. *Ekstraksi fitur*, dilakukan secara otomatis oleh model *CNN* yang mempelajari karakteristik visual dari dataset pelatihan.

Klasifikasi emosi, yaitu proses penentuan kategori emosi berdasarkan probabilitas keluaran model, kemudian dipilih kelas dengan nilai probabilitas tertinggi sebagai hasil prediksi [13], [16].

Dalam pendekatan *multimodal* seperti yang digunakan pada penelitian ini, deteksi emosi melalui wajah dipadukan dengan deteksi emosi melalui suara. Ekspresi wajah dan karakteristik vokal saling melengkapi dalam merepresentasikan kondisi emosional seseorang. Oleh karena itu, penggabungan kedua modalitas mampu mengurangi keterbatasan yang dimiliki masing-masing sistem, seperti gangguan akibat pencahayaan yang kurang baik pada sistem visual maupun kebisingan lingkungan pada sistem audio [17], [18].

Pemanfaatan deteksi emosi melalui wajah telah diterapkan pada berbagai bidang, seperti *human-computer interaction*, layanan pelanggan, sistem keamanan, pendidikan, serta layanan kesehatan. Kemampuan sistem dalam mengenali emosi secara otomatis memungkinkan terciptanya interaksi antara manusia dan komputer yang lebih adaptif, alami, dan responsif terhadap kondisi emosional pengguna [14], [18].

2.3.1 Citra

Citra merupakan representasi visual dari suatu objek yang diperoleh melalui perangkat akuisisi, seperti kamera digital, kemudian disimpan dalam bentuk data digital yang tersusun atas sejumlah piksel (*pixel*). Setiap piksel memiliki nilai intensitas tertentu yang merepresentasikan tingkat kecerahan maupun warna pada suatu lokasi gambar. Dalam bidang *computer vision*, citra digital menjadi sumber informasi utama yang digunakan untuk melakukan analisis, pengenalan objek, segmentasi, serta klasifikasi berbagai pola visual [13], [17].

Proses pembentukan citra digital dilakukan melalui digitalisasi gambar analog menggunakan proses *sampling* dan *quantization*. Proses *sampling* mengubah sinyal visual kontinu menjadi kumpulan piksel, sedangkan *quantization* menentukan tingkat intensitas atau warna pada setiap piksel. Hasil dari kedua proses tersebut adalah citra digital yang dapat diproses menggunakan berbagai metode pengolahan citra maupun algoritma *deep learning* [14], [18].

Dalam pengolahan citra digital dikenal dua teknik perubahan resolusi, yaitu *downsampling* dan *upsampling*. *Downsampling* merupakan proses mengurangi jumlah piksel atau resolusi citra sehingga ukuran gambar menjadi lebih kecil, sedangkan *upsampling* merupakan proses meningkatkan jumlah piksel untuk menghasilkan resolusi citra yang lebih besar. Kedua teknik tersebut sering digunakan pada tahap *pre-processing* maupun dalam arsitektur *Convolutional Neural Network* (*CNN*) untuk menyesuaikan ukuran citra sebelum dilakukan proses klasifikasi [15], [16].

Pada penelitian ini, citra wajah diperoleh menggunakan kamera *webcam* kemudian diproses menjadi citra digital yang digunakan sebagai masukan pada model *CNN*. Seluruh citra wajah disesuaikan ukurannya dengan dataset *FER2013* sehingga proses pelatihan dan pengujian dapat dilakukan secara konsisten.

Selanjutnya, model *CNN* mempelajari karakteristik visual dari setiap citra untuk mengklasifikasikan empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad* [16], [18].

2.4 Perkembangan *Classifier Learning*

Classifier learning merupakan salah satu cabang penting dalam *machine learning* yang berfokus pada pengembangan model untuk mengelompokkan data ke dalam kelas-kelas tertentu berdasarkan pola yang dipelajari dari data pelatihan (*training data*). Tujuan utama proses klasifikasi adalah membangun model yang mampu memberikan prediksi secara akurat terhadap data baru yang belum pernah dipelajari sebelumnya. Berbagai algoritma klasifikasi telah dikembangkan untuk menangani berbagai jenis data, seperti citra, suara, teks, maupun sinyal fisiologis [19].

Perkembangan *classifier learning* diawali dengan penggunaan algoritma klasifikasi konvensional, seperti *K-Nearest Neighbor* (*KNN*), *Naïve Bayes*, *Decision Tree*, dan *Support Vector Machine* (*SVM*). Algoritma tersebut banyak digunakan karena memiliki proses pelatihan yang relatif sederhana dan mampu memberikan performa yang baik pada dataset dengan jumlah fitur yang terbatas. Namun demikian, performanya cenderung menurun ketika dihadapkan pada data berdimensi tinggi atau data dengan karakteristik yang kompleks [19], [20].

Seiring berkembangnya teknologi komputasi, metode *deep learning* mulai mendominasi berbagai bidang klasifikasi. Model seperti *Convolutional Neural Network* (*CNN*) terbukti mampu melakukan ekstraksi fitur secara otomatis pada data citra, sedangkan *Recurrent Neural Network* (*RNN*) dan *Long Short-Term Memory* (*LSTM*) banyak digunakan untuk menganalisis data sekuensial, termasuk sinyal suara. Kemampuan tersebut menjadikan metode *deep learning* sebagai salah satu pendekatan utama dalam sistem *Emotion Recognition*, baik berbasis suara maupun citra wajah [20].

Selain metode *deep learning*, algoritma *ensemble learning* juga mengalami perkembangan yang cukup pesat. Salah satu algoritma yang banyak digunakan adalah *Random Forest*, yaitu metode klasifikasi yang menggabungkan sejumlah *decision tree* untuk menghasilkan keputusan berdasarkan mekanisme *majority voting*. Pendekatan ini mampu meningkatkan akurasi klasifikasi, mengurangi risiko *overfitting*, serta memberikan performa yang stabil pada data berdimensi tinggi maupun data hasil ekstraksi fitur, seperti *Mel-Frequency Cepstral Coefficients* (*MFCC*) [21].

2.4 Perkembangan *Classifier Learning*

Classifier learning merupakan salah satu cabang dalam *machine learning* yang berfokus pada pengembangan model untuk mengelompokkan data ke dalam

kelas-kelas tertentu berdasarkan pola yang dipelajari dari data pelatihan (*training data*). Tujuan utama proses klasifikasi adalah membangun model yang mampu memberikan prediksi secara akurat terhadap data baru yang belum pernah dipelajari sebelumnya. Metode klasifikasi telah banyak diterapkan pada berbagai bidang, seperti pengenalan wajah, pengenalan suara, deteksi objek, analisis teks, hingga sistem *Emotion Recognition* [19].

Perkembangan *classifier learning* diawali dengan penggunaan algoritma klasifikasi konvensional, seperti *K-Nearest Neighbor* (*KNN*), *Naïve Bayes*, *Decision Tree*, dan *Support Vector Machine* (*SVM*). Algoritma-algoritma tersebut memiliki proses pelatihan yang relatif sederhana dan mampu memberikan hasil klasifikasi yang baik pada dataset dengan jumlah fitur yang terbatas. Namun, performanya cenderung menurun ketika dihadapkan pada data berdimensi tinggi atau data yang memiliki hubungan nonlinier sehingga diperlukan metode klasifikasi yang lebih adaptif [19], [20].

Seiring meningkatnya kemampuan komputasi, perkembangan *deep learning* membawa perubahan signifikan dalam bidang klasifikasi. Model seperti *Convolutional Neural Network* (*CNN*) mampu melakukan ekstraksi fitur secara otomatis pada data citra, sedangkan *Recurrent Neural Network* (*RNN*) dan *Long Short-Term Memory* (*LSTM*) banyak digunakan untuk memproses data sekuensial, termasuk sinyal suara. Kemampuan tersebut menjadikan metode *deep learning* sebagai salah satu pendekatan utama dalam berbagai penelitian *Emotion Recognition* karena mampu menghasilkan tingkat akurasi yang lebih tinggi dibandingkan metode konvensional [20].

Selain metode *deep learning*, algoritma berbasis *ensemble learning* juga berkembang pesat sebagai alternatif dalam proses klasifikasi. Salah satu algoritma yang banyak digunakan adalah *Random Forest*, yaitu metode yang menggabungkan sejumlah *decision tree* untuk menghasilkan keputusan berdasarkan mekanisme *majority voting*. Pendekatan ini memiliki keunggulan dalam mengurangi *overfitting*, meningkatkan kemampuan generalisasi model, serta memberikan performa yang stabil pada data hasil ekstraksi fitur, seperti *Mel-Frequency Cepstral Coefficients* (*MFCC*) [22]

Dalam penelitian ini digunakan dua pendekatan klasifikasi yang berbeda sesuai dengan karakteristik data yang diolah. Data suara diklasifikasikan menggunakan algoritma *Random Forest* berdasarkan fitur hasil ekstraksi *MFCC*, sedangkan data citra wajah diproses menggunakan model *Convolutional Neural Network* (*CNN*). Hasil klasifikasi dari kedua modalitas tersebut kemudian digabungkan menggunakan metode *Weighted Late Fusion* untuk menghasilkan sistem deteksi emosi berbasis *audio-visual* yang lebih akurat dan stabil dibandingkan penggunaan satu modalitas saja.

Perkembangan berbagai algoritma klasifikasi tersebut menunjukkan bahwa tidak ada satu metode yang mampu memberikan performa terbaik untuk seluruh

jenis data. Oleh karena itu, pemilihan algoritma harus disesuaikan dengan karakteristik data dan tujuan penelitian. Berdasarkan pertimbangan tersebut, penelitian ini memanfaatkan kombinasi *Random Forest* untuk klasifikasi suara dan *CNN* untuk klasifikasi citra wajah karena kedua metode tersebut telah terbukti memberikan performa yang baik pada penelitian-penelitian terdahulu serta sesuai dengan karakteristik dataset *RAVDESS* dan *FER2013* yang digunakan.

2.5 *Random Forest Classifier*

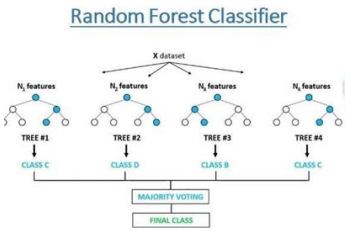
Random Forest merupakan salah satu algoritma *ensemble learning* yang digunakan untuk menyelesaikan permasalahan klasifikasi maupun regresi dengan menggabungkan sejumlah *decision tree* menjadi satu model yang lebih kuat. Algoritma ini diperkenalkan oleh Leo Breiman pada tahun 2001 dan hingga saat ini masih banyak digunakan karena memiliki tingkat akurasi yang tinggi, mampu menangani data berdimensi besar, serta memiliki ketahanan yang baik terhadap *overfitting*. Berbeda dengan *decision tree* tunggal yang rentan terhadap perubahan data pelatihan, *Random Forest* menghasilkan model yang lebih stabil melalui proses penggabungan hasil prediksi dari banyak pohon keputusan [23].

Prinsip utama *Random Forest* adalah membangun sejumlah *decision tree* menggunakan subset data dan subset fitur yang dipilih secara acak. Setiap pohon melakukan proses klasifikasi secara independen, kemudian seluruh hasil prediksi digabungkan menggunakan mekanisme *majority voting* untuk menentukan kelas akhir. Pendekatan ini memungkinkan model menghasilkan keputusan yang lebih akurat karena kesalahan pada satu pohon dapat dikompensasi oleh pohon lainnya [24].

Keunggulan lain dari *Random Forest* adalah kemampuannya menangani data dengan jumlah fitur yang banyak tanpa memerlukan proses seleksi fitur secara manual. Selain itu, algoritma ini mampu memberikan informasi mengenai tingkat kepentingan setiap fitur (*feature importance*) sehingga peneliti dapat mengetahui fitur mana yang paling berpengaruh terhadap proses klasifikasi. Kemampuan tersebut menjadikan *Random Forest* sebagai salah satu algoritma yang banyak digunakan pada berbagai bidang, seperti pengolahan citra, pengenalan suara, diagnosis medis, serta sistem *Emotion Recognition* [24].

Pada penelitian *Speech Emotion Recognition* (*SER*), *Random Forest* banyak digunakan karena mampu mengolah data hasil ekstraksi fitur *Mel-Frequency Cepstral Coefficients* (*MFCC*) secara efektif. Fitur *MFCC* memiliki dimensi yang cukup besar dan mengandung informasi spektral yang kompleks. Melalui mekanisme *ensemble learning*, *Random Forest* mampu mempelajari hubungan antarfitur tanpa memerlukan asumsi distribusi data tertentu sehingga menghasilkan performa klasifikasi yang baik pada berbagai dataset emosi suara, termasuk *RAVDESS* [25][26].

Pada penelitian ini, *Random Forest* digunakan sebagai model klasifikasi untuk data audio yang telah diekstraksi menggunakan metode *MFCC*. Pemilihan algoritma ini didasarkan pada kemampuannya dalam mengurangi risiko *overfitting*, menangani data dengan jumlah fitur yang tinggi, serta menghasilkan proses klasifikasi yang stabil terhadap variasi sinyal suara. Hasil klasifikasi dari model *Random Forest* kemudian digabungkan dengan hasil klasifikasi model *Convolutional Neural Network* (*CNN*) menggunakan metode *Weighted Late Fusion* untuk memperoleh keputusan akhir pada sistem deteksi emosi berbasis *audio-visual*.



Gambar 2. 2 Cara kerja *Random Forest Classifier*

2.5.1 *Bootstrap Sampling*

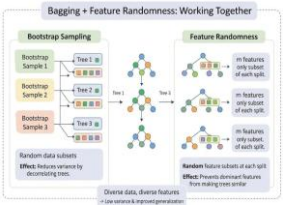
Bootstrap sampling merupakan tahap awal dalam pembentukan model *Random Forest*. Teknik ini dilakukan dengan mengambil sampel data pelatihan (*training data*) secara acak menggunakan metode *sampling with replacement*, yaitu setiap data yang telah terpilih masih memiliki peluang untuk dipilih kembali pada proses pengambilan sampel berikutnya. Akibatnya, satu data dapat muncul lebih dari satu kali dalam satu subset, sementara sebagian data lainnya tidak terpilih sama sekali. Setiap subset hasil *bootstrap sampling* selanjutnya digunakan untuk membangun satu *decision tree* yang berbeda [25], [26].

Penerapan *bootstrap sampling* bertujuan untuk meningkatkan keragaman (*diversity*) antar *decision tree* dalam *Random Forest*. Karena setiap pohon dilatih menggunakan kombinasi data yang berbeda, model yang dihasilkan memiliki karakteristik pembelajaran yang beragam. Keragaman tersebut menjadi salah satu faktor utama yang membuat *Random Forest* mampu mengurangi risiko *overfitting* dibandingkan *decision tree* tunggal, sekaligus meningkatkan kemampuan model dalam melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya [25].

Selain meningkatkan keragaman model, proses *bootstrap sampling* juga menghasilkan sejumlah data yang tidak terpilih pada setiap subset pelatihan. Data tersebut dikenal sebagai *Out-of-Bag* (*OOB*) dan dimanfaatkan sebagai data validasi internal untuk mengestimasi performa model tanpa memerlukan pembagian

dataset secara terpisah. Pendekatan ini membuat proses evaluasi menjadi lebih efisien karena pelatihan dan validasi dapat dilakukan secara bersamaan selama pembentukan *Random Forest* [24].

Pada penelitian ini, mekanisme *bootstrap sampling* digunakan secara otomatis oleh algoritma *Random Forest* pada saat proses pelatihan data audio yang telah diekstraksi menggunakan metode *Mel-Frequency Cepstral Coefficients* (*MFCC*). Setiap *decision tree* dibangun menggunakan subset data hasil *bootstrap sampling*, sehingga model mampu mempelajari variasi karakteristik sinyal suara secara lebih baik. Pendekatan ini membantu meningkatkan stabilitas model dalam mengklasifikasikan empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad* [26].



Gambar 2. 3 Proses Bootstrap Sampling pada Random Forest.

2.5.2 *Decision Tree*

Decision Tree merupakan algoritma klasifikasi yang bekerja dengan membentuk struktur pohon (*tree*) untuk memisahkan data berdasarkan karakteristik atau atribut tertentu. Struktur pohon terdiri atas *root node*, *internal node*, cabang (*branch*), dan *leaf node*. *Root node* merupakan simpul awal yang berisi seluruh data pelatihan, sedangkan *internal node* berfungsi sebagai titik pengambilan keputusan berdasarkan nilai suatu fitur. Proses klasifikasi berakhir pada *leaf node* yang menunjukkan hasil prediksi dari kelas tertentu [25].

Pada setiap proses percabangan, *Decision Tree* memilih atribut yang mampu memberikan pemisahan data terbaik berdasarkan ukuran impuritas. Salah satu metode yang paling banyak digunakan dalam algoritma *Random Forest* adalah *Gini Index*, yaitu ukuran yang menunjukkan tingkat ketidakmurnian data pada suatu node. Semakin kecil nilai *Gini Index*, maka semakin baik kualitas pemisahan data yang dihasilkan sehingga atribut tersebut dipilih sebagai dasar pembentukan percabangan berikutnya[23].

Persamaan *Gini Index* dapat dinyatakan sebagai berikut.

$$Gini_{split} = \frac{N_{left}}{N} Gini_{left} + \frac{N_{right}}{N} Gini_{right}$$

Keterangan:

N = jumlah seluruh data pada node
NleftN_{left}Nleft = jumlah data pada cabang kiri
NrightN_{right}Nright = jumlah data pada cabang kanan
GinileftGini_{left}Gini left = nilai Gini cabang kiri
GinirightGini_{right}Giniright = nilai Gini cabang kana

Nilai *Gini Index* berada pada rentang 0 hingga 1. Nilai yang mendekati 0 menunjukkan bahwa seluruh data pada *node* berasal dari kelas yang sama (*pure node*), sedangkan nilai yang semakin besar menunjukkan bahwa data masih terdiri atas beberapa kelas yang berbeda. Oleh karena itu, algoritma akan memilih atribut yang menghasilkan nilai *Gini Index* paling kecil agar pemisahan data menjadi lebih optimal [23], [24].

Pada penelitian ini, setiap *Decision Tree* dibangun menggunakan data hasil *bootstrap sampling* dan subset fitur yang dipilih secara acak (*random feature selection*). Proses tersebut menghasilkan banyak pohon keputusan dengan struktur yang berbeda-beda sehingga meningkatkan kemampuan generalisasi model dan mengurangi risiko *overfitting*. Selanjutnya, seluruh hasil prediksi dari masing-masing *Decision Tree* digabungkan menggunakan mekanisme *majority voting* untuk menentukan kategori emosi akhir, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad* [24], [25].

2.5.3 *Random Feature Selection*

Random feature selection merupakan salah satu karakteristik utama yang membedakan *Random Forest* dari algoritma *Decision Tree* tunggal. Pada setiap proses pembentukan *node*, algoritma tidak mengevaluasi seluruh fitur yang tersedia, melainkan hanya memilih sebagian fitur secara acak (*random subset of features*). Dari kumpulan fitur tersebut, algoritma kemudian menentukan fitur terbaik yang digunakan sebagai dasar pemisahan (*split*) berdasarkan nilai *Gini Index* terkecil [23], [25] .

Pemilihan fitur secara acak bertujuan untuk meningkatkan keragaman (*diversity*) antar *decision tree*. Apabila seluruh pohon menggunakan fitur yang sama pada setiap proses pemisahan, maka struktur pohon yang dihasilkan akan cenderung serupa sehingga mengurangi manfaat dari pendekatan *ensemble learning*. Dengan adanya *random feature selection*, setiap pohon memiliki pola pembelajaran yang berbeda sehingga mampu meningkatkan kemampuan generalisasi model terhadap data baru [25].

Pada algoritma *Random Forest*, jumlah fitur yang dipilih pada setiap *split* ditentukan oleh parameter *max_features*. Untuk kasus klasifikasi, jumlah fitur yang dipilih umumnya menggunakan akar kuadrat dari jumlah seluruh fitur (*Vp*), dengan *p* menyatakan jumlah total fitur pada dataset. Pendekatan ini menghasilkan keseimbangan antara keragaman model dan kemampuan setiap pohon dalam mempelajari karakteristik data [24].

Dalam penelitian *Speech Emotion Recognition* (*SER*), proses *random feature selection* memberikan keuntungan karena fitur hasil ekstraksi *Mel-Frequency Cepstral Coefficients* (*MFCC*) memiliki dimensi yang relatif tinggi. Pemilihan sebagian fitur secara acak pada setiap percabangan memungkinkan model memanfaatkan berbagai kombinasi koefisien *MFCC* tanpa bergantung pada fitur tertentu saja. Dengan demikian, model menjadi lebih robust terhadap variasi data suara dan mampu menghasilkan proses klasifikasi yang lebih stabil [26].

2.5.4 *Bagging*

Bagging atau *Bootstrap Aggregating* merupakan teknik *ensemble learning* yang digunakan untuk meningkatkan akurasi dan stabilitas model klasifikasi dengan cara menggabungkan hasil prediksi dari beberapa *decision tree*. Pada metode ini, setiap *decision tree* dibangun menggunakan subset data yang berbeda hasil dari proses *bootstrap sampling*. Karena setiap pohon dilatih menggunakan data yang berbeda, model yang dihasilkan memiliki karakteristik pembelajaran yang beragam sehingga mampu mengurangi varians dan risiko *overfitting* yang umumnya terjadi pada *decision tree* tunggal [23], [25].

Setelah seluruh *decision tree* selesai dibangun, masing-masing pohon akan memberikan hasil prediksi terhadap data yang diuji. Pada kasus klasifikasi, hasil akhir ditentukan menggunakan mekanisme *majority voting*, yaitu kelas yang memperoleh jumlah suara terbanyak dari seluruh *decision tree* akan dipilih sebagai hasil prediksi akhir. Pendekatan ini menghasilkan keputusan yang lebih stabil karena kesalahan prediksi pada satu pohon dapat dikoreksi oleh pohon-pohon lainnya [23], [24].

Konsep *bagging* memberikan keuntungan dalam menangani data yang memiliki variasi tinggi atau mengandung *noise*. Dengan menggabungkan banyak

model yang berbeda, *Random Forest* mampu menghasilkan performa klasifikasi yang lebih baik dibandingkan penggunaan satu *decision tree*. Selain itu, teknik ini meningkatkan kemampuan generalisasi model sehingga lebih mampu memberikan prediksi yang akurat pada data baru yang belum pernah digunakan selama proses pelatihan [24].

Pada penelitian *Speech Emotion Recognition* (*SER*), mekanisme *bagging* sangat sesuai digunakan karena karakteristik fitur *Mel-Frequency Cepstral Coefficients* (*MFCC*) memiliki variasi yang cukup tinggi. Penggabungan hasil prediksi dari beberapa *decision tree* memungkinkan model mengenali pola emosi secara lebih konsisten meskipun terdapat variasi suara akibat perbedaan intonasi, tempo bicara, maupun kebisingan lingkungan [26].

2.5.5 *Out-of-Bag (OOB) Error Estimation*

Out-of-Bag (*OOB*) merupakan salah satu mekanisme evaluasi internal yang dimiliki oleh algoritma *Random Forest*. Konsep ini muncul sebagai konsekuensi dari proses *bootstrap sampling*, di mana setiap *decision tree* hanya menggunakan sebagian data pelatihan yang dipilih secara acak dengan metode *sampling with replacement*. Akibatnya, terdapat sejumlah data yang tidak terpilih dalam pembentukan suatu pohon. Data tersebut disebut sebagai *Out-of-Bag* (*OOB*) dan dimanfaatkan sebagai data pengujian untuk mengevaluasi performa pohon tersebut [25], [26].

Pada setiap proses pelatihan, sekitar 63% data digunakan untuk membangun satu *decision tree*, sedangkan sekitar 37% sisanya tidak terpilih dan secara otomatis menjadi data *Out-of-Bag*. Proses ini berlangsung pada setiap pohon sehingga setiap sampel data memiliki kesempatan menjadi data pelatihan maupun data pengujian pada pohon yang berbeda. Dengan demikian, *Random Forest* dapat memperkirakan tingkat kesalahan model tanpa memerlukan pembagian dataset secara terpisah menjadi data pelatihan dan data pengujian [23].

Keunggulan utama metode *Out-of-Bag* adalah kemampuannya memberikan estimasi performa model secara efisien tanpa meningkatkan beban komputasi secara signifikan. Nilai *OOB Error* sering digunakan sebagai indikator awal untuk mengevaluasi kemampuan generalisasi model terhadap data baru. Semakin kecil nilai *OOB Error*, maka semakin baik kemampuan model dalam melakukan klasifikasi terhadap data yang belum pernah dipelajari sebelumnya [24], [26]

2.5.6 *Feature Important*

Feature Importance merupakan salah satu keunggulan utama algoritma *Random Forest* yang digunakan untuk mengukur tingkat kontribusi setiap fitur dalam proses klasifikasi. Nilai *feature importance* diperoleh berdasarkan seberapa besar suatu fitur mampu mengurangi tingkat ketidakmurnian (*impurity*) pada setiap *decision tree* selama proses pembentukan model. Semakin besar kontribusi suatu fitur dalam menghasilkan pemisahan data yang baik, maka semakin tinggi pula nilai *feature importance* yang dimilikinya [24], [25].

Pada algoritma *Random Forest*, nilai *feature importance* dihitung dengan mengakumulasi penurunan nilai *Gini Index* yang dihasilkan oleh setiap fitur pada seluruh *decision tree*. Fitur yang sering digunakan sebagai dasar pemisahan (*split*) dan mampu menghasilkan penurunan *impurity* yang besar akan memperoleh nilai *feature importance* yang lebih tinggi dibandingkan fitur lainnya. Dengan demikian, *feature importance* dapat digunakan untuk mengetahui fitur-fitur yang paling berpengaruh terhadap proses klasifikasi [23].

Analisis *feature importance* memiliki peran penting dalam pengolahan data berdimensi tinggi karena membantu mengidentifikasi fitur yang memberikan kontribusi terbesar terhadap performa model. Selain meningkatkan interpretasi hasil klasifikasi, informasi tersebut juga dapat dimanfaatkan untuk proses seleksi fitur (*feature selection*), sehingga model menjadi lebih sederhana tanpa mengurangi tingkat akurasi secara signifikan [24].

Pada penelitian *Speech Emotion Recognition* (*SER*), analisis *feature importance* sangat bermanfaat karena data suara yang telah diekstraksi menggunakan metode *Mel-Frequency Cepstral Coefficients* (*MFCC*) menghasilkan sejumlah koefisien yang merepresentasikan karakteristik spektral sinyal suara. Setiap koefisien *MFCC* memiliki kontribusi yang berbeda dalam membedakan kategori

emosi. Melalui *feature importance*, algoritma *Random Forest* dapat mengidentifikasi koefisien *MFCC* yang paling berpengaruh terhadap proses klasifikasi emosi [26].

2.5.8 *Hyperparameters Random Forest*

Hyperparameter merupakan parameter yang ditentukan sebelum proses pelatihan (*training*) dimulai dan berfungsi untuk mengatur cara kerja algoritma *Random Forest*. Berbeda dengan parameter model yang dipelajari secara otomatis selama proses pelatihan, *hyperparameter* ditetapkan oleh pengguna untuk mengontrol kompleksitas model, proses pembentukan *decision tree*, serta mekanisme klasifikasi. Pemilihan *hyperparameter* yang tepat berpengaruh terhadap tingkat akurasi, kemampuan generalisasi, dan efisiensi komputasi dari model *Random Forest* [25], [26].

Beberapa *hyperparameter* yang umum digunakan pada algoritma *Random Forest* dijelaskan sebagai berikut.

1. *n_estimators*

n_estimators merupakan parameter yang menentukan jumlah *decision tree* yang akan dibangun dalam satu model *Random Forest*. Semakin banyak jumlah pohon yang digunakan, maka hasil klasifikasi umumnya menjadi lebih stabil dan memiliki kemampuan generalisasi yang lebih baik. Namun, peningkatan jumlah pohon juga akan meningkatkan waktu pelatihan (*training time*) dan kebutuhan komputasi [23].

2. *max_depth*

max_depth merupakan parameter yang mengatur kedalaman maksimum setiap *decision tree*. Nilai yang terlalu besar dapat menyebabkan model mengalami *overfitting*, sedangkan nilai yang terlalu kecil dapat menyebabkan

underfitting karena model tidak mampu mempelajari pola data secara optimal [24].

3. *max_features*

max_features menentukan jumlah fitur yang dipilih secara acak pada setiap proses pemisahan (*split*) di dalam *decision tree*. Pemilihan sebagian fitur secara acak bertujuan untuk meningkatkan keragaman antar pohon sehingga mampu mengurangi korelasi antar *decision tree* dan meningkatkan performa model secara keseluruhan [25].

4. *min_samples_split*

min_samples_split merupakan jumlah minimum sampel yang diperlukan agar suatu *node* dapat dibagi menjadi dua *child node*. Parameter ini digunakan untuk mengontrol proses pembentukan percabangan sehingga pohon yang dihasilkan tidak terlalu kompleks dan mampu mengurangi risiko *overfitting* [24].

5. *min_samples_leaf*

min_samples_leaf menentukan jumlah minimum sampel yang harus dimiliki oleh setiap *leaf node*. Pengaturan parameter ini membantu menghasilkan struktur pohon yang lebih seimbang dan meningkatkan kemampuan model dalam melakukan generalisasi terhadap data baru [26].

6. *random_state*

random_state merupakan parameter yang digunakan untuk mengontrol proses pengacakan (*randomization*) selama pembentukan *Random Forest*. Penggunaan nilai *random_state* yang tetap (*fixed seed*) memungkinkan proses pelatihan menghasilkan model yang konsisten sehingga penelitian dapat direplikasi dengan hasil yang sama [23].

2.6 *Convolutional Neural Network (CNN)*

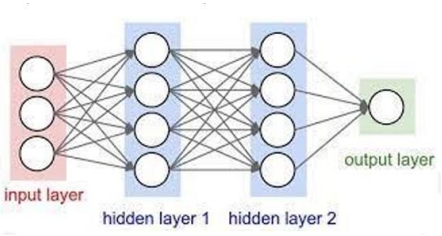
Convolutional Neural Network (*CNN*) merupakan salah satu metode *deep learning* yang dirancang khusus untuk memproses data berbentuk citra. Berbeda dengan jaringan saraf tiruan konvensional (*Multilayer Perceptron*/*MLP*), *CNN* mampu mempertahankan informasi spasial pada citra sehingga proses pembelajaran menjadi lebih efektif dalam mengenali pola visual. Kemampuan tersebut menjadikan *CNN* sebagai salah satu algoritma yang paling banyak digunakan dalam berbagai aplikasi *computer vision*, seperti klasifikasi citra, deteksi objek, segmentasi citra, serta *Facial Emotion Recognition* (*FER*) [27], [28].

Arsitektur *CNN* terdiri atas beberapa lapisan (*layer*) yang bekerja secara bertahap untuk mengekstraksi dan mengklasifikasikan informasi dari citra masukan. Secara umum, proses pada *CNN* dibagi menjadi dua tahap utama, yaitu *feature learning* dan *classification*. Tahap *feature learning* bertujuan mengekstraksi karakteristik penting dari citra menggunakan beberapa lapisan, seperti *Convolution Layer*, fungsi aktivasi *Rectified Linear Unit* (*ReLU*), dan *Pooling Layer*. Selanjutnya, pada tahap *classification*, fitur hasil ekstraksi diteruskan menuju *Fully Connected Layer* untuk menghasilkan probabilitas setiap kelas menggunakan fungsi aktivasi *Softmax* [27], [28].

Keunggulan utama *CNN* dibandingkan metode klasifikasi citra konvensional adalah kemampuannya melakukan ekstraksi fitur secara otomatis (*automatic feature extraction*). Model tidak memerlukan proses rekayasa fitur (*feature engineering*) secara manual karena filter konvolusi akan mempelajari pola-pola penting, seperti tepi (*edges*), tekstur, bentuk, maupun karakteristik objek selama proses pelatihan. Semakin dalam arsitektur jaringan, semakin kompleks pula fitur yang dapat dipelajari oleh *CNN* [27], [29].

Pada sistem *Facial Emotion Recognition* (*FER*), *CNN* mampu mengenali perubahan karakteristik wajah, seperti posisi alis, bentuk mata, kontur hidung,

serta perubahan bentuk mulut yang berkaitan dengan ekspresi emosi. Informasi visual tersebut dipelajari secara bertahap sehingga model mampu membedakan berbagai kategori emosi dengan tingkat akurasi yang tinggi. Oleh karena itu, *CNN* menjadi salah satu metode yang paling banyak digunakan dalam penelitian deteksi emosi berbasis citra wajah [28], [30].



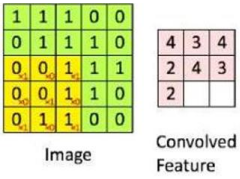
Gambar 2. 4 *Arsitektur MLP*

2.6.1 *Convolution Layer*

Convolution Layer merupakan lapisan utama pada *Convolutional Neural Network* (*CNN*) yang berfungsi untuk mengekstraksi karakteristik penting (*feature extraction*) dari citra masukan. Pada lapisan ini, citra diproses menggunakan sejumlah filter (*kernel*) yang berukuran lebih kecil daripada citra input. Setiap *kernel* digeser secara sistematis pada seluruh area citra untuk menghasilkan *feature map* yang merepresentasikan pola-pola tertentu, seperti tepi (*edges*), sudut (*corners*), tekstur, maupun bentuk objek [27], [28].

Operasi konvolusi dilakukan dengan mengalikan setiap elemen *kernel* terhadap nilai piksel pada area citra yang bersesuaian, kemudian menjumlahkan seluruh hasil perkalian tersebut untuk menghasilkan satu nilai keluaran. Proses ini dilakukan secara berulang hingga seluruh area citra diproses sehingga terbentuk *feature map*. Setiap *kernel* memiliki bobot (*weights*) yang dipelajari secara

otomatis selama proses pelatihan, sehingga *CNN* mampu mengenali berbagai karakteristik visual tanpa memerlukan proses ekstraksi fitur secara manual [28].



Gambar 2. 5 Operasi *Konvolusi*

Pada Gambar 2.5 kotak kuning merupakan kernel dan kotak hijau offset dari data citra. Kernel bergerak dari kiri atas menuju kanan bawah. Konvolusi bertujuan untuk mengekstraksi fitur dari citra input. Konvolusi akan menghasilkan transformasi linear dari data input sesuai informasi spasial pada data. Bobot pada layer tersebut menspesifikasikan kernel konvolusi yang digunakan, sehingga kernel konvolusi dapat dilatih berdasarkan input pada *CNN*.

2.6.2 *Activation Function (Relu)*

Setelah citra diproses pada *Convolution Layer*, hasil konvolusi akan diteruskan ke *Activation Layer*. Fungsi aktivasi berperan untuk memperkenalkan sifat nonlinier (*non-linearity*) ke dalam jaringan sehingga *Convolutional Neural Network* (*CNN*) mampu mempelajari hubungan yang kompleks pada data citra. Tanpa fungsi aktivasi, seluruh lapisan pada jaringan hanya akan membentuk transformasi linear sehingga kemampuan model dalam mengenali pola visual menjadi sangat terbatas [27], [28].

Fungsi aktivasi yang paling banyak digunakan pada arsitektur *CNN* adalah *Rectified Linear Unit* (*ReLU*). Fungsi *ReLU* memiliki proses komputasi yang sederhana dan mampu mempercepat proses pelatihan dibandingkan fungsi

aktivasi lain, seperti *Sigmoid* maupun *Hyperbolic Tangent* (*Tanh*). Selain itu, *ReLU* juga mampu mengurangi permasalahan *vanishing gradient* yang sering terjadi pada jaringan saraf dengan banyak lapisan (*deep neural network*) [28], [31].

Penggunaan *ReLU* juga menghasilkan representasi fitur yang lebih jarang (*sparse representation*), karena sebagian neuron akan menghasilkan nilai nol. Kondisi tersebut mampu mengurangi kompleksitas komputasi serta meningkatkan kemampuan model dalam mempelajari karakteristik visual yang penting, seperti bentuk mata, alis, hidung, dan mulut pada proses *Facial Emotion Recognition* (*FER*) [29].

2.6.3 *Pooling Layer*

Pooling Layer merupakan salah satu lapisan pada *Convolutional Neural Network* (*CNN*) yang berfungsi untuk mengurangi dimensi (*downsampling*) dari *feature map* yang dihasilkan oleh *Convolution Layer*. Tujuan utama penggunaan *Pooling Layer* adalah mengurangi jumlah parameter dan kompleksitas komputasi tanpa menghilangkan informasi penting yang telah dipelajari oleh jaringan. Selain itu, proses *pooling* juga membantu meningkatkan kemampuan model dalam mengenali objek meskipun terjadi perubahan posisi, skala, maupun rotasi pada citra [27], [28].

Jenis *pooling* yang paling banyak digunakan pada arsitektur *CNN* adalah *Max Pooling*. Metode ini bekerja dengan membagi *feature map* menjadi beberapa wilayah kecil (*pooling window*), kemudian memilih nilai maksimum dari setiap wilayah sebagai keluaran. Nilai maksimum tersebut diasumsikan mewakili fitur yang paling dominan sehingga informasi penting tetap dipertahankan meskipun ukuran data menjadi lebih kecil [28], [31].

Sebagai contoh, apabila digunakan ukuran *pooling window* sebesar **2 × 2** dengan *stride* sebesar **2**, maka setiap empat nilai piksel pada *feature map* akan direduksi menjadi satu nilai terbesar. Proses ini menghasilkan *feature map* dengan dimensi yang lebih kecil sehingga mempercepat proses pelatihan serta mengurangi penggunaan memori tanpa mengurangi kemampuan model dalam mengenali pola visual [31].

2.6.4 *Batch Normalization*

Batch Normalization merupakan salah satu teknik pada *Convolutional Neural Network* (*CNN*) yang digunakan untuk menormalkan distribusi data pada setiap *mini-batch* selama proses pelatihan. Teknik ini diperkenalkan untuk mengatasi permasalahan *internal covariate shift*, yaitu perubahan distribusi data masukan pada setiap lapisan jaringan yang dapat memperlambat proses pembelajaran. Dengan melakukan normalisasi, proses pelatihan menjadi lebih stabil, konvergensi model lebih cepat, dan performa klasifikasi dapat ditingkatkan [27], [31].

Pada setiap proses pelatihan, *Batch Normalization* menghitung nilai rata-rata (*mean*) dan simpangan baku (*standard deviation*) dari data pada satu *mini-batch*. Selanjutnya, data dinormalisasi sehingga memiliki distribusi dengan rata-rata mendekati nol dan simpangan baku mendekati satu. Setelah proses normalisasi, dilakukan transformasi menggunakan parameter yang dapat dipelajari (*learnable parameters*) agar model tetap memiliki fleksibilitas dalam mempelajari karakteristik data [28].

2.6.5 *Dropout*

Dropout merupakan salah satu teknik regularisasi yang digunakan pada *Convolutional Neural Network* (*CNN*) untuk mengurangi risiko *overfitting* selama proses pelatihan. Teknik ini bekerja dengan cara menonaktifkan (*deactivate*) sebagian neuron secara acak pada setiap iterasi pelatihan sehingga neuron-neuron tersebut tidak berpartisipasi dalam proses *forward propagation* maupun *backpropagation*. Dengan demikian, model tidak bergantung pada neuron tertentu dan mampu mempelajari representasi fitur yang lebih umum (*generalized features*) [27], [31].

Pada setiap proses pelatihan, *Dropout* akan memilih sejumlah neuron berdasarkan probabilitas tertentu untuk dinonaktifkan sementara. Neuron yang tidak dipilih tetap aktif dan digunakan dalam proses pembelajaran. Pada tahap pengujian (*testing*), seluruh neuron akan diaktifkan kembali sehingga model dapat memanfaatkan seluruh informasi yang telah dipelajari selama proses pelatihan [28].

2.6.6 *Fully Connected Layer*

Fully Connected Layer (*FC Layer*) merupakan lapisan akhir pada arsitektur *Convolutional Neural Network* (*CNN*) yang berfungsi menggabungkan seluruh fitur hasil ekstraksi dari lapisan sebelumnya untuk menghasilkan keputusan klasifikasi. Berbeda dengan *Convolution Layer* yang hanya memproses sebagian wilayah citra

menggunakan *kernel*, pada *Fully Connected Layer* setiap neuron terhubung dengan seluruh neuron pada lapisan sebelumnya. Struktur ini memungkinkan jaringan mempelajari hubungan global antar fitur sehingga mampu membedakan setiap kategori kelas secara lebih akurat [27], [28]

Sebelum data diproses oleh *Fully Connected Layer*, keluaran dari lapisan konvolusi dan *pooling* terlebih dahulu diubah menjadi vektor satu dimensi melalui proses *Flatten*. Proses *Flatten* mengubah *feature map* berdimensi dua atau tiga menjadi sebuah vektor sehingga dapat digunakan sebagai masukan bagi jaringan saraf pada *Fully Connected Layer*. Tahap ini menjadi penghubung antara proses ekstraksi fitur (*feature learning*) dan proses klasifikasi (*classification*) [28].

Pada sistem klasifikasi citra, *Fully Connected Layer* berfungsi mengintegrasikan seluruh informasi visual yang telah dipelajari oleh lapisan sebelumnya, seperti bentuk mata, posisi alis, kontur hidung, dan bentuk mulut. Informasi tersebut kemudian digunakan untuk membedakan karakteristik setiap kategori emosi sehingga model mampu menghasilkan keputusan klasifikasi yang tepat[29].

2.6.7 *Softmax* Function

Softmax merupakan fungsi aktivasi yang digunakan pada lapisan keluaran (*output layer*) dalam *Convolutional Neural Network* (*CNN*) untuk menyelesaikan permasalahan klasifikasi multikelas (*multi-class classification*). Fungsi ini mengubah nilai keluaran (*logits*) dari *Fully Connected Layer* menjadi nilai probabilitas yang berada pada rentang 0 hingga 1. Jumlah seluruh probabilitas yang dihasilkan selalu bernilai satu sehingga setiap nilai dapat diinterpretasikan sebagai peluang suatu data termasuk ke dalam kelas tertentu [27], [28].

Pada proses klasifikasi, setiap neuron pada lapisan keluaran menghasilkan nilai yang merepresentasikan tingkat keyakinan model terhadap masing-masing kelas. Fungsi *Softmax* kemudian melakukan normalisasi terhadap nilai-nilai tersebut sehingga kelas dengan probabilitas terbesar dipilih sebagai hasil prediksi akhir. Pendekatan ini sangat sesuai untuk sistem *Facial Emotion Recognition* (*FER*) karena mampu memberikan probabilitas untuk setiap kategori emosi yang diprediksi [27], [28].

2.6.8 *Fully Connected Layer*

Fully connected (FC) layer berfungsi sebagai bagian akhir dari *CNN* yang mengumpulkan seluruh fitur yang telah dipelajari sebelumnya. Neuron pada FC layer memiliki hubungan penuh dengan neuron pada lapisan sebelumnya sehingga lapisan ini mampu menggabungkan pola global dari citra [32].

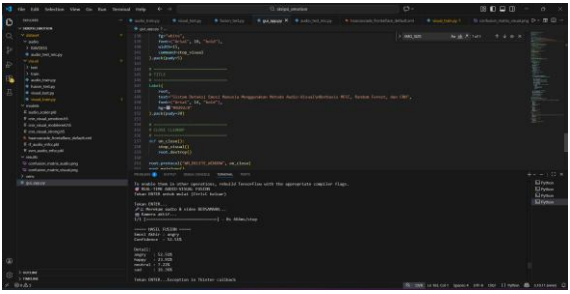
Pada klasifikasi emosi wajah, FC layer akan menentukan hasil akhir berdasarkan fitur ekspresi yang telah dikenali pada lapisan-lapisan sebelumnya.

2.7 Bahasa Pemrograman Python

Python merupakan bahasa pemrograman tingkat tinggi (*high-level programming language*) yang bersifat *interpreted*, *object-oriented*, dan *open source*. Python dikembangkan dengan sintaks yang sederhana sehingga mudah dipelajari serta banyak digunakan dalam pengembangan perangkat lunak, analisis data, *artificial intelligence*, *machine learning*, maupun *deep learning*. Dukungan komunitas yang luas serta ketersediaan berbagai pustaka (*library*) menjadikan Python sebagai salah satu bahasa pemrograman yang paling populer dalam penelitian berbasis kecerdasan buatan [29], [30].

Dalam bidang *machine learning* dan *computer vision*, Python menyediakan berbagai *library* yang mendukung proses pengolahan data, ekstraksi fitur, pelatihan model, hingga visualisasi hasil. Framework seperti TensorFlow, Keras, Scikit-learn, OpenCV, NumPy, dan Librosa memungkinkan pengembangan sistem pengenalan emosi secara lebih efisien dibandingkan bahasa pemrograman lainnya [27], [30].

Pada penelitian ini, Python digunakan sebagai bahasa pemrograman utama dalam pengembangan sistem *Audio-Visual Emotion Recognition* (*AVER*). Seluruh proses, mulai dari akuisisi data suara dan citra, ekstraksi fitur *Mel-Frequency Cepstral Coefficients* (*MFCC*), pelatihan model *Random Forest* dan *Convolutional Neural Network* (*CNN*), hingga implementasi antarmuka pengguna (*Graphical User Interface*/*GUI*) dilakukan menggunakan Python. Pemilihan Python didasarkan pada kemudahan implementasi, kompatibilitas dengan berbagai *library machine learning*, serta kemampuannya dalam mendukung pengembangan sistem secara *real-time*.



Gambar 2. 6 Pemograman Python pada Visual Studio Code

2.7.1 *Visual Studio Code*

Visual Studio Code (VSCode) adalah sebuah editor kode sumber teks yang sangat populer yang dikembangkan oleh Microsoft. VSCode dirancang untuk menjadi ringan, cepat, dan sangat dapat disesuaikan, sehingga memungkinkan para pengembang dan programmer untuk mengembangkan berbagai jenis proyek perangkat Lunak.

2.7.2 *Library* pada *Python*

A. *Librosa*

Librosa merupakan *library* Python yang digunakan untuk analisis sinyal suara dan ekstraksi fitur audio. *Librosa* menyediakan berbagai metode pemrosesan sinyal, seperti ekstraksi *Mel-Frequency Cepstral Coefficients* (*MFCC*), spektrogram, *Zero Crossing Rate* (*ZCR*), dan *Chroma Features*. Pada penelitian ini, *Librosa* digunakan untuk mengekstraksi 40 koefisien *MFCC* dari sinyal suara sebelum dilakukan proses klasifikasi menggunakan *Random Forest* [25].

B. Soundfile

SoundFile digunakan untuk membaca dan menyimpan berkas audio dalam berbagai format, khususnya format WAV. Pada penelitian ini, *SoundFile* digunakan untuk menyimpan hasil perekaman suara sebelum dilakukan proses ekstraksi fitur menggunakan *Librosa* [29].

C. NumPy

NumPy merupakan *library* komputasi numerik yang menyediakan struktur data *array* multidimensi dan berbagai fungsi matematika. Pada penelitian ini, *NumPy* digunakan untuk mengolah data hasil ekstraksi *MFCC*, melakukan operasi matriks, serta mendukung proses komputasi pada model *Random Forest* dan *CNN* [27].

D. Py*Audio*

Py*Audio* adalah perpustakaan Python yang digunakan untuk berinteraksi dengan perangkat *Audio* dan mikrofon komputer. Perpustakaan ini memungkinkan untuk merekam *Audio* dari mikrofon, memainkan suara melalui speaker, serta melakukan operasi terkait *Audio* lainnya dalam aplikasi Python [26].

E. Tensor Flow

TensorFlow merupakan *framework deep learning* yang digunakan untuk membangun, melatih, dan mengevaluasi model *Convolutional Neural Network*. Keras yang terintegrasi di dalam TensorFlow menyediakan antarmuka tingkat tinggi sehingga mempermudah proses perancangan arsitektur *CNN* [30].

F. OS *(Operation System)*

Library ini menyediakan fungsi-fungsi untuk berinteraksi dengan sistem operasi (OS), seperti mengelola file dan direktori .

G. Skitch Leaning

Scikit-learn merupakan *library machine learning* yang menyediakan berbagai algoritma klasifikasi, termasuk *Random Forest*. Pada penelitian ini, Scikit-learn digunakan untuk membangun model *RandomForestClassifier*, melakukan pembagian dataset, serta mengevaluasi performa model menggunakan *Confusion Matrix* dan *Classification Report* [25].

H. OpenCV

OpenCV merupakan *library computer vision* yang digunakan untuk menangkap citra dari kamera, mendeteksi wajah, serta melakukan *pre-processing* citra sebelum diproses oleh model *CNN*. Library ini mendukung implementasi sistem deteksi emosi secara *real-time* menggunakan kamera *webcam* [28].

I. Joblib

Library ini menyediakan fungsionalitas untuk menyimpan dan memuat objek Python yang besar ke/dari disk secara efisien .

J. Matplotlib

Library yang umum digunakan untuk *Visual*isasi data dalam bentuk grafik 2D, seperti plot garis, scatter plot, histogram, dll

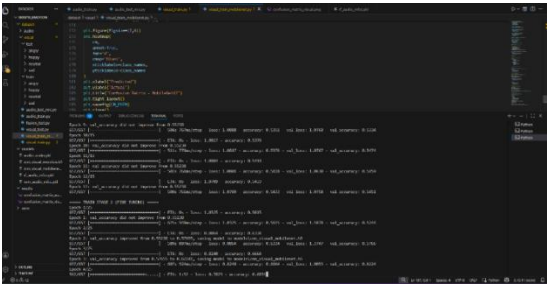
K. Tkinter

Tkinter merupakan *library* standar Python yang digunakan untuk membangun *Graphical User Interface* (*GUI*). Pada penelitian ini, *Tkinter* digunakan untuk merancang antarmuka aplikasi sehingga pengguna dapat melakukan proses deteksi emosi melalui kamera dan mikrofon secara lebih mudah.

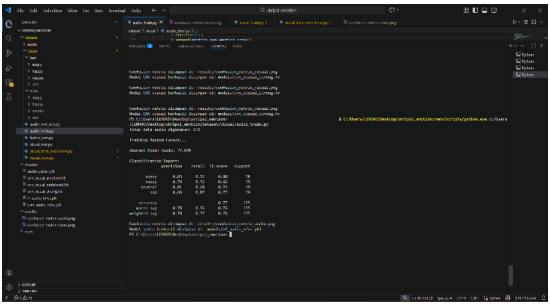
2.8 Evaluasi Model pada Classifier Learning

2.8.1 Train, Test, Data Validation

Evaluasi model merupakan tahapan yang dilakukan untuk mengukur kemampuan algoritma dalam melakukan klasifikasi terhadap data yang belum pernah dipelajari sebelumnya. Proses evaluasi bertujuan untuk mengetahui tingkat akurasi, kemampuan generalisasi, serta performa model dalam mengenali setiap kelas pada dataset. Pada penelitian berbasis *machine learning* dan *deep learning*, evaluasi model umumnya dilakukan menggunakan pembagian dataset menjadi data pelatihan (*training*), validasi (*validation*), dan pengujian (*testing*), serta menggunakan metrik evaluasi seperti *Confusion Matrix*, *Accuracy*, *Precision*, *Recall*, dan *F1-Score* [30], [31].



Gambar 2. 7 Contoh hasil dari percobaan Train.



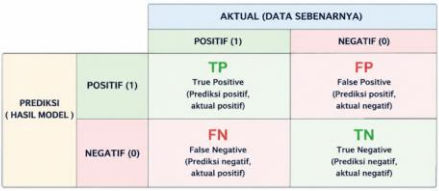
Gambar 2. 8 Contoh hasil dan percobaan *Train*

2.8.2 Con*Fusion* Matrix

Pada proses pengembangan model *machine learning*, dataset umumnya dibagi menjadi tiga bagian utama, yaitu *training dataset*, *validation dataset*, dan *testing dataset*. Pembagian tersebut bertujuan agar model dapat mempelajari pola data, melakukan penyesuaian parameter, serta dievaluasi menggunakan data yang belum pernah digunakan selama proses pelatihan [31].

Training dataset merupakan kumpulan data yang digunakan untuk melatih model agar mampu mengenali pola hubungan antara fitur masukan dan kelas keluaran. Selama proses pelatihan, parameter model diperbarui secara bertahap untuk meminimalkan kesalahan prediksi.

Testing dataset merupakan kumpulan data yang digunakan setelah proses pelatihan selesai. Data ini tidak pernah digunakan selama proses pelatihan maupun validasi sehingga dapat memberikan gambaran mengenai kemampuan generalisasi model terhadap data baru [29].



Gambar 2. 9 *ConFusion Matrix*

Pada Gambar 2.6 dapat dijelaskan dengan *True positive* (TP) yaitu jumlah data yang kelas aktualnya kelas positif dan kelas prediksinya kelas positif. Dapat di interpretasikan dengan memprediksi positif dan kenyataannya benar. *False negative* (FN) yaitu jumlah data yang kelas aktualnya kelas positif dan kelas prediksinya kelas negatif. Dapat di interpretasikan dengan memprediksi negatif dan kenyataannya salah. *False positive* (FP) yaitu jumlah data yang kelas aktualnya kelas negatif dan kelas prediksinya kelas positif. Dapat di interpretasikan dengan memprediksi positif dan kenyataannya salah. *True negative* (TN) yaitu jumlah data yang kelas aktualnya kelas negatif dan kelas prediksinya kelas negatif. Dapat di interpretasikan dengan memprediksi negatif dan kenyataannya benar.

A. Akurasi

Accuracy menunjukkan persentase jumlah data yang berhasil diklasifikasikan dengan benar terhadap seluruh data pengujian., Untuk penghitungan dapat dilihat pada Persamaan di bawah.

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

B. Presisi

Precision menunjukkan tingkat ketepatan model dalam memberikan prediksi positif.Untuk penghitungan dapat dilihat pada Persamaan di bawah.

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

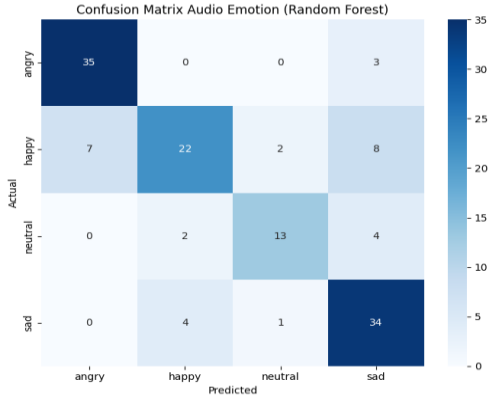
C. Recall

Ijiji *Recall* menunjukkan kemampuan model dalam menemukan seluruh data yang benar-benar termasuk ke dalam kelas tertentu. Untuk penghitungan dapat dilihat pada Persamaan di bawah.

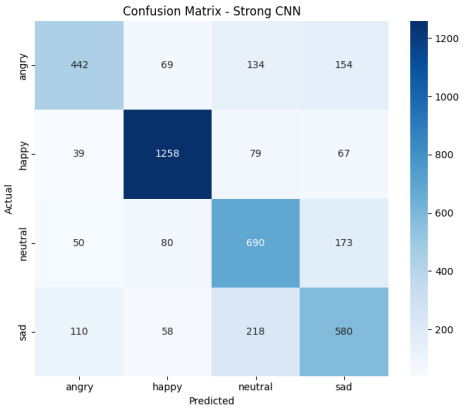
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

D. F1-Scroe

F1-Score merupakan rata-rata harmonik antara Precision dan Recall. Metrik ini banyak digunakan pada kasus klasifikasi yang memiliki distribusi kelas tidak seimbang karena mampu memberikan gambaran performa model secara lebih komprehensi



Gambar 2. 10 Contoh *ConFusion Matrix* dari *Audio Emotion*



Gambar 2. 11 Contoh *ConFusion Matrix* dari *Video Emotion*

2.9 Penelitian Terbaru

Pada bagian ini disajikan beberapa penelitian terdahulu yang berkaitan dengan *Speech Emotion Recognition*, *Facial Emotion Recognition*, dan *Audio-Visual Emotion Recognition*. Studi literatur tersebut digunakan sebagai dasar dalam menentukan metode, algoritma, dataset, serta pendekatan yang digunakan pada penelitian ini. Selain itu, perbandingan beberapa penelitian sebelumnya memberikan gambaran mengenai perkembangan teknologi deteksi emosi serta menunjukkan kontribusi penelitian yang dilakukan.

Berikut merupakan table study literatur yang mana jadi acuan dalam pembuatan alat dan dasar pembuatan dalam skripsi. Dan ada table perbandingan dari beberapa journal sebelumnya.

Tabel 2. 1 Studi Literature Acuan

No	Penulis	Tahun	Metode	Dataset	Hasil Penelitian
1	Khaireddin & Chen	2021	CNN	FER2013	CNN berhasil meningkatkan performa pengenalan ekspresi wajah dan mencapai akurasi

					73,28% pada dataset FER2013. (arXiv)
2	Sharafi et al.	2022	CNN + LSTM + Audio-Visual Fusion	SAVEE, RAVDESS, RML	Pendekatan multimodal mampu meningkatkan performa pengenalan emosi dengan akurasi hingga 94,99% pada dataset RAVDESS. (ScienceDirect)
3	Alsharhan et al.	2023	MFCC + Random Forest + CNN	RAVDESS, REMA-D, TESS, SAVEE	Perbandingan menunjukkan CNN memiliki performa lebih tinggi dibandingkan Random Forest, sedangkan Random Forest tetap efektif untuk klasifikasi berbasis MFCC. (PLOS)
4	Waleed & Shaker	2025	CNN	MELD, RAVDESS	CNN menghasilkan akurasi 91,90% pada pengenalan emosi berbasis suara menggunakan dataset RAVDESS. (MDPI)
5	Ouyang	2025	MFCC + CNN-LSTM	SAVEE, RAVDESS	Kombinasi MFCC dan CNN-LSTM mampu mengenali emosi suara dengan memanfaatkan fitur MFCC sebagai representasi sinyal audio. (arXiv)
6	Recent Multimodal Emotion Recogniti on	2025	Multi-View Audio-Visual Fusion	RAVDESS	Pendekatan multimodal menunjukkan akurasi hingga 95,83%, membuktikan bahwa penggabungan audio dan visual memberikan performa lebih baik dibandingkan modalitas tunggal. (PMC)
7	Priyadarshini, M. A., et al.	2024	MFCC + CNN	RAVDESS ,CREMA-D, TESS, SAVEE	Menggunakan ekstraksi fitur MFCC dan CNN dengan data augmentation untuk meningkatkan performa Speech Emotion Recognition pada beberapa dataset.
8	Tang, X., Lin, Y., Dang, T.,	2024	CNN+ Transformer	IEMOCAP, Emo-DB	Menggabungkan CNN dan Transformer sehingga mampu meningkatkan performa

	Zhang, Y., & Cheng, J.				pengenalan emosi suara dibandingkan metode CNN konvensional.
9	Reghunathan, R. K., et al.	2024	CNN (Pre-trained Architecture)	FER2013	Penggunaan model CNN pralatih (pre-trained) mampu meningkatkan performa Facial Expression Recognition dan mendukung implementasi real-time.
10	Tang, S., et al.	2024	MFCC + 1D-CNN	EMO-DB, EMOVO, SAVEE, RAVDESS	Kombinasi MFCC dengan 1D-CNN menghasilkan akurasi hingga **98,1%** pada beberapa dataset dan menunjukkan peningkatan dibandingkan penggunaan MFCC saja.

Tabel 2. 2 Perbandingan antar Journal**.**

Spesifikasi	**Sistem yang Dikembangkan**	**Jurnal 1** **Khairuddin & Chen (2021)**	**Jurnal 2** **Sharafi et al. (2022)**	**Jurnal 3** **Alsharhan et al. (2023)**	**Jurnal 4** **Waleed & Shaker (2025)**	**Jurnal 5** **Ouyang (2025)**
Metode Input	Audio + Wajah	Wajah	Audio + Wajah	Audio	Audio	Audio
Metode/ Fitur	MFCC + Random Forest + CNN + Weighted Late Fusion	CNN	CNN + LSTM + Audio-Visual Fusion	MFCC + Random Forest + CNN	CNN	MFCC + CNN-LSTM

Dataset	RAVDESS & FER2013	FER2013	SAVEE, RAVDESS, RML	RAVDESS, CREMA-D, TESS, SAVEE	MELD, RAVDESS	SAVEE, RAVDESS
Realtime	Ya	Tidak	Tidak	Tidak	Ya	Tidak
Akurasi	Audio **75,56%**Visual **87,48%**	73,28%	94,99%	CNN lebih baik daripada Random Forest	91,90%	>90%
Kompleksitas	Sedang	Sedang	Tinggi	Sedang	Tinggi	Tinggi
Usability	Tinggi	Sedang	Sedang	Sedang	Tinggi	Sedang
Interface	GUI (Python Tkinter)	Tidak disebutkan	Tidak disebutkan	Tidak disebutkan	Tidak disebutkan	Tidak disebutkan
Kelebihan	Mampu melakukan deteksi emosi secara real-time menggunakan	Akurat untuk Facial Expression Recognition	Multimodal meningkatkan performa	Random Forest cepat dan CNN akurat	Akurasi tinggi pada Speech Emotion Recognition	CNN-LSTM mampu menangkap pola temporal suara

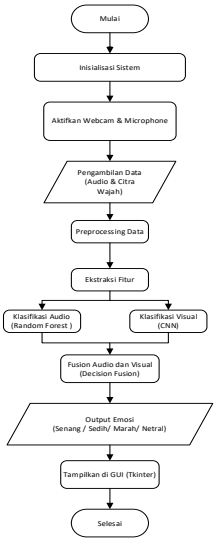
	audio dan visual					
Kekurangan	Membutuhkan webcam dan mikrofon	Dipengaruhi pencahayaan dan pose wajah	Komputasi tinggi	Sensitif terhadap kualitas fitur MFCC	Membutuhkan GPU dan komputasi tinggi	Waktu pelatihan relatif lebih lama

Bab 3. Metodologi Penelitian

3.1. Perancangan

Perancangan sistem pada penelitian ini mencakup penyusunan tahapan penelitian untuk menjawab rumusan masalah yang telah ditetapkan, serta perancangan perangkat lunak yang digunakan dalam pengolahan dan klasifikasi emosi berbasis *Audio* dan citra wajah.

Perangkat keras yang digunakan hanya berfungsi sebagai media input, yaitu webcam dan *microphone* untuk memperoleh data. Alur kerja sistem di*Visual*isasikan dalam bentuk diagram alir (*flowchart*) yang menggambarkan proses mulai dari pengambilan data hingga menghasilkan output berupa klasifikasi emosi.



Gambar 3. 1 *Flowchart* Perancangan Sistem

3.2 Alat dan Bahan

Alat dan bahan yang digunakan pada penelitian ini terdiri dari p erangkat keras (*Hardware*) yakni : Laptop merk Lenovo Ideapad Gaming 4 dengan RAM 4GB dan *AMD Ryzen 5 4500H* , kamera laptop *SunplusIT* microfone Laptop *Microphone array.*

Lalu, perangkat Lunak (*Software*) yang digunakan terdiri dari *Operating System* Windows 10 64-bit, *Visual* Studio Code, Python 3.8.10,dengan *Library* yakni; *Librosa, Soundfile, OS, Glob, NumPy, Scikit-learn (Sklearn), Scipy, PyAudio , Wave, Matplotlib, Joblib, Serial, Time, Warning, Imblearn, Tkinter*

3.3 Tahap dan Alur Penelitian

Tahapan penelitian dalam sistem deteksi emosi ini disusun secara sistematis untuk menghasilkan model yang mampu mengenali emosi manusia berdasarkan *Audio* dan citra wajah. Secara umum, alur penelitian dimulai dari pengumpulan data hingga evaluasi model.

3.3.1 Pembuatan Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset RAVDESS (Ryerson *Audio* -*Visual* Database of Emotional Speech and Song), yang berisi rekaman *Audio* dan video dari 24 aktor profesional dengan berbagai ekspresi emosi.

Dataset ini mencakup beberapa kategori emosi, namun pada penelitian ini difokuskan pada empat emosi utama, yaitu:

- Senang
- Sedih
- Marah
- Netral

Data yang digunakan berupa:

- *Audio* (format .wav)
- Citra wajah (frame dari video atau dataset wajah)

Preprocessing Data

Pada tahap preprocessing data, dilakukan serangkaian proses awal untuk mempersiapkan data sebelum digunakan pada tahap ekstraksi fitur dan pelatihan model.

Langkah pertama adalah melakukan pemanggilan *library* Python yang diperlukan untuk mendukung proses pengolahan data, ekstraksi fitur, pelatihan model, serta evaluasi sistem. *Library* yang digunakan antara lain mencakup pengolahan *Audio* , pengolahan citra, dan machine learning.

Selanjutnya, dilakukan penentuan parameter dan konfigurasi sistem, termasuk pemilihan kategori emosi yang akan dianalisis, yaitu senang, sedih, marah, dan netral. Setiap kategori emosi kemudian diberikan label numerik untuk mempermudah proses klasifikasi oleh model.

Dataset yang digunakan dimuat dari sumber RAVDESS, kemudian dilakukan proses pemetaan kode emosi ke dalam label yang telah ditentukan. Tahap ini bertujuan untuk memastikan kesesuaian antara data dan label yang digunakan dalam proses pelatihan model.

Ekstraksi Fitur

Tahap ekstraksi fitur merupakan proses untuk mengambil informasi penting dari data *Audio* dan citra wajah yang digunakan dalam penelitian. Proses ini bertujuan untuk merepresentasikan karakteristik data dalam bentuk numerik sehingga dapat diproses oleh model klasifikasi.

Pada data *Audio* , ekstraksi fitur dilakukan menggunakan metode *Mel-Frequency Cepstral Coefficients (MFCC)*. Metode ini digunakan karena mampu merepresentasikan karakteristik suara yang berkaitan dengan persepsi pendengaran manusia. Hasil dari proses ini berupa vektor fitur yang menggambarkan pola akustik dari sinyal suara.

Sementara itu, pada data citra wajah, ekstraksi fitur dilakukan menggunakan metode *Convolutional Neural Network* (*CNN*). *CNN* mampu mengekstraksi fitur secara otomatis dari citra dengan memanfaatkan operasi konvolusi untuk menangkap pola spasial seperti bentuk mata, mulut, dan ekspresi wajah lainnya. Hasil dari proses ini berupa *feature map* yang merepresentasikan ciri *Visual* dari ekspresi wajah.

Perancangan Model

Pada tahap perancangan model, dilakukan pembangunan sistem klasifikasi emosi berdasarkan data *Audio* dan citra wajah yang telah melalui proses ekstraksi fitur. Model dirancang secara terpisah untuk masing-masing jenis data, kemudian hasilnya akan digabungkan pada tahap selanjutnya.

Untuk data *Audio* , model klasifikasi yang digunakan adalah algoritma *Random Forest*. Model ini dipilih karena memiliki kemampuan yang baik dalam menangani data dengan jumlah fitur yang cukup banyak serta mampu mengurangi risiko overfitting. Input yang digunakan pada model ini berupa fitur *MFCC* yang

telah diekstraksi sebelumnya, kemudian diproses untuk menghasilkan prediksi emosi.

Sementara itu, untuk data citra wajah digunakan model *Convolutional Neural Network* (*CNN*). Model ini mampu melakukan klasifikasi citra dengan cara mempelajari pola *Visual* dari ekspresi wajah secara otomatis. *CNN* akan mengolah feature map hasil ekstraksi sebelumnya untuk menentukan kategori emosi berdasarkan karakteristik *Visual* yang terdeteksi.

Pada tahap ini juga dilakukan pembagian dataset menjadi data pelatihan (*Training*) dan data pengujian (*Testing*). Proses pelatihan dilakukan untuk membangun model yang mampu mengenali pola emosi, sedangkan data pengujian digunakan untuk mengevaluasi performa model. Selain itu, dilakukan penyesuaian parameter (hyperparameter tuning) untuk memperoleh hasil klasifikasi yang optimal.

3.3.5 *Fusion* Model

Pada tahap ini dilakukan proses penggabungan hasil klasifikasi dari model *Audio* dan model citra wajah untuk menghasilkan keputusan emosi yang lebih akurat. Metode yang digunakan adalah *decision-level Fusion* atau *late Fusion*, yaitu penggabungan dilakukan pada tahap akhir setelah masing-masing model menghasilkan prediksi.

Model *Audio* yang menggunakan algoritma *Random Forest* akan menghasilkan prediksi emosi berdasarkan fitur suara, sedangkan model citra wajah berbasis *Convolutional Neural Network (CNN)* akan menghasilkan prediksi emosi berdasarkan ekspresi wajah. Kedua hasil prediksi tersebut kemudian digabungkan untuk menentukan keputusan akhir emosi.

Proses penggabungan dilakukan dengan mempertimbangkan hasil prediksi dari kedua model, misalnya melalui metode *majority voting* atau pembobotan (*weighted voting*). Dengan pendekatan ini, sistem dapat memanfaatkan kelebihan masing-masing modalitas sehingga mampu meningkatkan akurasi dibandingkan penggunaan satu jenis data saja.

Tahap *Fusion* ini menjadi komponen penting dalam sistem karena mampu mengatasi keterbatasan yang dimiliki oleh masing-masing model, seperti gangguan suara pada data *Audio* atau kondisi pencahayaan yang kurang baik pada data citra wajah.

3.3.6 Evaluasi Model

Tahap evaluasi model dilakukan untuk mengukur kinerja sistem dalam mendeteksi emosi berdasarkan data *Audio* , citra wajah, maupun hasil penggabungan keduanya. Evaluasi ini bertujuan untuk mengetahui tingkat akurasi serta kemampuan model dalam mengklasifikasikan emosi secara tepat.

Metode evaluasi yang digunakan meliputi beberapa metrik, yaitu *accuracy*, *precision*, dan *recall*. *Accuracy* digunakan untuk mengukur tingkat ketepatan model dalam memprediksi seluruh data, sedangkan *precision* menunjukkan ketepatan model dalam mengklasifikasikan suatu kelas tertentu. *Recall* digunakan untuk mengukur kemampuan model dalam menemukan seluruh data yang termasuk dalam suatu kelas.

Selain itu, digunakan juga *conFusion matrix* untuk memberikan gambaran lebih rinci mengenai hasil klasifikasi model. Con*Fusion* matrix menampilkan jumlah prediksi yang benar dan salah pada setiap kategori emosi, sehingga dapat diketahui kelas mana yang memiliki tingkat kesalahan paling tinggi. Evaluasi dilakukan pada tiga kondisi, yaitu model berbasis *Audio* , model berbasis citra wajah, serta model hasil penggabungan (*Fusion*). Hasil evaluasi ini kemudian digunakan sebagai dasar untuk menganalisis performa sistem serta membandingkan efektivitas masing-masing pendekatan dalam mendeteksi emosi.

3.3 Perancangan *Interface*

Perancangan antarmuka (*Interface*) dilakukan untuk memudahkan pengguna dalam berinteraksi dengan sistem deteksi emosi yang dikembangkan. Antarmuka dirancang menggunakan *library* Tkinter pada bahasa pemrograman Python karena bersifat sederhana, ringan, dan mudah diimplementasikan.

Interface yang dibangun memiliki beberapa fungsi utama, yaitu sebagai media untuk menampilkan hasil deteksi emosi secara real-time serta mengontrol proses pengambilan data dari webcam dan *microphone*. Melalui *Interface* ini, pengguna dapat mengaktifkan kamera dan mikrofon untuk memulai proses deteksi emosi.

Selain itu, *Interface* juga menampilkan hasil klasifikasi emosi dalam bentuk label emosi, seperti senang, sedih, marah, dan netral. Untuk meningkatkan kejelasan informasi, sistem juga dapat menampilkan nilai kepercayaan (*confidence* score) dari hasil prediksi.

Perancangan *Interface* dibuat sesederhana mungkin agar mudah digunakan oleh pengguna tanpa memerlukan pengetahuan teknis yang mendalam. Dengan demikian, sistem diharapkan memiliki tingkat usability yang tinggi serta mampu digunakan secara efektif dalam berbagai kondisi.

3.4 Flowchart *Training* Model

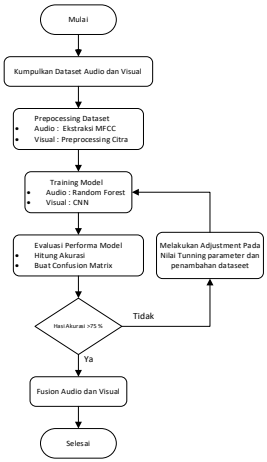
Flowchart *Training* model digunakan untuk menggambarkan proses pelatihan model dalam sistem deteksi emosi berbasis *Audio* . Proses ini dimulai dari tahap pemuatan dataset hingga menghasilkan model yang siap digunakan untuk proses pengujian.

Tahap pertama adalah memuat dataset RAVDESS yang telah disiapkan sebelumnya. Dataset tersebut kemudian diproses untuk mengekstraksi fitur *Audio* menggunakan metode *Mel-Frequency Cepstral Coefficients (MFCC)*. Selain MFCC, dapat digunakan fitur tambahan seperti Mel Spectrogram dan Chroma untuk memperkaya representasi data.

Setelah proses ekstraksi fitur, data dibagi menjadi dua bagian, yaitu data pelatihan (*Training*) dan data pengujian (*Testing*). Data pelatihan digunakan untuk melatih model, sedangkan data pengujian digunakan untuk mengevaluasi performa model.

Selanjutnya dilakukan proses pelatihan menggunakan algoritma *Random Forest*. Pada tahap ini, model dilatih untuk mengenali pola emosi berdasarkan fitur yang telah diekstraksi. Untuk meningkatkan performa model, dilakukan proses penyesuaian parameter (hyperparameter tuning) menggunakan metode seperti Grid Search.

Setelah model dilatih, dilakukan evaluasi menggunakan metrik seperti akurasi dan con*Fusion* matrix untuk mengetahui tingkat keberhasilan model dalam mengklasifikasikan emosi. Model yang telah memiliki performa terbaik kemudian disimpan dan digunakan pada tahap pengujian sistem secara keseluruhan.



Gambar 3. 2*Flowchart Training Model*

3.5 Flowchart *Testing* Model

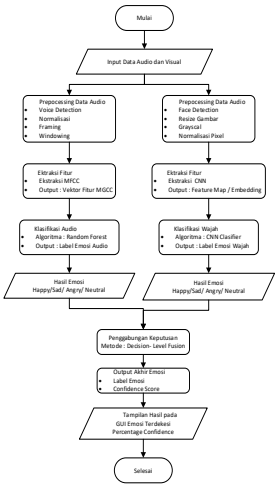
Flowchart *Testing* sistem digunakan untuk menggambarkan proses pengujian sistem deteksi emosi secara keseluruhan setelah model selesai dilatih. Tahap ini bertujuan untuk menguji kemampuan sistem dalam mendeteksi emosi secara real-time berdasarkan input *Audio* dan citra wajah.

Proses dimulai dengan pengambilan data dari pengguna melalui *microphone* dan webcam. Data *Audio* dan citra wajah yang diperoleh kemudian melalui tahap preprocessing untuk memastikan kualitas data yang digunakan dalam proses deteksi. Selanjutnya, dilakukan ekstraksi fitur, di mana data *Audio* diproses menggunakan metode *Mel-Frequency Cepstral Coefficients (MFCC)*, sedangkan citra wajah diproses menggunakan model *Convolutional Neural Network* (*CNN*).

Hasil ekstraksi fitur kemudian diklasifikasikan menggunakan model yang telah dilatih sebelumnya, yaitu *Random Forest* untuk data *Audio* dan *CNN* untuk data citra wajah. Masing-masing model menghasilkan prediksi emosi berdasarkan input yang diterima.

Selanjutnya, dilakukan proses penggabungan hasil klasifikasi menggunakan metode decision-level *Fusion* untuk memperoleh keputusan akhir emosi. Hasil akhir ini kemudian ditampilkan melalui *Interface* dalam bentuk label emosi beserta tingkat kepercayaan (*confidence* score).

Tahap *Testing* ini menjadi bagian penting dalam penelitian karena digunakan untuk mengevaluasi performa sistem dalam kondisi nyata serta memastikan bahwa sistem dapat bekerja sesuai dengan tujuan yang telah ditetapkan.



Gambar 3. 3 Flowchart Testing Model

Bab 4. Hasil dan Pembahasan

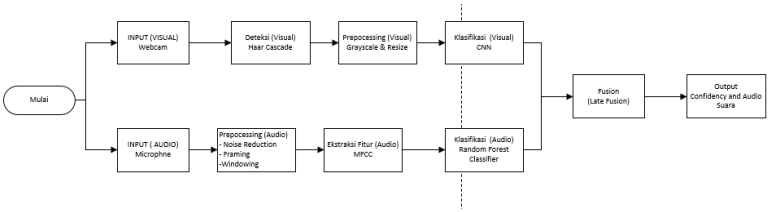
4.1 Implementasi Sistem

Pada penelitian ini telah berhasil dibuat sistem deteksi emosi manusia berbasis *Audio-Visual* menggunakan kombinasi metode *Mel-Frequency Cepstral Coefficients (MFCC)*, *Random Forest*, dan Convolutional Neural Network (*CNN*). Sistem dikembangkan menggunakan bahasa pemrograman Python dengan bantuan beberapa library seperti TensorFlow, OpenCV, Librosa, Scikit-learn, dan Tkinter.

Sistem yang dibangun mampu melakukan deteksi emosi manusia secara realtime melalui dua sumber input utama yaitu suara dan ekspresi wajah. Selain itu, sistem juga menerapkan metode *Fusion* untuk menggabungkan hasil prediksi *Audio* dan *Visual* sehingga menghasilkan prediksi emosi yang lebih stabil.

Secara umum, sistem terdiri dari tiga mode utama yaitu mode *Visual*, mode *Audio*, dan mode *Fusion*. Pada mode *Visual*, sistem mendeteksi emosi berdasarkan ekspresi wajah pengguna menggunakan webcam dan model *CNN*. Pada mode *Audio*, sistem mendeteksi emosi berdasarkan karakteristik suara menggunakan metode MFCC dan algoritma *Random Forest*. Sedangkan pada mode *Fusion*, sistem menggabungkan hasil prediksi *Audio* dan *Visual* menggunakan metode late *Fusion*.

Antarmuka sistem dibuat menggunakan Tkinter sehingga pengguna dapat menjalankan sistem dengan lebih mudah melalui tampilan GUI secara interaktif.



Gambar 4. 1 Sistem Deteksi Emosi *Audio* & *Visual*

4.1.1 Implementasi Sistem Audio

Implementasi sistem *Audio* dilakukan menggunakan microphone sebagai perangkat input suara. Suara yang diterima sistem kemudian diproses menggunakan metode *Mel-Frequency Cepstral Coefficients (MFCC)* untuk mengekstraksi ciri-ciri penting dari sinyal *Audio*.

Metode MFCC digunakan karena mampu merepresentasikan karakteristik suara manusia dengan baik pada domain frekuensi sehingga cocok digunakan pada sistem pengenalan emosi berbasis suara. Pada penelitian ini digunakan algoritma *Random Forest* sebagai classifier untuk melakukan pengenalan emosi berdasarkan fitur MFCC yang diperoleh.

Secara umum, proses kerja MFCC dimulai dari sinyal suara yang diterima microphone. Sinyal suara tersebut kemudian diubah dari domain waktu menjadi domain frekuensi menggunakan proses Fast Fourier Transform (FFT). Setelah itu dilakukan proses Mel Filterbank untuk menyesuaikan karakteristik pendengaran manusia. Hasil proses tersebut kemudian diubah menggunakan Discrete Cosine Transform (DCT) sehingga menghasilkan koefisien MFCC yang digunakan sebagai representasi fitur suara.

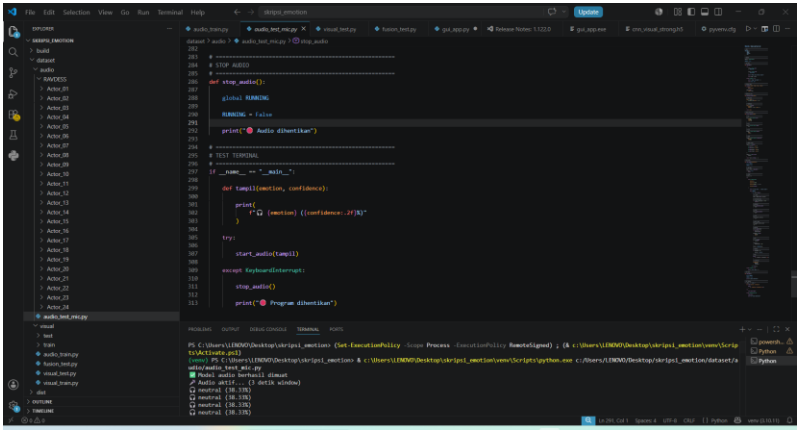
Melalui proses tersebut, MFCC mampu menangkap karakteristik penting dari suara seperti intonasi, energi, dan pola frekuensi yang berbeda pada setiap emosi. Fitur MFCC yang diperoleh kemudian dinormalisasi menggunakan StandardScaler sebelum dilakukan proses klasifikasi menggunakan *Random Forest*.

Pada penelitian ini digunakan dataset RAVDESS dengan total data *Audio* sebanyak 1440 data dan empat kategori emosi yaitu *Angry*, *Happy*, *Neutral*, dan *Sad*.

Tahapan implementasi sistem *Audio* terdiri dari:

1. Pengambilan suara menggunakan microphone
2. Preprocessing *Audio*
3. Ekstraksi fitur MFCC
4. Klasifikasi menggunakan *Random Forest*
5. Menampilkan hasil emosi dan confidence score

Gambar 4. 2 Hasil *Train Audio*



Gambar 4. 3 Implementasi Sistem *Audio* Realtime

Tabel 4. 1 Parameter Sistem *Audio*

Parameter	**Nilai**
Dataset	RAVDESS
Total Dataset *Audio*	1440
Jumlah Emosi	4
Kategori Emosi	*Angry*, *Happy*, *Neutral*, *Sad*
Sampling Rate	16000 Hz
Metode Ekstraksi Fitur	MFCC
Jumlah MFCC	40
Algoritma Klasifikasi	*Random Forest*
Jumlah Estimator	1000
Input Sistem	Microphone
Output Sistem	Emosi dan Confidence Score

Berdasarkan Tabel 4.1, sistem *Audio* menggunakan dataset RAVDESS dengan total data *Audio* sebanyak 1440 data dan empat kategori emosi yaitu *Angry*, *Happy*, *Neutral*, dan *Sad*. Proses ekstraksi fitur dilakukan menggunakan

metode MFCC sebanyak 40 koefisien untuk merepresentasikan karakteristik suara pengguna.

Pada proses klasifikasi digunakan algoritma *Random Forest* dengan jumlah estimator sebanyak 1000 pohon keputusan untuk meningkatkan stabilitas model dalam mengenali emosi suara pengguna secara realtime.

4.1.2 Implementasi Sistem *Visual*

Implementasi sistem *Visual* dilakukan menggunakan webcam sebagai perangkat input untuk mendeteksi ekspresi wajah pengguna secara realtime. Sistem *Visual* menggunakan metode *Convolutional Neural Network (CNN)* untuk melakukan klasifikasi emosi berdasarkan ekspresi wajah.Sebelum proses klasifikasi dilakukan, sistem terlebih dahulu melakukan deteksi wajah menggunakan Haar Cascade Classifier dari OpenCV. Metode ini digunakan untuk menentukan posisi wajah pada frame video yang ditangkap oleh webcam.

Setelah wajah berhasil dideteksi, citra wajah kemudian dilakukan preprocessing berupa konversi citra menjadi grayscale, resize gambar menjadi ukuran 96×96 piksel, dan normalisasi nilai piksel. Tahapan preprocessing dilakukan agar data citra sesuai dengan input model *CNN*.

CNN bekerja dengan melakukan ekstraksi fitur citra secara otomatis melalui beberapa lapisan convolution dan pooling. Pada proses convolution, sistem akan mengambil pola penting dari wajah seperti bentuk mata, mulut, alis, dan ekspresi wajah lainnya. Selanjutnya hasil fitur tersebut digunakan pada proses klasifikasi untuk menentukan kategori emosi pengguna.Pada penelitian ini digunakan empat kategori emosi yaitu *Angry*, *Happy*, *Neutral*, dan *Sad*. Model *CNN* dilatih menggunakan dataset ekspresi wajah sehingga mampu mengenali pola ekspresi dari masing-masing emosi.

Deteksi wajah menggunakan Haar Cascade

1. Preprocessing citra wajah
2. Ekstraksi fitur menggunakan *CNN*
3. Klasifikasi emosi
4. Menampilkan hasil emosi dan confidence score

Sistem *Visual* berhasil dijalankan secara realtime menggunakan webcam laptop. Hasil prediksi emosi kemudian ditampilkan pada GUI beserta confidence score untuk menunjukkan tingkat keyakinan model terhadap hasil klasifikasi.

Tabel 4. 2 Jumlah Datasheet Test

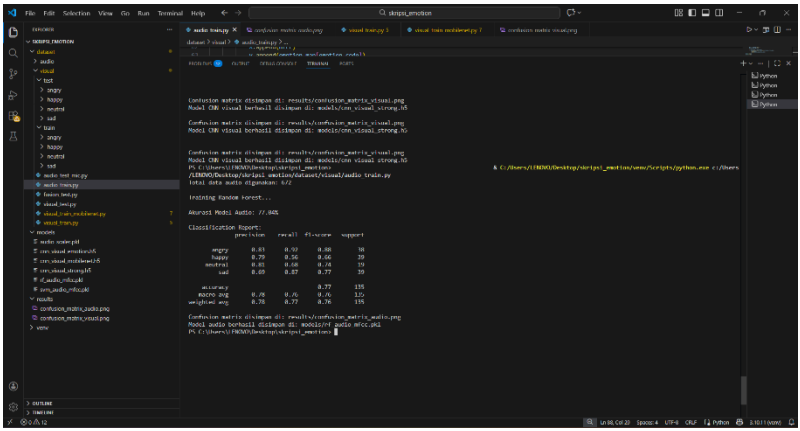
Jenis Emosi	**Jumlah Datasheet (Test)**
Angry	958
Happy	1774
Neutral	1233
Sad	1247

Tabel 4. 3 Jumlah Datasheet *Train*

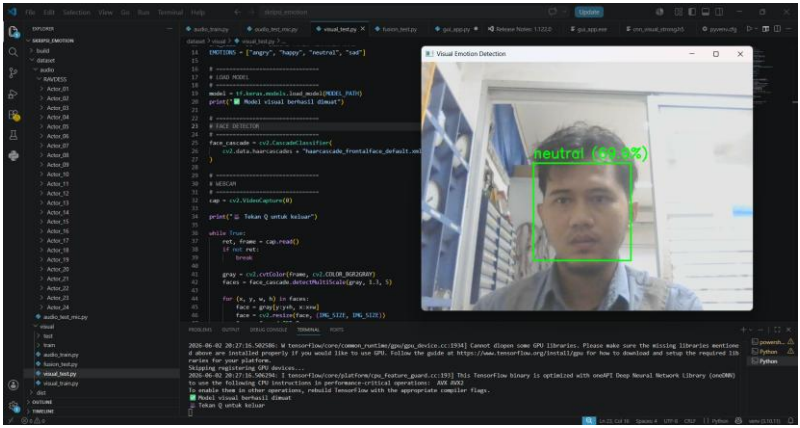
Jenis Emosi	**Jumlah Datasheet (*Train*)**
Angry	3995
Happy	7215
Neutral	4965
Sad	4830

Berdasarkan Tabel 4.2 dan Tabel 4.3, jumlah dataset terbesar terdapat pada kategori *Happy* sedangkan jumlah dataset terkecil terdapat pada kategori *Angry*. Perbedaan jumlah dataset pada masing-masing kategori emosi dapat mempengaruhi performa model *CNN* dalam mengenali ekspresi wajah tertentu.

Total dataset *Visual* yang digunakan pada penelitian ini terdiri dari 21005 data *train*ing dan 5212 data testing. Dataset tersebut digunakan untuk melatih dan mengevaluasi model *CNN* pada proses deteksi emosi berbasis *Visual* secara realtime.



Gambar 4. 4 Hasil *Train Visual*



Gambar 4. 5 Implementasi Sistem *Visual* Realtime

Tabel 4. 4 Parameter Sistem *Visual*

Parameter	Nilai
Metode Klasifikasi	*CNN*
Deteksi Wajah	Haar Cascade
Input Citra	96 × 96 Piksel
Jenis Citra	Grayscale
Jumlah Emosi	4
Kategori Emosi	*Angry*, *Happy*, *Neutral*, *Sad*

Total Data *Train*ing	21005
Total Data Testing	5212
Epoch	50
Batch Size	32
Optimizer	Adam
Loss Function	Categorical Crossentropy
Framework Deep Learning	TensorFlow / Keras
Input Sistem	Webcam
Output Sistem	Emosi dan Confidence Score
Bahasa Pemrograman	Python

Berdasarkan Tabel 4.4, sistem *Visual* menggunakan metode *CNN* sebagai model utama untuk melakukan klasifikasi emosi berdasarkan ekspresi wajah pengguna. Proses deteksi wajah dilakukan menggunakan Haar Cascade Classifier sebelum citra diproses oleh model *CNN*.

Pada penelitian ini digunakan citra grayscale berukuran 96 × 96 piksel sebagai input model. Dataset *Visual* terdiri dari empat kategori emosi yaitu *Angry*, *Happy*, *Neutral*, dan *Sad* dengan total 21005 data *train*ing dan 5212 data testing.

Proses *train*ing model dilakukan menggunakan optimizer Adam dan loss function categorical crossentropy. Hasil klasifikasi kemudian ditampilkan pada GUI dalam bentuk label emosi dan confidence score secara realtime.

4.1.3 Implementasi Sistem *Fusion*

Implementasi sistem *Fusion* dilakukan dengan menggabungkan hasil prediksi dari sistem *Audio* dan sistem *Visual* untuk menghasilkan keputusan emosi akhir yang lebih stabil. Metode *Fusion* yang digunakan pada penelitian ini adalah late *Fusion* dengan pendekatan weighted average.

Pada metode late *Fusion*, sistem *Audio* dan *Visual* melakukan proses klasifikasi secara terpisah. Sistem *Audio* menghasilkan probabilitas emosi berdasarkan fitur MFCC dan *Random Forest*, sedangkan sistem *Visual* menghasilkan probabilitas emosi berdasarkan model *CNN*.

Hasil probabilitas dari kedua sistem kemudian digabungkan menggunakan pembobotan tertentu untuk menentukan emosi akhir pengguna. Pada penelitian ini digunakan pembobotan sebesar 75% untuk sistem *Visual* dan 25% untuk sistem *Audio*.

Pembobotan lebih besar diberikan pada sistem *Visual* karena model *CNN* memiliki performa yang lebih stabil dibandingkan sistem *Audio* realtime yang masih dipengaruhi oleh noise lingkungan dan kualitas microphone.

Persamaan *Fusion* yang digunakan pada penelitian ini yaitu:

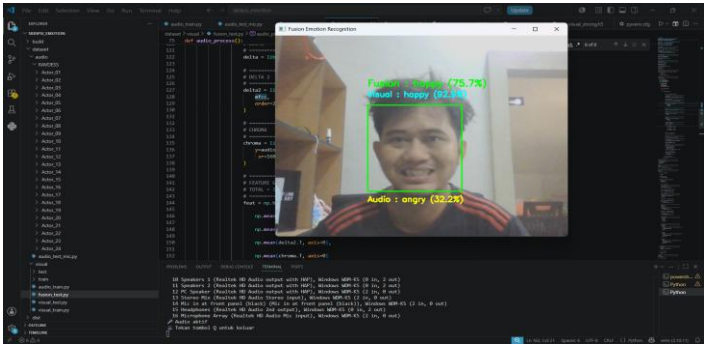
$$Fusion=(0.75\times Visual)+(0.25\times Audio)$$

Hasil probabilitas *Fusion* kemudian digunakan untuk menentukan kategori emosi akhir dengan memilih nilai probabilitas tertinggi menggunakan metode argmax.

Tahapan implementasi sistem *Fusion* terdiri dari:

1. Pengambilan input *Audio* dan *Visual*
2. Prediksi emosi *Audio* menggunakan *Random Forest*
3. Prediksi emosi *Visual* menggunakan *CNN*
4. Penggabungan probabilitas menggunakan weighted *Fusion*
5. Menentukan emosi akhir
6. Menampilkan hasil emosi pada GUI

Sistem *Fusion* berhasil dijalankan secara realtime menggunakan webcam dan microphone secara bersamaan. Hasil *Fusion* menunjukkan prediksi emosi yang lebih stabil dibandingkan penggunaan sistem *Audio* atau *Visual* secara terpisah.



Gambar 4. 6 Implementasi *Fusion* di Real Time

Tabel 4. 5 Parameter Sistem *Fusion*

Parameter	**Nilai**
Metode *Fusion*	Late *Fusion*
Metode Penggabungan	Weighted Average
Bobot *Visual*	75%
Bobot *Audio*	25%
Model *Audio*	MFCC + *Random Forest*
Model *Visual*	*CNN*
Input Sistem	Webcam dan Microphone
Output Sistem	Emosi dan Confidence Score
Jumlah Emosi	4
Kategori Emosi	*Angry*, *Happy*, *Neutral*, *Sad*

Berdasarkan Tabel 4.5, sistem *Fusion* menggunakan metode late *Fusion* dengan pendekatan weighted average untuk menggabungkan hasil prediksi *Audio* dan *Visual*. Pembobotan lebih besar diberikan pada sistem *Visual* karena memiliki tingkat stabilitas yang lebih baik dibandingkan sistem *Audio* realtime.

Penggunaan metode *Fusion* pada penelitian ini membantu meningkatkan kestabilan hasil prediksi emosi terutama pada kondisi tertentu seperti noise lingkungan atau pencahayaan yang kurang baik.

4.1.4 Implementasi GUI

Implementasi GUI sistem dilakukan menggunakan library Tkinter pada bahasa pemrograman Python. GUI dirancang sebagai antarmuka utama agar pengguna dapat menjalankan sistem deteksi emosi dengan lebih mudah dan interaktif.

GUI sistem mampu menampilkan hasil deteksi emosi secara realtime menggunakan webcam dan microphone. Selain itu, GUI juga digunakan untuk menampilkan hasil prediksi emosi dari sistem *Audio*, *Visual*, maupun *Fusion Audio-Visual*.

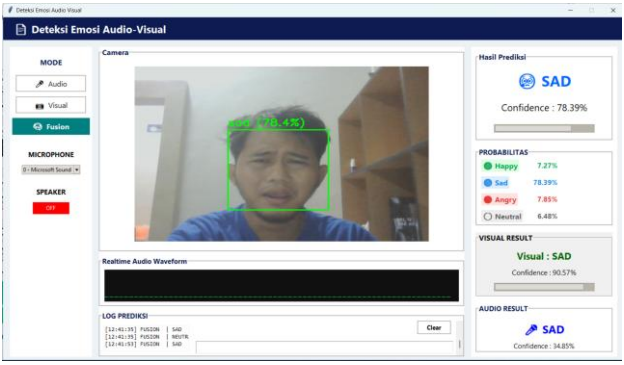
Pada implementasi GUI, sistem memiliki beberapa mode utama yaitu:

1. Mode *Visual*
2. Mode *Audio*
3. Mode *Fusion*

Mode *Visual* digunakan untuk mendeteksi emosi berdasarkan ekspresi wajah menggunakan webcam dan model *CNN*. Mode *Audio* digunakan untuk mendeteksi emosi berdasarkan suara menggunakan MFCC dan *Random Forest*. Sedangkan mode *Fusion* digunakan untuk menggabungkan hasil prediksi *Audio* dan *Visual* secara realtime.

GUI juga menampilkan confidence score untuk menunjukkan tingkat keyakinan model terhadap hasil prediksi emosi. Selain itu, sistem menampilkan emoticon sesuai kategori emosi yang terdeteksi sehingga tampilan GUI menjadi lebih interaktif.

Implementasi GUI berhasil dijalankan secara realtime dan mampu menampilkan hasil deteksi emosi secara langsung menggunakan webcam dan microphone secara bersamaan.



Gambar 4. 7 Implementasi GUI

4.2 Hasil Pengujian Sistem

Pada penelitian ini dilakukan pengujian terhadap sistem deteksi emosi berbasis *Audio*, *Visual*, dan *Fusion Audio-Visual* untuk mengetahui performa sistem dalam mengenali emosi manusia secara realtime.

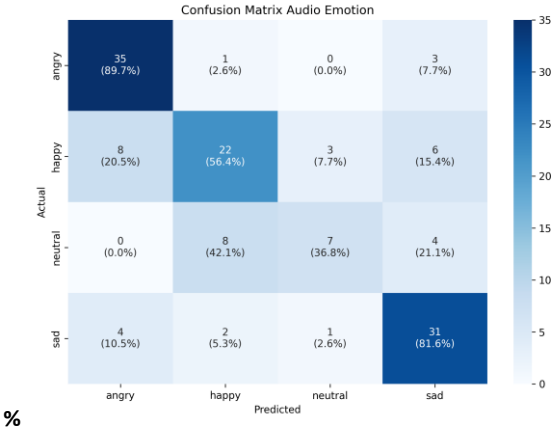
Pengujian dilakukan menggunakan dataset *Audio* dan *Visual* yang telah melalui proses *train*ing sebelumnya. Selain itu, pengujian juga dilakukan secara realtime menggunakan webcam dan microphone untuk mengetahui kemampuan sistem dalam mendeteksi emosi pengguna secara langsung.

Evaluasi sistem dilakukan menggunakan beberapa parameter seperti accuracy, precision, recall, F1-score, dan con*Fusion* matrix. Pengujian dilakukan pada tiga sistem utama yaitu sistem *Audio*, sistem *Visual*, dan sistem *Fusion Audio-Visual*.

4.2.1 Hasil Pengujian *Audio*

Pengujian sistem *Audio* dilakukan menggunakan metode ekstraksi fitur MFCC dan algoritma *Random Forest* sebagai classifier untuk mengenali emosi berdasarkan suara pengguna.

Proses pengujian dilakukan menggunakan dataset RAVDESS dengan empat kategori emosi yaitu *Angry*, *Happy*, *Neutral*, dan *Sad*. Pengujian dilakukan setelah model selesai melalui proses *train*ing menggunakan data *Audio* yang telah diekstraksi menggunakan MFCC. Berdasarkan hasil pengujian, model *Random Forest* mampu mengenali emosi suara dengan cukup baik dengan tingkat akurasi sebesar: **75.56**



%

Gambar 4. 8 Confussion Matrix *Audio*

Hasil tersebut menunjukkan bahwa metode MFCC dan *Random Forest* mampu digunakan untuk sistem pengenalan emosi berbasis *Audio*.

Tabel 4. 6 Hasil Evaluasi *Audio*

Emosi	Precision	Recall	F1-Score
Angry	0.85	0.90	0.88
Happy	0.77	0.51	0.62
Neutral	0.81	0.68	0.74
Sad	0.65	0.89	0.76

Berdasarkan hasil evaluasi pada Tabel 4.6, model *Audio* memiliki performa terbaik pada kategori *Angry* dengan nilai precision dan recall yang cukup tinggi. Hal tersebut menunjukkan bahwa karakteristik suara pada emosi *Angry* lebih mudah dikenali oleh model dibandingkan kategori emosi lainnya.

Pada kategori *Happy* masih ditemukan beberapa kesalahan klasifikasi akibat kemiripan pola suara dengan kategori emosi lain seperti *Sad* dan *Neutral*. Selain itu, faktor noise lingkungan dan kualitas microphone juga mempengaruhi performa sistem *Audio* realtime.

Tabel 4. 7 Pengujian Deteksi Emosi melalui mode *Visual*

NO		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10

1	*Happy*	*Angry* 44.78	*Sad* 34.45	*Happy* 32.10	*Sad* 32.44	*Sad* 33.49	*Sad* 37.35	*Sad* 42.43	*Sad* 38.54	*Sad* 42.81	*Sad* 33.78
2		*Angry* 38.57	*Sad* 37.09	*Sad* 38.66	*Sad* 36.30	*Happy* 34.01	*Sad* 41.52	*Sad* 33.28	*Sad* 48.81	*Sad* 37.26	*Sad* 43.54
3		*Sad* 38.54	*Sad* 42.58	*Sad* 38.84	*Sad* 34.58	*Sad* 43.53	*Sad* 33.82	*Sad* 37.86	*Sad* 45.72	*Sad* 39.30	*Sad* 47.81
4		*Sad* 48.81	*Sad* 42.81	*Sad* 36.84	*Sad* 32.81	*Happy* 40.06	*Angry* 39.33	*Sad* 34.80	*Sad* 38.27	*Sad* 41.58	*Sad* 35.86
5		*Sad* 45.72	*Sad* 35.03	*Happy* 35.93	*Sad* 35.03	*Sad* 31.78	*Sad* 36.57	*Sad* 32.58	*Sad* 35.35	*Sad* 39.81	*Sad* 37.30
6		*Happy* 33.72	*Angry* 34.17	*Happy* 33.27	*Angry* 38.17	*Sad* 32.54	*Angry* 34.29	*Sad* 41.81	*Sad* 38.53	*Sad* 48.58	*Sad* 33.58
7		*Angry* 33.83	*Sad* 39.87	*Sad* 39.17	*Sad* 37.87	*Sad* 39.81	*Sad* 35.44	*Sad* 40.58	*Sad* 42.99	*Happy* 38.65	*Sad* 41.81
8		*Angry* 38.18	*Sad* 45.84	*Sad* 39.30	*Sad* 36.84	*Sad* 38.81	*Sad* 32.30	*Angry* 37.18	*Sad* 31.22	*Sad* 32.81	*Sad* 41.58
9		*Angry* 33.69	*Angry* 37.40	*Sad* 40.52	*Angry* 34.40	*Sad* 32.72	*Angry* 34.58	*Sad* 37.43	*Sad* 32.58	*Sad* 35.03	*Sad* 37.64
10		*Sad* 33.78	*Angry* 36.20	*Sad* 38.02	*Angry* 39.20	*Happy* 37.72	*Sad* 32.81	*Sad* 43.28	*Sad* 41.81	*Sad* 32.58	*Sad* 34.81
11		*Sad* 38.54	*Sad* 39.14	*Sad* 35.33	*Sad* 38.14	*Sad* 34.53	*Sad* 42.58	*Sad* 39.86	*Happy* 38.86	*Sad* 41.81	*Sad* 34.87
12		*Sad* 48.81	*Sad* 47.19	*Sad* 32.52	*Sad* 37.19	*Sad* 37.25	*Sad* 37.81	*Sad* 35.30	*Sad* 41.81	*Happy* 38.86	*Sad* 31.41
13		*Happy* 33.72	*Sad* 43.53	*Sad* 31.28	*Sad* 33.53	*Sad* 35.20	*Sad* 35.03	*Sad* 33.58	*Happy* 38.86	*Happy* 37.25	*Sad* 35.04
14		*Angry* 33.69	*Sad* 39.25	*Sad* 39.42	*Sad* 39.25	*Sad* 31.53	*Angry* 38.17	*Sad* 41.81	*Happy* 37.25	*Happy* 35.66	*Sad* 33.53
15		*Sad* 38.54	*Sad* 41.20	*Sad* 37.04	*Sad* 31.20	*Sad* 33.25	*Sad* 31.22	*Sad* 42.58	*Sad* 34.87	*Happy* 37.25	*Sad* 36.17
16	*Neutral*	*Sad* 37.5	*Sad* 37.35	*Sad* 37.45	*Sad* 34.5	*Sad* 32.54	*Sad* 37.43	*Angry* 38.18	*Sad* 40.81	*Sad* 39.86	*Sad* 39.30
17		*Sad* 42.99	*Sad* 34.81	*Sad* 38.53	*Sad* 32.99	*Sad* 34.23	*Sad* 43.8	*Sad* 37.43	*Sad* 42.86	*Sad* 38.36	*Sad* 39.42
18		*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*

		41.81	27.87	42.99	31.81	35.68	39.86	43.8	47.30	36.00	33.78
19		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*
		35.86	37.41	31.22	35.86	39.27	35.30	39.86	43.58	37.69	37.54
20		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		37.04	43.04	36.43	34.04	36.17	33.58	35.30	45.58	41.78	39.81
21		*Sad*	*Happy*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		37.45	40.56	42.8	33.45	38.30	41.81	33.58	39.42	38.54	45.86
22		*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		38.5	33.83	36.86	34.81	41.32	42.58	41.88	35.86	39.27	44.30
23		*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*
		42.99	38.18	44.91	37.87	40.02	38.18	43.98	34.04	36.17	41.58
24		*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Angry*	*Angry*	*Sad*	*Sad*	*Sad*
		41.81	33.69	40.59	27.41	37.53	33.69	38.18	33.45	38.30	43.78
25		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		35.86	33.78	39.44	33.04	36.17	34.58	35.43	34.81	41.32	43.54
26		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		37.04	38.54	42.24	38.04	39.30	37.64	43.8	37.87	40.02	40.81
27		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		34.45	38.81	36.56	35.45	39.42	37.45	39.86	27.41	37.53	42.86
28		*Sad*	*Happy*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		37.5	38.72	37.51	36.99	32.65	38.45	35.30	35.86	39.27	47.30
29		*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		35.99	32.69	36.86	32.81	37.53	34.35	33.58	34.04	36.17	43.58
30		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		39.81	37.54	38.91	37.99	37.21	35.81	41.81	33.45	38.30	45.58
31		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		31.78	40.56	30.59	34.41	39.30	31.87	42.58	34.81	41.32	39.42
32		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*
		31.5	40.41	29.44	40.66	40.52	32.41	38.18	37.87	40.02	33.78
33	*Angry*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*
		34.8	38.18	43.24	30.71	36.33	30.59	37.43	27.41	37.53	37.54
34		*Angry*	*Angry*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		35.8	33.69	37.56	37.28	37.25	33.38	43.8	32.58	39.25	43.88
35		*Happy*	*Sad*	*Sad*	*Angry*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		31.2	33.78	37.04	32.66	38.18	38.24	39.86	41.81	37.28	45.13

36		*Angry*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		48.6	38.54	34.45	32.77	33.69	38.51	35.30	40.58	34.41	44.45
37		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*
		30.2	38.81	37.5	31.52	33.78	36.76	33.58	37.18	32.61	41.57
38		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Happy*	*Sad*
		33.4	31.78	31.78	33.61	38.54	38.91	41.81	35.61	41.82	43.79
39		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Happy*	*Sad*	*Sad*
		39.4	31.5	34.5	31.78	38.81	30.59	42.58	30.72	36.92	43.44
40		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*
		35.1	35.3	25.5	34.53	31.78	31.53	38.18	38.92	38.41	40.71
41		*Angry*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		32.5	29.9	33.3	35.31	31.51	34.84	37.43	38.41	37.17	42.86
42		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*
		37.2	36.0	27.9	33.61	31.78	35.88	43.8	37.17	41.24	47.30
43		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*
		38.3	35.5	34.30	36.78	34.53	32.54	39.86	33.61	38.56	37.28
44		*Happy*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		35.5	32.4	29.59	31.53	25.52	37.27	35.30	36.78	39.71	34.41
45		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		29.5	29.7	37.0	34.31	33.37	38.37	33.58	31.53	40.86	43.88
46	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*
		35.3	36.1	33.5	33.11	27.92	32.59	41.81	39.81	37.28	45.13
47		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		29.9	37.1	30.3	39.14	34.30	37.21	42.58	45.86	35.45	44.45
48		*Sad*	*Angry*	*Sad*	*Angry*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*
		36.0	33.69	35.1	32.39	29.59	38.36	38.18	44.30	37.61	41.57
49		*Happy*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Happy*	*Sad*
		52.3	33.78	34.5	33.48	36.0	35.17	37.43	41.58	41.61	43.79
50		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*
		33.1	38.54	33.4	34.44	35.5	32.56	43.8	43.78	34.77	43.44
51		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		30.3	37.04	27.7	34.09	32.4	37.28	39.86	39.81	31.52	5 3.88
		Sad	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		31.1	34.45	34.1	35.45	29.7	34.51	35.30	45.86	35.61	43.13
53		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*

		35.5	37.5	36.1	36.15	31.12	35.61	33.58	44.30	31.78	52.45
54		*Sad*	*Sad*	*Happy*	*Sad*	*Sad*	*Happy*	*Sad*	*Sad*	*Sad*	*Sad*
		32.4	35.99	29.7	32.78	37.04	30.72	41.81	41.58	34.53	49.57
55		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		29.7	39.81	37.9	35.72	34.45	38.92	42.58	43.78	39.31	50.79
56		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*
		36.1	37.2	35.1	36.13	37.5	38.41	38.18	43.54	33.61	43.44
57		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		31.1	38.3	29.9	37.23	31.78	37.17	37.43	40.81	36.78	40.71
58		*Happy*	*Happy*	*Sad*	*Happy*	*Sad*	*Angry*	*Sad*	*Sad*	*Sad*	*Sad*
		29.7	35.5	36.0	33.15	34.45	38.15	43.8	42.86	31.53	49.86
59		*Sad*	*Sad*	*Happy*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		32.9	29.5	52.3	24.67	37.5	34.23	39.86	47.30	37.5	53.30
60		*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*	*Sad*
		34.1	35.2	36.1	33.42	37.5	36.35	35.30	44.30	32.99	55.28

Tingngkat keberhasilan sistem pada Mode *Audio* digunakan untuk mengetahui kemampuan sistem dalam mengenali emosi berdasarkan sinyal suara yang direkam melalui mikrofon. Pengujian dilakukan terhadap 10 responden, dimana setiap responden melakukan 60 kali pengujian yang terdiri atas 15 data emosi *Happy*, 15 data emosi *Neutral*, 15 data emosi *Angry*, dan 15 data emosi *Sad*. Dengan demikian, jumlah data pengujian untuk setiap kategori emosi adalah 150 data, sedangkan jumlah keseluruhan data pengujian adalah 600 data.

Tingkat keberhasilan dihitung berdasarkan jumlah prediksi yang sesuai dengan label sebenarnya dibandingkan dengan jumlah seluruh data yang diuji. Persamaan yang digunakan untuk menghitung tingkat keberhasilan ditunjukkan pada Persamaan berikut.

$$\text{Tingkat Keberhasilan} = \frac{\text{Jumlah Prediksi Benar}}{\text{Jumlah Data Pengujian}} \times 100\%$$

Sebagai contoh, pada pengujian Mode *Audio* untuk emosi *Happy*, sistem berhasil mengklasifikasikan 16 data dengan benar dari total 150 data pengujian. Oleh karena itu, tingkat keberhasilan sistem untuk emosi *Happy* dihitung sebagai berikut.

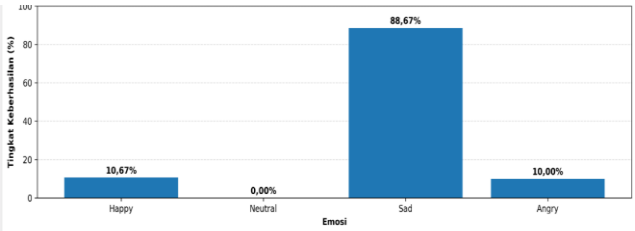
$$\text{Tingkat Keberhasilan Happy} = \frac{16}{150} \times 100\% = 10{,}67\%$$

Perhitungan yang sama dilakukan pada kategori emosi *Neutral*, *Angry*, dan *Sad* sehingga diperoleh tingkat keberhasilan masing-masing kategori emosi. Hasil perhitungan tersebut kemudian digunakan untuk mengevaluasi kemampuan sistem dalam mengenali emosi berdasarkan informasi suara pada kondisi pengujian secara real-time.

Tabel 4. 8 Tingkat Keberhasilan Mode *Audio* tiap Emosi

Emosi	Total Data	Benar	Salah	Tingkat Keberhasilan
Happy	150	16	134	10,67%
Neutral	150	0	150	0,00%
Angry	150	15	135	10,00%
Sad	150	133	17	88,67%

Gambar 4. 9 Grafik Tingkat Keberhasilan *Audio*

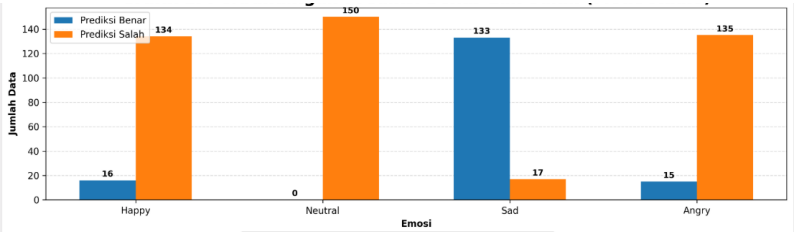


Berdasarkan Gambar 4.9, tingkat keberhasilan sistem pada Mode *Audio* menunjukkan perbedaan yang cukup signifikan untuk setiap kategori emosi. Emosi *Sad* memperoleh tingkat keberhasilan tertinggi yaitu 88,67%, sedangkan emosi *Happy* dan *Angry* hanya memperoleh tingkat keberhasilan masing-masing sebesar 10,67% dan 10,00%. Sementara itu, emosi *Neutral* tidak berhasil dikenali oleh sistem sehingga memperoleh tingkat keberhasilan 0,00%.

Tingginya tingkat keberhasilan pada emosi *Sad* menunjukkan bahwa karakteristik suara sedih, seperti intonasi yang lebih rendah, tempo bicara yang lambat, dan energi suara yang relatif stabil, lebih mudah dipelajari oleh model *Random Forest* menggunakan fitur MFCC. Sebaliknya, karakteristik suara pada emosi *Happy*, *Neutral*, dan *Angry* memiliki kemiripan pola akustik sehingga model masih mengalami kesalahan dalam proses klasifikasi.

Selain karakteristik suara, performa sistem juga dipengaruhi oleh beberapa faktor lain, seperti kualitas mikrofon, kondisi lingkungan yang mengandung noise, jarak responden terhadap mikrofon, serta perbedaan intonasi dan cara berbicara setiap responden. Faktor-faktor tersebut menyebabkan fitur MFCC yang diekstraksi dari sinyal suara menjadi kurang representatif sehingga menurunkan tingkat keberhasilan sistem pada beberapa kategori emosi.

Gambar 4. 10 Grafik Perbandingan Benar dan Salah Hasil Prediksi Mode *Audio*



Tabel 4. 9 Tingkat Keberhasilan Tiap Responden Mode *Audio*

Responden	***Happy***	***Neutral***	***Angry***	***Sad***
S1	13,33	0,00	20,00	80,00
S2	0,00	0,00	13,33	86,67
S3	20,00	0,00	0,00	80,00
S4	0,00	0,00	13,33	86,67
S5	20,00	0,00	13,33	100,00
S6	0,00	0,00	20,00	73,33
S7	0,00	0,00	13,33	86,67
S8	20,00	0,00	6,67	100,00
S9	33,33	0,00	0,00	93,33
S10	0,00	0,00	0,00	100,00

Berdasarkan Gambar 4.10, jumlah prediksi benar dan prediksi salah menunjukkan bahwa model *Audio* memiliki performa yang belum merata pada setiap kategori emosi. Emosi *Sad* memiliki jumlah prediksi benar sebanyak 133 data dari 150 data pengujian, sedangkan jumlah prediksi salah hanya 17 data. Sebaliknya, emosi *Happy* hanya berhasil diklasifikasikan dengan benar sebanyak 16 data, emosi *Angry* sebanyak 15 data, dan emosi *Neutral* tidak memiliki prediksi benar sama sekali.

Hasil tersebut menunjukkan bahwa model cenderung memberikan prediksi ke kelas *Sad* dibandingkan kelas emosi lainnya. Kondisi ini mengindikasikan bahwa model masih mengalami bias klasifikasi (classification bias), sehingga kemampuan sistem dalam membedakan karakteristik suara antar emosi belum optimal. Oleh karena itu, penggunaan modalitas *Audio* saja masih memiliki keterbatasan dalam mendeteksi emosi secara akurat.

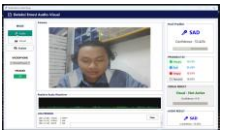
Secara keseluruhan, hasil pengujian menunjukkan bahwa sistem deteksi emosi berbasis *Audio* masih memiliki performa yang lebih rendah dibandingkan sistem *Visual*. Hal ini disebabkan karena sinyal suara lebih mudah dipengaruhi oleh faktor eksternal, seperti kebisingan lingkungan, kualitas perangkat perekam, serta variasi karakteristik suara setiap individu. Oleh karena itu, pada penelitian ini digunakan metode *Audio-Visual Fusion*, yaitu menggabungkan hasil prediksi *Audio* dan *Visual*, sehingga keputusan akhir sistem menjadi lebih stabil dan mampu meningkatkan keandalan deteksi emosi secara real-time.

Tabel 4. 10 Hasil Pengujian Deteksi Emosi melalui mode Visual

NO	Responden 1	Responden 2	Responden 3	Responden 4
1				
2				
3				
4				
5				
6				
7				

8				
NO	Responden 5	Responden 6	Responden 7	Responden 8
1				
2				
3				
4				
5				

6				
7				
8				
No	Responden 9	Responden 10		
1				
2				
3				
4				

5		
6		
7		
8		

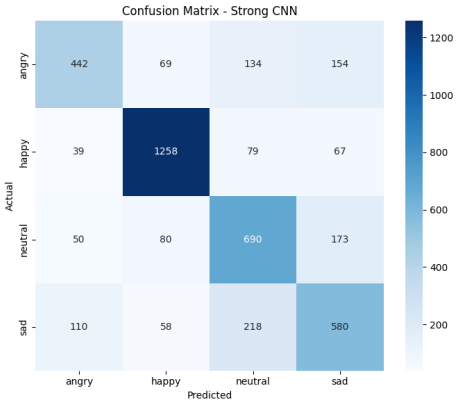
4.2.2 Hasil Pengujian Visual

Pengujian sistem *Visual* dilakukan menggunakan model *Convolutional Neural Network (CNN)* untuk mendeteksi emosi berdasarkan ekspresi wajah pengguna secara realtime. Sistem *Visual* menggunakan webcam sebagai perangkat input utama untuk menangkap citra wajah pengguna.

Pada proses pengujian, sistem terlebih dahulu melakukan deteksi wajah menggunakan Haar Cascade Classifier sebelum citra diproses oleh model *CNN*. Setelah wajah berhasil dideteksi, citra dilakukan preprocessing berupa grayscale, resize, dan normalisasi sebelum masuk ke tahap klasifikasi emosi.

Pengujian dilakukan menggunakan dataset *Visual* yang telah dibagi menjadi data *train*ing dan data testing dengan empat kategori emosi yaitu *Angry*, *Happy*, *Neutral*, dan *Sad*.

Berdasarkan hasil pengujian, model *CNN* mampu mengenali ekspresi wajah pengguna dengan cukup baik dan menghasilkan performa yang lebih stabil dibandingkan sistem *Audio* realtime.



Gambar 4. 11 Confussion Matrix *Visual*

Tabel 4. 11 Hasil Evaluasi *Visual*

Emosi	**Precision**	**Recall**	**F1-Score**
Angry	0.85	0.90	0.87
Happy	0.79	0.86	0.82
Neutral	0.89	0.92	0.90
Sad	0.90	0.80	0.85

Berdasarkan hasil pengujian, model *CNN* berhasil mencapai tingkat akurasi sebesar 87.48% pada proses klasifikasi emosi wajah. Hasil tersebut menunjukkan bahwa model *CNN* mampu mengenali ekspresi wajah pengguna dengan cukup baik.

Kategori *Neutral* memiliki performa terbaik dengan nilai precision dan recall yang tinggi. Hal tersebut menunjukkan bahwa model mampu mengenali pola ekspresi wajah *Neutral* dengan lebih stabil dibandingkan kategori emosi lainnya.

Tabel 4. 12 Pengujian untuk *Visual* dengan Responden

NO		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10
1		*Happy* 87.2	*Happy* 90.0	*Happy* 78.93	*Happy* 79.25	*Happy* 65.25	*Happy* 72.30	*Happy* 98.93	*Happy* 94.43	*Happy* 95.24	*Happy* 98.89
2		*Happy* 99.9	*Happy* 92.0	*Happy* 81.03	*Happy* 76.52	*Happy* 78.89	*Happy* 75.73	*Happy* 89.93	*Happy* 98.65	*Happy* 94.62	*Happy* 99.02
3		*Happy* 99.2	*Happy* 95.0	*Happy* 80.79	*Happy* 89.02	*Happy* 80.52	*Happy* 60.23	*Happy* 90.27	*Happy* 96.27	*Happy* 95.67	*Happy* 98.67
4		*Happy* 85.2	*Happy* 96.3	*Happy* 99.91	*Happy* 80.17	*Happy* 58.13	*Happy* 70.59	*Happy* 85.32	*Happy* 90.32	*Happy* 97.25	*Happy* 97.29
5		*Happy* 99.7	*Happy* 95.4	*Happy* 98.13	*Happy* 80.64	*Happy* 60.46	*Happy* 60.91	*Happy* 92.92	*Happy* 94.72	*Happy* 96.74	*Happy* 96.74
6		*Happy* 81.4	*Happy* 93.1	*Happy* 97.83	*Happy* 80.80	*Happy* 67.17	*Happy* 78.23	*Neutral* 93.57	*Happy* 93.67	*Happy* 94.58	*Happy* 74.48
7		*Neutral* 60.82	*Happy* 80.21	*Happy* 90.72	*Happy* 76.55	*Happy* 72.70	*Happy* 67.83	*Happy* 95.11	*Happy* 95.21	*Happy* 95.93	*Happy* 85.31
8	*Happy*	*Neutral* 70.5	*Happy* 77.5	*Happy* 81.22	*Happy* 74.21	*Happy* 74.21	*Happy* 70.72	*Happy* 86.78	*Happy* 89.78	*Happy* 98.42	*Happy* 88.78
9		*Happy* 69.2	*Happy* 71.4	*Happy* 86.7	*Happy* 73.13	*Happy* 78.13	*Happy* 72.80	*Happy* 91.72	*Happy* 91.52	*Happy* 95.77	*Happy* 99.77
10		*Happy* 62.6	*Happy* 85.4	*Happy* 75.39	*Happy* 78.23	*Happy* 80.37	*Happy* 76.55	*Happy* 78.33	*Happy* 97.33	*Happy* 96.42	*Happy* 99 .42
11		*Happy* 87.2	*Neutral* 51.4	*Neutral* 95.8	*Neutral* 75.51	*Neutral* 90.66	*Happy* 73.21	*Happy* 88.82	*Happy* 95.82	*Happy* 94.21	*Happy* 99.66
12		*Happy* 62.5	*Happy* 73.4	*Happy* 83.7	*Happy* 70.11	*Happy* 8.22	*Happy* 74.13	*Happy* 94.59	*Happy* 93.59	*Happy* 93.52	*Happy* 98.37
13		*Happy* 82.5	*Neutral* 83.5	*Neutral* 82.8	*Happy* 96.17	*Happy* 86.37	*Happy* 66.23	*Happy* 81.45	*Happy* 94.33	*Happy* 97.31	*Happy* 83.34
14		*Happy* 63.8	*Neutral* 86.2	*Neutral* 93.5	*Happy* 8.64	*Happy* 88.08	*Happy* 63.80	*Happy* 84.47	*Happy* 90.47	*Happy* 94.11	*Happy* 94.11
15		*Happy* 82.5	*Neutral* 52.4	*Neutral* 99.7	*Neutral* 90.67	*Neutral* 78.63	*Happy* 76.55	*Happy* 81.98	*Happy* 96.98	*Happy* 95.85	*Happy* 97.53
16	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*

		86.9	81.0	46.92	69.74	96.29	49.92	86.89	85.49	81.48	96.48
17		*Neutral* 83.4	*Neutral* 75.4	*Neutral* 44.00	*Neutral* 82.19	*Neutral* 94.57	*Neutral* 51.06	*Neutral* 83.46	*Neutral* 83.31	*Happy* 80.24	*Neutral* 95.24
18		*Neutral* 82.6	*Neutral* 71.0	*Happy* 52.46	*Neutral* 79.96	*Neutral* 95.37	*Happy* 47.46	*Neutral* 81.65	*Neutral* 80.33	*Neutral* 88.89	*Neutral* 94.89
19		*Neutral* 77.7	*Neutral* 59.5	*Happy* 52.91	*Neutral* 85.36	*Neutral* 93.27	*Happy* 45.71	*Neutral* 90.72	*Neutral* 81.97	*Happy* 85.77	*Happy* 85.77
20		*Neutral* 81.4	*Neutral* 73.5	*Neutral* 65.83	*Neutral* 81.40	*Neutral* 92.45	*Neutral* 46.32	*Neutral* 81.48	*Neutral* 81.38	*Neutral* 86.57	*Neutral* 96.54
21		*Happy* 85.2	*Neutral* 70.1	*Neutral* 75.78	*Happy* 85.54	*Neutral* 96.38	*Neutral* 47.92	*Happy* 80.24	*Happy* 78.41	*Neutral* 92.63	*Neutral* 98.32
22		*Neutral* 81.4	*Neutral* 60.0	*Neutral* 80.33	*Neutral* 81.4	*Neutral* 95.83	*Neutral* 44.00	*Neutral* 84.89	*Neutral* 70.39	*Neutral* 85.28	*Neutral* 93.25
23		*Happy* 85.2	*Happy* 45.1	*Neutral* 54.43	*Happy* 85.2	*Neutral* 94.57	*Happy* 52.46	*Happy* 85.77	*Happy* 75.44	*Neutral* 86.15	*Neutral* 96.75
24		*Neutral* 72.5	*Neutral* 92.5	*Neutral* 76.61	*Neutral* 84.97	*Neutral* 96.83	*Happy* 40.91	*Neutral* 96.57	*Neutral* 86.77	*Neutral* 87.46	*Neutral* 92.51
25		*Neutral* 70.6	*Neutral* 80.8	*Neutral* 60.66	*Neutral* 74.75	*Neutral* 96.75	*Neutral* 46.92	*Neutral* 91.66	*Neutral* 81.53	*Neutral* 85.45	*Neutral* 95.45
26		*Neutral* 75.4	*Neutral* 81.2	*Neutral* 74.5	*Neutral* 85.97	*Neutral* 95.75	*Neutral* 46.23	*Neutral* 85.78	*Neutral* 75.64	*Neutral* 88.72	*Neutral* 95.24
27		*Neutral* 63.5	*Neutral* 79.3	*Neutral* 70.18	*Neutral* 84.97	*Neutral* 93.28	*Happy* 51.46	*Neutral* 83.75	*Neutral* 80.72	*Neutral* 93.88	*Neutral* 94.89
28		*Neutral* 82.5	*Neutral* 95.7	*Neutral* 60.01	*Neutral* 77.65	*Neutral* 94.73	*Happy* 45.91	*Neutral* 92.55	*Neutral* 72.67	*Neutral* 89.95	*Happy* 85.77
29		*Neutral* 77.3	*Neutral* 93.9	*Neutral* 73.53	*Neutral* 87.46	*Neutral* 93.92	*Neutral* 42.92	*Neutral* 87.65	*Neutral* 84.95	*Neutral* 90.55	*Neutral* 97.44
30		*Neutral* 81.4	*Neutral* 91.9	*Neutral* 70.15	*Neutral* 91.7	*Neutral* 94.29	*Neutral* 56.00	*Neutral* 89.36	*Neutral* 76.45	*Neutral* 87.26	*Neutral* 96.37
31		*Angry* 62.5	*Angry* 50.6	*Sad* 40.70	*Happy* 52.29	*Neutral* 90.27	*Sad* 55.10	*Angry* 75.28	*Neutral* 56.67	*Neutral* 91.40	*Neutral* 95.24
32	*Angry*	*Happy* 63.8	*Sad* 37.2	*Sad* 44.94	*Happy* 60.64	*Neutral* 91.45	*Sad* 45.94	*Angry* 81.14	*Neutral* 48.54	*Neutral* 87.46	*Neutral* 94.89
33		*Angry* 62.5	*Neutral* 38.3	*Sad* 60.75	*Happy* 73.80	*Neutral* 93.38	*Sad* 55.75	*Angry* 62.57	*Neutral* 53.21	*Neutral* 85.45	*Happy* 85.77

34		*Angry*	*Angry*	*Neutral*	*Happy*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		67.4	42.5	48.77	76.55	92.83	45.70	668.6	46.22	88.72	83.14
35		*Angry*	*Angry*	*Happy*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		62.5	36.54	48.67	56.33	91.29	47.35	75.12	39.96	93.88	85.51
36		*Angry*	*Angry*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		48.6	42.7	56.72	48.8	80.57	43.94	83.95	49.70	89.22	80.35
37		*Angry*	*Angry*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		50.2	55.7	33.14	53.21	89.37	50.75	79.77	36.23	75.41	88.28
38		*Angry*	*Angry*	*Sad*	*Neutral*	*Neutral*	*Happy*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		53.4	52.4	39.88	46.22	85.27	46.67	68.56	42.53	80.72	87.42
39		*Angry*	*Sad*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		59.4	32.5	38.65	45.63	82.45	56.72	67.72	37.22	85.64	80.69
40		*Angry*	*Angry*	*Sad*	*Neutral*	*Neutral*	*Neutral*	*Angry*	*Neutral*	*Neutral*	*Happy*
		55.1	44.1	60.59	48.70	86.38	45.63	75.15	45.93	83.54	85.46
41		*Angry*	*Neutral*	*Sad*	*Neutral*	*Neutral*	*Neutral*	*Angry*	*Happy*	*Neutral*	*Neutral*
		62.5	47.2	65.99	47.70	76.73	43.72	73.55	68.79	82.83	86.59
42		*Angry*	*Neutral*	*Sad*	*Neutral*	*Neutral*	*Neutral*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		65.2	50.0	49.13	43.98	83.92	45.63	80.72	40.21	88.29	88.65
43		*Happy*	*Angry*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		93.5	42.2	60.17	45.42	86.28	48.70	83.48	42.22	80.37	82.25
44		*Angry*	*Angry*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		62.5	42.6	54.02	47.66	85.63	55.99	75.66	47.43	89.43	86.75
45		*Happy*	*Angry*	*Happy*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Happy*	*Neutral*	*Neutral*
		99.5	53.2	56.9	44.23	94.57	39.13	69.74	98.79	81.27	89..25
46	*Sad*	*Sad*	*Angry*	*Sad*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Sad*	*Sad*	*Neutral*
		75.3	53.2	54.38	54.36	66.83	40.17	73.15	41.30	47.25	80.91
47		*Sad*	*Neutral*	*Sad*	*Sad*	*Happy*	*Sad*	*Sad*	*Sad*	*Sad*	*Neutral*
		49.9	74.0	50.91	47.90	68.79	54.02	75.37	34.28	48.91	82.33
48		*Sad*	*Angry*	*Sad*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Sad*	*Sad*	*Neutral*
		56.0	40.3	57.63	56.0	67.33	40.3	65.95	34.36	43.63	86.59
49		*Neutral*	*Angry*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Sad*	*Sad*	*Sad*	*Neutral*
		51.3	42.2	47.35	51.3	72.8	47.15	70.89	37.90	46.35	79.55
50		*Neutral*	*Neutral*	*Sad*	*Neutral*	*Neutral*	*Sad*	*Sad*	*Sad*	*Sad*	*Neutral*
		53.42	36.7	56.89	53.42	71.21	45.32	74.23	42.03	43.79	72.25

51		*Neutral* 60.3	*Neutral* 35.5	*Sad* 43.78	*Neutral* 40.3	*Neutral* 73.22	*Sad* 56.13	*Sad* 76.11	*Sad* 39.30	*Sad* 45.19	*Neutral* 76.75
52		*Sad* 71.1	*Neutral* 40.3	*Sad* 68.99	*Sad* 47.15	*Neutral* 63.21	*Sad* 47.53	*Sad* 68.78	*Sad* 36.39	*Sad* 49.64	*Neutral* 80..25
53		*Sad* 65.5	*Angry* 40.7	*Sad* 45.58	*Sad* 45.32	*Neutral* 56.22	*Neutral* 62.83	*Sad* 67.58	*Sad* 36.28	*Sad* 44.59	*Neutral* 82.51
54		*Sad* 64.4	*Angry* 44.0	*Sad* 56.9	*Sad* 41.13	*Neutral* 65.63	*Happy* 66.79	*Sad* 65.45	*Sad* 37.28	*Sad* 41.33	*Neutral* 74.35
55		*Sad* 60.5	*Angry* 50.2	*Sad* 60.5	*Sad* 40.53	*Neutral* 68.70	*Neutral* 61.33	*Sad* 69.89	*Sad* 42.91	*Sad* 50.28	*Neutral* 75.68
56		*Sad* 55.1	*Angry* 49.2	*Sad* 55.1	*Sad* 49.11	*Neutral* 57.70	*Neutral* 51.23	*Neutral* 60.71	*Sad* 37.63	*Sad* 44.38	*Neutral* 87.42
57		*Sad* 49.1	*Sad* 36.8	*Sad* 46.81	*Sad* 44.19	*Neutral* 53.98	*Neutral* 50.35	*Sad* 75.55	*Sad* 34.35	*Sad* 49.66	*Neutral* 80.69
58		*Neutral* 60.1	*Angry* 43.22	*Happy* 45.99	*Neutral* 47.50	*Neutral* 65.42	*Neutral* 48.49	*Sad* 78.63	*Sad* 38.79	*Sad* 46.20	*Neutral* 75.28
59		*Sad* 54.0	*Angry* 46.3	*Neutral* 53.64	*Sad* 51.30	*Neutral* 73.98	*Neutral* 49.53	*Sad* 69.79	*Sad* 40.35	*Sad* 40.68	*Neutral* 77.42
60		*Sad* 71.1	*Angry* 48.48	*Sad* 45.36	*Sad* 44.28	*Neutral* 65.42	*Neutral* 58.32	*Sad* 76.12	*Sad* 39.79	*Sad* 52.36	*Neutral* 87.69

Tingkat keberhasilan sistem digunakan untuk mengetahui kemampuan sistem dalam mengenali emosi sesuai dengan kondisi sebenarnya (ground truth). Pengujian dilakukan terhadap 10 responden, dimana setiap responden melakukan 60 kali pengujian yang terdiri dari 15 data emosi *Happy*, 15 data emosi *Neutral*, 15 data emosi *Angry*, dan 15 data emosi *Sad*. Dengan demikian, jumlah data pengujian untuk setiap kategori emosi adalah 150 data, sedangkan total keseluruhan data pengujian adalah 600 data.

Tingkat keberhasilan dihitung berdasarkan jumlah prediksi yang sesuai dengan label sebenarnya dibandingkan dengan jumlah seluruh data yang diuji. Persamaan yang digunakan untuk menghitung tingkat keberhasilan ditunjukkan pada Persamaan berikut.

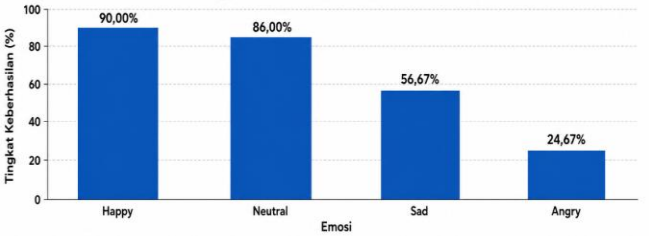
$$\text{Tingkat Keberhasilan} = \frac{\text{Jumlah Prediksi Benar}}{\text{Jumlah Data Pengujian}} \times 100\%$$

Sebagai contoh, pada pengujian emosi *Happy* diperoleh sebanyak 135 data berhasil dikenali dengan benar dari total 150 data pengujian. Maka tingkat keberhasilan sistem pada emosi *Happy* dapat dihitung sebagai berikut.

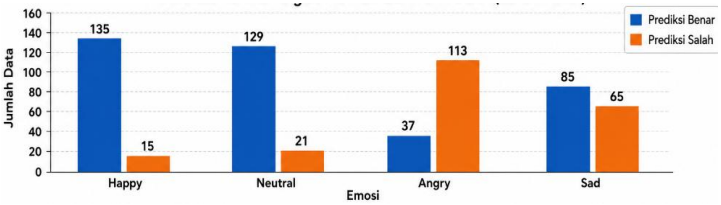
$$\text{Tingkat Keberhasilan Happy} = \frac{135}{150} \times 100\% = 90\%$$

Perhitungan yang sama dilakukan pada kategori emosi *Neutral*, *Angry*, dan *Sad* sehingga diperoleh tingkat keberhasilan masing-masing kategori emosi. Hasil perhitungan tersebut kemudian digunakan untuk mengevaluasi performa sistem dalam mendeteksi emosi pada kondisi pengujian secara real-time.

Gambar 4. 12 Grafik Tingkat Keberhasilan Mode *Visual*



Gambar 4. 13 Grafik Perbandingan Benar dan Salah Hasil Prediksi Mode *Visual*



Tabel 4. 13 Tingkat Keberhasilan Mode *Visual* tiap Emosi

Emosi	Total Data	Benar	Salah	Tingkat Keberhasilan
Happy	150	135	15	90.00%
Neutral	150	129	21	86.00%
Angry	150	37	113	24.67%
Sad	150	85	65	56.67%

Berdasarkan hasil perhitungan tingkat keberhasilan, sistem memiliki performa terbaik dalam mengenali emosi *Happy* dengan tingkat keberhasilan sebesar 90,00%, diikuti oleh emosi *Neutral* sebesar 86,00%. Hal ini menunjukkan bahwa karakteristik *Visual* pada kedua emosi tersebut lebih mudah dikenali oleh

model *CNN*. Sementara itu, emosi *Sad* memperoleh tingkat keberhasilan sebesar 56,67%, sedangkan emosi *Angry* memiliki tingkat keberhasilan paling rendah yaitu 24,67%. Rendahnya tingkat keberhasilan pada emosi *Angry* menunjukkan bahwa ekspresi marah yang ditampilkan oleh responden memiliki variasi yang cukup besar dan masih memiliki kemiripan dengan ekspresi emosi lainnya, sehingga sistem masih mengalami kesalahan klasifikasi pada kondisi pengujian secara real-time.

Tabel 4. 14 Tingkat Keberhasilan Tiap Responden Mode *Visual*

Responden	*Happy*	*Neutral*	*Angry*	*Sad*
S1	86.67	86.67	80.00	73.33
S2	73.33	93.33	66.67	6.67
S3	73.33	86.67	0.00	86.67
S4	86.67	86.67	0.00	73.33
S5	86.67	100.00	0.00	0.00
S6	100.00	60.00	0.00	33.33
S7	93.33	86.67	100.00	93.33
S8	100.00	86.67	0.00	100.00
S9	100.00	86.67	0.00	100.00
S10	100.00	86.67	0.00	0.00

Berdasarkan Tabel 4.14, tingkat keberhasilan deteksi emosi pada setiap responden menunjukkan variasi performa untuk masing-masing kategori emosi. Secara umum, emosi *Happy* dan *Neutral* memiliki tingkat keberhasilan yang lebih tinggi dibandingkan emosi *Angry* dan *Sad*. Perbedaan tersebut menunjukkan bahwa karakteristik *Visual* pada emosi *Happy* dan *Neutral* lebih mudah dikenali oleh model *CNN*, sedangkan emosi *Angry* dan *Sad* memiliki kemiripan ekspresi dengan emosi lain sehingga masih terjadi kesalahan klasifikasi pada beberapa responden.

Tabel 4. 15 Hasil Pengujian Deteksi Emosi melalui mode Visual

NO	Responden 1	Responden 2	Responden 3	Responden 4
1				
2				
3				
4				
5				
6				
7				

8				
NO	Responden 5	Responden 6	Responden 7	Responden 8
1				
2				
3				
4				
5				
6				

7				
8				

No	Responden 9	Responden 10
1		
2		
3		
4		
5		

6			
7			
8			

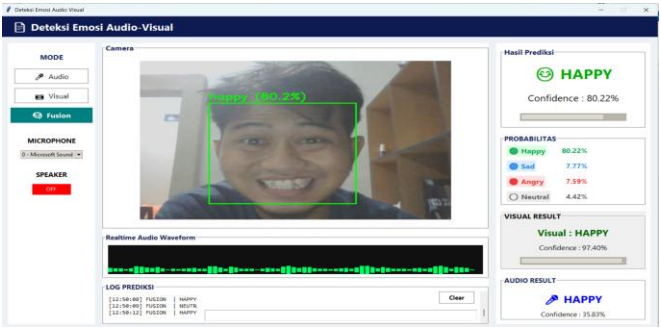
4.2.3 Hasil Pengujian Fusion

Pengujian sistem *Fusion* dilakukan dengan menggabungkan hasil prediksi dari sistem *Audio* dan sistem *Visual* menggunakan metode late *Fusion* dengan pendekatan weighted average.

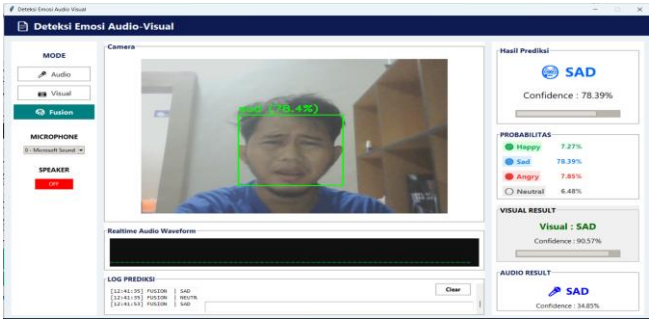
Pada penelitian ini digunakan pembobotan sebesar 75% untuk sistem *Visual* dan 25% untuk sistem *Audio*. Pembobotan lebih besar diberikan pada sistem *Visual* karena model *CNN* memiliki performa yang lebih stabil dibandingkan sistem *Audio* realtime.

Proses *Fusion* dilakukan dengan menggabungkan probabilitas hasil prediksi dari model *Audio* dan model *Visual*, kemudian sistem menentukan emosi akhir berdasarkan nilai probabilitas tertinggi.

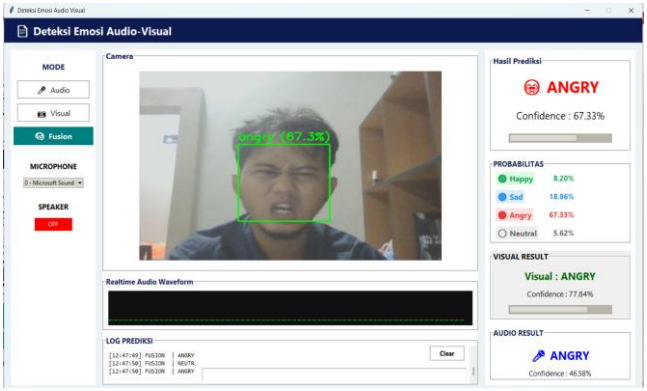
Pengujian *Fusion* dilakukan secara realtime menggunakan webcam dan microphone secara bersamaan. Sistem mampu menampilkan hasil deteksi emosi *Audio*, *Visual*, dan hasil *Fusion* secara langsung pada GUI.



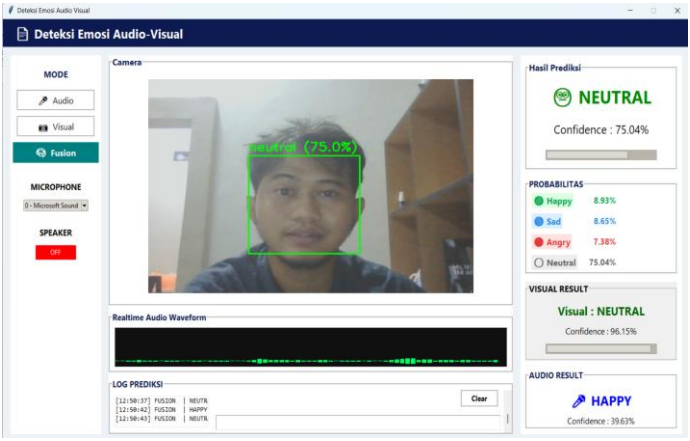
Gambar 4. 14 Hasil Pengujian *Fusion* Ouput (*Happy*)



Gambar 4. 15 Hasil Pengujian *Fusion* Ouput (*Sad*)



Gambar 4. 16 Hasil Pengujian *Fusion* Ouput (*Angry*)



Gambar 4. 17Hasil Pengujian *Fusion* Ouput (*Neutral*)

Gambar 4.8 menunjukkan hasil implementasi sistem *Fusion Audio-Visual* secara realtime menggunakan webcam dan microphone. Sistem berhasil mendeteksi ekspresi wajah pengguna dan mengklasifikasikan emosi sebagai *Happy* dengan confidence score sebesar 75.3%. Pada pengujian realtime, sistem mampu menampilkan hasil deteksi emosi secara langsung melalui GUI yang telah dibuat menggunakan Tkinter. GUI menampilkan hasil klasifikasi emosi, confidence score, mode sistem, serta deteksi wajah pengguna secara realtime.

Selain itu, sistem juga mendukung beberapa mode pengujian yaitu mode *Visual*, mode *Audio*, dan mode *Fusion* sehingga pengguna dapat melakukan pengujian sesuai kebutuhan. Penggunaan metode *Fusion* membantu meningkatkan kestabilan hasil prediksi dibandingkan penggunaan sistem *Audio* atau *Visual* secara terpisah.

Tabel 4. 16Pengujian untuk *Fusion* dengan Responden

NO		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10
1	*Happy*	*Happy* 62.1	*Happy* 68.7	*Neutral* 45.23	*Happy* 79.73	*Neutral* 45.22	*Happy* 71.34	*Happy* 78.21	*Happy* 81.36	*Happy* 77.28	*Happy* 62.58
2		*Happy* 66.4	*Happy* 74.66	*Neutral* 34.28	*Happy* 80.56	*Neutral* 52.81	*Happy* 73.20	*Happy* 76.4	*Happy* 76.51	*Happy* 72.58	*Happy* 70.53
3		*Happy* 73.5	*Happy* 75.34	*Neutral* 48.56	*Happy* 77.24	*Neutral* 63.53	*Happy* 74.93	*Happy* 80.45	*Happy* 78.33	*Happy* 71.53	*Happy* 73.65

4		*Happy* 73.8	*Happy* 74.06	*Neutral* 53.21	*Happy* 59.39	*Neutral* 62.84	*Happy* 80.98	*Happy* 70.72	*Happy* 76.19	*Happy* 73.65	*Happy* 80.22
5		*Happy* 82.8	*Neutral* 70.24	*Happy* 78.48	*Neutral* 77.13	*Neutral* 64.94	*Happy* 65.63	*Happy* 82.28	*Happy* 72.32	*Happy* 83.22	*Happy* 65.58
6		*Happy* 74.7	*Neutral* 72.86	*Neutral* 74.91	*Neutral* 80.56	*Neutral* 56.17	*Neutral* 72.84	*Happy* 73.56	*Happy* 73.28	*Happy* 75.76	*Happy* 69.53
7		*Happy* 81.6	*Happy* 67.85	*Happy* 76.34	*Happy* 81.55	*Neutral* 73.81	*Neutral* 56.94	*Happy* 75.56	*Happy* 75.37	*Happy* 84.58	*Happy* 73.45
8		*Happy* 78.5	*Happy* 56.3	*Happy* 70.20	*Happy* 56.3	*Neutral* 66.12	*Neutral* 66.17	*Happy* 78.36	*Happy* 80.26	*Happy* 72.58	*Happy* 83.22
9		*Happy* 80.5	*Happy* 49.38	*Happy* 79.93	*Happy* 77.59	*Neutral* 75.21	*Neutral* 77.81	*Happy* 74.51	*Happy* 65.48	*Happy* 71.53	*Happy* 68.86
10		*Happy* 81.3	*Happy* 62.44	*Happy* 80.98	*Happy* 73.33	*Neutral* 63.78	*Neutral* 61.84	*Happy* 81.33	*Happy* 69.43	*Happy* 73.65	*Happy* 58.46
11		*Happy* 80.9	*Happy* 63.02	*Happy* 75.63	*Happy* 74.35	*Neutral* 70.21	*Neutral* 69.94	*Happy* 72.19	*Happy* 61.75	*Happy* 83.22	*Happy* 80.51
12		*Happy* 83.3	*Happy* 64.05	*Happy* 73.56	*Happy* 82.47	*Neutral* 74.22	*Happy* 70.56	*Happy* 83.32	*Happy* 80.22	*Happy* 75.76	*Happy* 81.33
13		*Happy* 81.2	*Happy* 59.52	*Happy* 78.86	*Happy* 80.46	*Neutral* 76.23	*Happy* 67.24	*Happy* 73.21	*Happy* 81.86	*Happy* 78.46	*Happy* 71.19
14		*Happy* 71.3	*Happy* 47.78	*Happy* 71.65	*Happy* 80.74	*Neutral* 64.78	*Happy* 62.39	*Happy* 75.37	*Happy* 78.56	*Happy* 80.51	*Happy* 69.32
15		*Happy* 81.7	*Happy* 67.85	*Happy* 77.08	*Happy* 80.76	*Happy* 56.12	*Happy* 60.39	*Happy* 78.72	*Happy* 78.86	*Happy* 81.33	*Happy* 72.19
16	*Neutral*	*Happy* 81.9	*Neutral* 54.28	*Happy* 77.27	*Neutral* 44.65	*Neutral* 41.02	*Happy* 62.14	*Neutral* 47.32	*Neutral* 58.32	*Happy* 72.19	*Neutral* 70.76
17		*Happy* 79.4	*Neutral* 40.28	*Happy* 57.52	*Neutral* 52.81	*Neutral* 66.12	*Happy* 64.45	*Neutral* 56.28	*Neutral* 68.93	*Neutral* 47.65	*Neutral* 69.21
18		*Happy* 81.6	*Neutral* 60.91	*Neutral* 66.7	*Neutral* 47.53	*Neutral* 75.21	*Happy* 67.45	*Neutral* 60.13	*Neutral* 75.16	*Neutral* 53.81	*Neutral* 71.22
19		*Happy* 81.4	*Neutral* 61.79	*Happy* 80.98	*Neutral* 61.62	*Neutral* 63.78	*Happy* 59.28	*Neutral* 62.82	*Neutral* 60.55	*Neutral* 48.53	*Neutral* 72.23
20		*Happy* 78.4	*Neutral* 64.94	*Happy* 75.63	*Happy* 64.94	*Neutral* 70.21	*Happy* 68.38	*Neutral* 49.32	*Neutral* 69.62	*Neutral* 59.62	*Neutral* 71.16
21		*Happy*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Happy*	*Neutral*	*Neutral*	*Happy*	*Neutral*

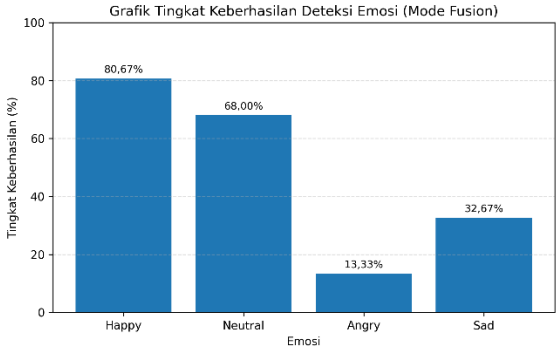
		75.3	74.91	66.73	62.11	74.22	73.27	55.28	72.13	71.09	73.55
22		*Happy* 80.1	*Neutral* 74.57	*Neutral* 66.88	*Neutral* 59.33	*Neutral* 76.23	*Happy* 61.56	*Neutral* 63.13	*Neutral* 64.82	*Happy* 75.73	*Happy* 69.62
23		*Happy* 83.6	*Neutral* 48.32	*Neutral* 58.33	*Neutral* 60.28	*Neutral* 70.21	*Sad* 53.40	*Neutral* 59.82	*Happy* 56.03	*Happy* 60.56	*Neutral* 60.91
24		*Happy* 79.4	*Neutral* 67.93	*Happy* 80.74	*Neutral* 62.66	*Neutral* 74.22	*Sad* 59.40	*Neutral* 48.32	*Neutral* 60.79	*Happy* 72.5	*Neutral* 61.79
25		*Happy* 79.3	*Neutral* 61.16	*Neutral* 70.1	*Neutral* 61.16	*Neutral* 76.23	*Happy* 63.67	*Neutral* 51.28	*Neutral* 68.32	*Neutral* 44.65	*Neutral* 60.94
26		*Happy* 78.8	*Neutral* 70.55	*Neutral* 59.01	*Neutral* 70.55	*Neutral* 70.21	*Happy* 62.82	*Neutral* 48.13	*Neutral* 67.93	*Neutral* 52.81	*Neutral* 71.91
27		*Happy* 81.09	*Neutral* 71.62	*Happy* 73.76	*Happy* 71.62	*Neutral* 74.22	*Happy* 61.72	*Neutral* 57.82	*Neutral* 51.16	*Neutral* 47.53	*Neutral* 73.57
28		*Happy* 75.7	*Neutral* 61.13	*Happy* 65.35	*Neutral* 56.13	*Neutral* 76.23	*Happy* 72.67	*Neutral* 48.72	*Neutral* 70.55	*Neutral* 43.62	*Neutral* 66.91
29		*Happy* 81.6	*Neutral* 64.82	*Neutral* 67.13	*Neutral* 54.78	*Neutral* 61.16	*Sad* 51.02	*Neutral* 46.28	*Neutral* 69.62	*Neutral* 48.65	*Neutral* 62.79
30		*Happy* 72.5	*Happy* 42.03	*Neutral* 65.06	*Happy* 49.12	*Neutral* 70.55	*Sad* 45.32	*Neutral* 55.23	*Neutral* 61.13	*Neutral* 51.81	*Neutral* 74.94
31	*Angry*	*Neutral* 47.2	*Neutral* 61.79	*Happy* 79.17	*Neutral* 41.02	*Happy* 71.62	*Sad* 51.95	*Angry* 60.01	*Neutral* 64.82	*Neutral* 73.53	*Neutral* 74.91
32		*Neutral* 42.1	*Neutral* 70.02	*Happy* 81.91	*Happy* 61.10	*Neutral* 56.13	*Sad* 47.18	*Sad* 57.72	*Happy* 42.03	*Neutral* 62.62	*Neutral* 74.57
33		*Sad* 48.8	*Neutral* 39.66	*Neutral* 66.72	*Happy* 61.56	*Neutral* 54.78	*Sad* 53.10	*Sad* 68.05	*Neutral* 61.79	*Neutral* 65.91	*Neutral* 60.91
34		*Neutral* 48.6	*Angry* 36.33	*Neutral* 46.36	*Happy* 62.13	*Neutral* 61.62	*Sad* 44.18	*Angry* 59.16	*Neutral* 72.13	*Neutral* 72.79	*Neutral* 61.79
35		Neutra 42.3	*Neutral* 34.75	*Neutral* 35.97	*Happy* 65.11	*Happy* 64.94	*Sad* 56.72	*Angry* 57.31	*Neutral* 64.42	*Neutral* 69.94	*Neutral* 64.94
36		*Neutral* 52.7	*Neutral* 38.65	*Neutral* 40.73	*Happy* 59.10	*Neutral* 62.11	*Neutral* 49.33	*Sad* 60.86	*Happy* 56.33	*Neutral* 71.91	*Neutral* 74.91
37		*Neutral* 42.00	*Angry* 36.55	*Neutral* 53.88	*Angry* 37.55	*Neutral* 59.33	*Sad* 47.82	*Sad* 58.05	*Neutral* 60.12	*Neutral* 74.57	*Neutral* 74.57
38		*Neutral* 56.9	*Sad* 35.68	*Neutral* 58.33	*Happy* 59.10	*Neutral* 60.28	*Neutral* 45.33	*Angry* 62.09	*Neutral* 68.65	*Neutral* 61.32	*Happy* 70.56

39		*Neutral*	*Neutral*	*Happy*	*Neutral*	*Neutral*	*Neutral*	*Angry*	*Neutral*	*Neutral*	*Happy*
		39.2	37.58	40.13	47.58	75.21	55.33	57.14	65.93	68.93	76.86
40		*Sad*	*Angry*	*Happy*	*Happy*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Happy*
		52.7	41.01	60.77	59.10	63.78	43.28	63.09	51.18	69.16	75.65
41		*Sad*	*Sad*	*Happy*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Angry*
		37.8	37.72	81.54	47.58	70.21	52.78	55.14	62.55	62.91	33.33
42		*Sad*	*Sad*	*Happy*	*Neutral*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		37.2	38.05	65.21	38.14	74.22	47.28	59.09	59.62	70.79	54.75
43		*Happy*	*Angry*	*Happy*	*Happy*	*Neutral*	*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*
		34.5	39.16	65.35	47.40	76.23	46.11	54.14	42.13	73.54	63.65
44		*Happy*	*Angry*	*Happy*	*Angry*	*Neutral*	*Neutral*	*Angry*	*Neutral*	*Neutral*	*Angry*
		40.5	37.31	80.60	37.31	64.78	49.33	60.09	54.82	74.91	32.55
45		*Sad*	*Sad*	*Happy*	*Neutral*	*Neutral*	*Neutral*	*Angry*	*Happy*	*Neutral*	*Sad*
		63.15	36.86	81.06	51.98	75.21	58.33	57.14	60.10	73.45	35.68
46		*Sad*	*Sad*	*Happy*	*Neutral*	*Neutral*	*Angry*	*Sad*	*Neutral*	*Neutral*	*Neutral*
		52.6	38.05	73.27	66.79	63.78	34.95	54.45	57.58	68.32	64.84
47		*Sad*	*Angry*	*Happy*	*Neutral*	*Neutral*	*Neutral*	*Sad*	*Neutral*	*Neutral*	*Happy*
		51.6	32.09	60.52	64.55	70.21	64.55	62.04	63.84	67.93	71.72
48		*Sad*	*Angry*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Sad*	*Happy*	*Neutral*	*Happy*
		56.4	36.14	72.11	60.22	71.22	60.22	65.33	71.52	61.16	65.35
49		*Happy*	*Sad*	*Happy*	*Happy*	*Neutral*	*Sad*	*Sad*	*Happy*	*Neutral*	*Happy*
		39.0	35.45	79.17	73.21	70.25	49.04	56.91	70.35	65.91	65.35
50		*Neutral*	*Neutral*	*Happy*	*Sad*	*Neutral*	*Sad*	*Sad*	*Happy*	*Neutral*	*Neutral*
		32.5	31.77	75.59	51.87	63.78	47.23	52.71	75.35	72.79	67.13
51	*sad*	*Neutral*	*Sad*	*Neutral*	*Sad*	*Neutral*	*Sad*	*Sad*	*Neutral*	*Neutral*	*Neutral*
		38.5	37.08	66.79	40.04	72.21	52.64	58.82	64.55	69.94	65.06
52		*Sad*	*Angry*	*Neutral*	*Sad*	*Neutral*	*Sad*	*Sad*	*Neutral*	*Neutral*	*Happy*
		52.0	34.95	64.55	47.20	56.68	51.69	65.27	60.22	71.91	73.21
53		*Sad*	*Neutral*	*Neutral*	*Sad*	*Neutral*	*Sad*	*Sad*	*Sad*	*Neutral*	*Happy*
		40.0	37.24	60.22	51.78	71.55	56.47	52.34	56.7	74.57	59.10
54		*Sad*	*Neutral*	*Happy*	*Neutral*	*Neutral*	*Happy*	*Sad*	*Sad*	*Neutral*	*Neutral*
		49.0	37.84	73.21	40.44	70.22	39.78	51.79	50.3	61.32	69.58
55		*Sad*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Neutral*	*Sad*	*Neutral*	*Neutral*	*Happy*
		51.0	40.44	79.34	44.12	72.23	42.85	53.57	67..13	68.31	60.10
56		*Sad*	*Neutral*	*Neutral*	*Sad*	*Neutral*	*Sad*	*Sad*	*Neutral*	*Sad*	*Neutral*

		41.8	44.12	64.84	32.70	6468	52.63	57.64	62.68	52.86	57.58
57		*Sad* 56.7	*Sad* 32.70	*Happy* 71.72	*Neutral* 67..13	*Neutral* 65.21	*Sad* 51.46	*Sad* 59.98	*Neutral* 57.23	*Sad* 51.65	*Neutral* 63.84
58		*Sad* 50.3	*Neutral* 39.92	*Happy* 65.35	*Neutral* 62.68	*Neutral* 67.57	*Sad* 53.56	*Sad* 60.88	*Neutral* 67..13	*Sad* 59.44	*Happy* 71.52
59		*Sad* 71.9	*Neutral* 38.15	*Neutral* 67.13	*Neutral* 57.23	*Neutral* 60.31	*Happy* 58.22	*Sad* 62.64	*Neutral* 62.68	*Happy* 42.0	*Happy* 70.35
60		*Sad* 60.1	*Neutral* 39.7	*Neutral* 65.06	*Neutral* 55.17	*Neutral* 71.22	*Neutral* 52.79	*Sad* 51.69	*Neutral* 57.23	*Sad* 54.43	*Happy* 75.35

Tabel 4. 17 Tingkat Keberhasilan Mode *Fusion* tiap Emosi

Emosi	**Total Data**	**Prediksi Benar**	**Prediksi Salah**	**Tingkat Keberhasilan**
Happy	150	121	29	80,67%
Neutral	150	102	48	68,00%
Angry	150	20	130	13,33%
Sad	150	49	101	32,67%



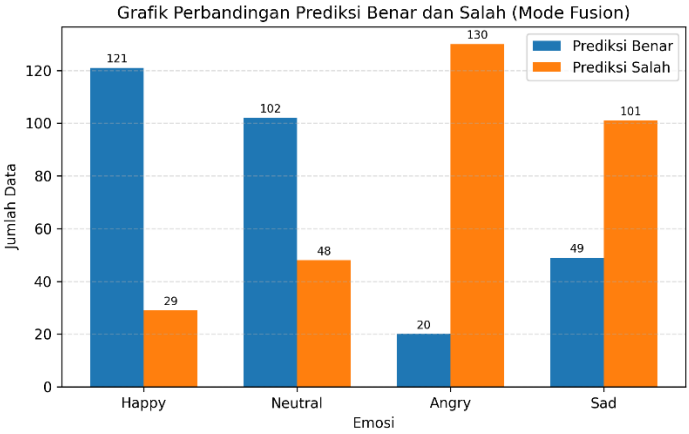
Gambar 4. 18 Tingkat Keberhasilan Mode *Fusion* tiap Emosi

Berdasarkan Tabel 4.17, hasil pengujian Mode *Fusion* menunjukkan bahwa sistem memiliki tingkat keberhasilan yang berbeda pada setiap kategori emosi. Emosi *Happy* memperoleh tingkat keberhasilan tertinggi sebesar 80,67%, dengan 121 prediksi benar dari 150 data pengujian. Selanjutnya, emosi *Neutral* memperoleh tingkat keberhasilan sebesar 68,00%, dengan 102 prediksi benar dari 150 data. Hasil tersebut menunjukkan bahwa metode *Weighted Late Fusion* mampu menggabungkan informasi dari modalitas *Audio* dan *Visual* dengan cukup baik sehingga sistem lebih mudah mengenali karakteristik emosi *Happy* dan *Neutral*.

Sebaliknya, emosi *Sad* hanya memperoleh tingkat keberhasilan sebesar 32,67%, sedangkan emosi *Angry* memiliki tingkat keberhasilan terendah yaitu 13,33%. Rendahnya tingkat keberhasilan pada kedua kategori emosi tersebut menunjukkan bahwa sistem masih mengalami kesalahan klasifikasi ketika menggabungkan hasil prediksi *Audio* dan *Visual*. Hal ini disebabkan oleh kemiripan karakteristik ekspresi wajah maupun pola suara pada emosi *Sad* dan *Angry*, sehingga keputusan akhir hasil *Fusion* masih belum mampu membedakan kedua emosi tersebut secara konsisten.

Tabel 4. 18 Tingkat Keberhasilan Tiap Responden Mode *Fusion*

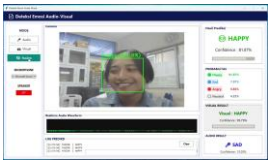
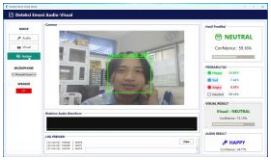
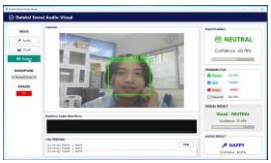
Responden	***Happy***	***Neutral***	***Angry***	***Sad***
Subject 1	100,00%	0,00%	0,00%	73,33%
Subject 2	86,67%	93,33%	33,33%	26,67%
Subject 3	66,67%	53,33%	0,00%	0,00%
Subject 4	86,67%	80,00%	13,33%	33,33%
Subject 5	6,67%	100,00%	0,00%	0,00%
Subject 6	60,00%	0,00%	0,00%	53,33%
Subject 7	100,00%	100,00%	73,33%	100,00%
Subject 8	100,00%	93,33%	0,00%	13,33%
Subject 9	100,00%	66,67%	0,00%	26,67%
Subject 10	100,00%	93,33%	13,33%	0,00%

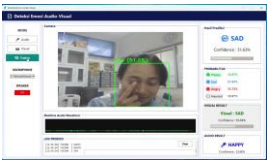


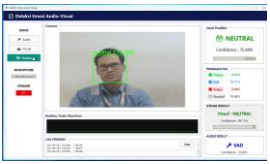
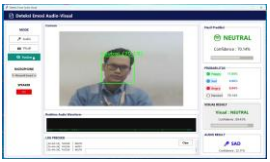
Gambar 4. 19 Tingkat Keberhasilan Tiap Responden Mode *Fusion*

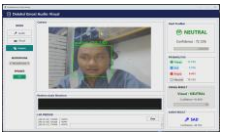
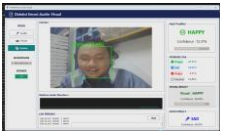
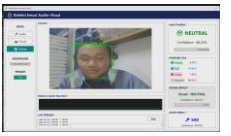
Secara keseluruhan, hasil pengujian menunjukkan bahwa metode *Weighted Late Fusion* memberikan performa yang baik pada emosi *Happy* dan *Neutral*, namun masih memerlukan pengembangan lebih lanjut untuk meningkatkan kemampuan sistem dalam mengenali emosi *Sad* dan *Angry*. Peningkatan performa dapat dilakukan melalui penambahan jumlah data latih, penggunaan teknik ekstraksi fitur yang lebih representatif, maupun penerapan metode *Fusion* yang lebih adaptif.

Tabel 4. 19 Hasil Pengujian Deteksi Emosi melalui mode Fusion

NO	Responden 1	Responden 2	Responden 3	Responden 4
1				
2				
3				
4				
5				
6				

NO	Responden 5	Responden 6	Responden 7	Responden 8
7				
8				
1				
2				
3				
4				

5				
6				
7				
8				
No	Responden 9	Responden 10		
1				
2				

3			
4			
5			
6			
7			
8			

4.3 Analisa Hasil

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan, sistem deteksi emosi berbasis *Audio-Visual* berhasil diimplementasikan dan dijalankan secara real-time menggunakan webcam dan mikrofon. Sistem mampu melakukan klasifikasi emosi melalui tiga mode pengujian, yaitu Mode *Audio*, Mode *Visual*, dan Mode *Fusion*, sehingga pengguna dapat memilih metode deteksi sesuai dengan kebutuhan.

- **Analisis Hasil *Train*ing *Audio***

Pada tahap pelatihan model *Audio*, metode *Mel-Frequency Cepstral Coefficients (MFCC)* digunakan untuk mengekstraksi ciri suara, sedangkan algoritma *Random Forest* digunakan sebagai metode klasifikasi. Berdasarkan hasil pelatihan, model *Audio* memperoleh tingkat akurasi sebesar 75,56%. Nilai tersebut menunjukkan bahwa sekitar 75 dari setiap 100 data uji berhasil diklasifikasikan sesuai dengan label emosi yang sebenarnya, sedangkan 24,44% data lainnya merupakan data yang salah diklasifikasikan (*misclassification*) oleh model.

Kesalahan klasifikasi tersebut menunjukkan bahwa model masih mengalami kesulitan dalam membedakan karakteristik suara pada beberapa kategori emosi. Kondisi ini disebabkan oleh adanya kemiripan pola akustik antar emosi, terutama pada emosi *Happy*, *Neutral*, dan *Angry*, sehingga model belum mampu mempelajari seluruh variasi karakteristik suara secara optimal. Selain itu, kualitas rekaman, kebisingan lingkungan (*noise*), variasi intonasi, serta perbedaan cara berbicara setiap individu juga memengaruhi proses ekstraksi fitur MFCC sehingga berdampak pada hasil klasifikasi.

Hasil tersebut kemudian dibuktikan melalui pengujian terhadap 10 responden dengan total 600 data pengujian. Berdasarkan hasil pengujian, sistem memiliki tingkat keberhasilan tertinggi pada emosi *Sad*, sedangkan emosi *Happy*, *Neutral*, dan *Angry* memiliki tingkat keberhasilan yang lebih rendah. Hal ini menunjukkan bahwa model *Audio* cenderung lebih mudah mengenali karakteristik suara pada emosi *Sad*, namun masih mengalami kesalahan dalam membedakan emosi lainnya. Dengan demikian, hasil pengujian terhadap responden mendukung

hasil evaluasi pada tahap pelatihan, dimana akurasi model sebesar 75,56% masih menunjukkan adanya data yang belum berhasil diklasifikasikan dengan benar.

- **Analisis Hasil *Train*ing *Visual***

Pada tahap pelatihan model *Visual*, digunakan metode *Convolutional Neural Network (CNN)* untuk mengenali pola ekspresi wajah berdasarkan dataset FER2013. Berdasarkan hasil pelatihan, model *Visual* memperoleh tingkat akurasi sebesar 87,48%. Nilai tersebut menunjukkan bahwa 87,48% data uji berhasil diklasifikasikan sesuai dengan label sebenarnya, sedangkan 12,52% sisanya merupakan data yang salah diklasifikasikan oleh model.

Kesalahan klasifikasi tersebut umumnya terjadi karena adanya kemiripan karakteristik ekspresi wajah antar beberapa kategori emosi, seperti *Angry* yang menyerupai *Sad* atau *Neutral*, serta perbedaan intensitas ekspresi pada setiap individu. Selain itu, variasi pencahayaan, posisi wajah, dan kualitas citra pada dataset juga memengaruhi kemampuan model dalam mempelajari pola ekspresi secara menyeluruh.

Hasil pelatihan tersebut sejalan dengan hasil pengujian terhadap 10 responden, dimana sistem *Visual* menunjukkan tingkat keberhasilan yang tinggi, khususnya pada emosi *Happy* dan *Neutral*. Tingginya tingkat keberhasilan tersebut menunjukkan bahwa model *CNN* mampu mengenali pola ekspresi wajah dengan baik pada kondisi pengujian real-time. Meskipun demikian, beberapa kesalahan klasifikasi masih ditemukan pada emosi *Sad* dan *Angry*, sehingga hasil pengujian tetap mencerminkan adanya 12,52% data yang belum berhasil dikenali dengan benar selama proses pelatihan.

- **Analisis Hasil Pengujian Responden**

Pengujian terhadap 10 responden dilakukan untuk mengevaluasi kemampuan sistem dalam kondisi penggunaan secara langsung. Setiap responden melakukan 60 kali pengujian, yang terdiri dari 15 data emosi *Happy*, 15 data *Neutral*, 15 data *Angry*, dan 15 data *Sad*, sehingga total data pengujian pada setiap mode berjumlah 600 data.

Berdasarkan hasil pengujian, Mode *Visual* menunjukkan tingkat keberhasilan yang lebih tinggi dibandingkan Mode *Audio*. Hal tersebut disebabkan karena ekspresi wajah memiliki karakteristik *Visual* yang lebih mudah dibedakan oleh model *CNN*, sedangkan karakteristik suara lebih dipengaruhi oleh faktor eksternal seperti noise, kualitas mikrofon, intonasi, serta cara berbicara setiap responden.

Pada Mode *Fusion*, hasil prediksi *Audio* dan *Visual* digabungkan menggunakan metode *Weighted Late Fusion* dengan pembobotan 75% untuk *Visual* dan 25% untuk *Audio*. Penggabungan kedua modalitas tersebut menghasilkan sistem yang lebih stabil dibandingkan penggunaan *Audio* saja, karena ketika salah satu modalitas mengalami penurunan performa, modalitas lainnya masih dapat memberikan informasi tambahan dalam proses pengambilan keputusan.

Secara keseluruhan, hasil pengujian terhadap responden menunjukkan bahwa sistem telah mampu melakukan deteksi emosi secara real-time, meskipun masih terdapat beberapa kesalahan klasifikasi pada kategori emosi tertentu. Hasil tersebut menunjukkan bahwa kombinasi metode MFCC, *Random Forest*, *CNN*, dan *Weighted Late Fusion* mampu meningkatkan keandalan sistem dibandingkan penggunaan satu modalitas saja.

Bab 5. Kesimpulan

5.1 Kesimpulan

Berdasarkan hasil penelitian, implementasi, dan pengujian yang telah dilakukan, dapat disimpulkan bahwa sistem deteksi emosi berbasis *Audio-Visual* berhasil dikembangkan menggunakan metode *Mel-Frequency Cepstral Coefficients (MFCC)* dan algoritma *Random Forest* pada sistem *Audio* serta *Convolutional Neural Network (CNN)* pada sistem *Visual*. Kedua modalitas tersebut kemudian diintegrasikan menggunakan metode *Weighted Late Fusion* sehingga mampu melakukan deteksi emosi secara real-time melalui webcam dan mikrofon.

Berdasarkan hasil pelatihan model, sistem *Audio* memperoleh tingkat akurasi sebesar 75,56%, sedangkan sistem *Visual* memperoleh tingkat akurasi sebesar 87,48%. Hasil tersebut menunjukkan bahwa model *Visual* memiliki kemampuan klasifikasi yang lebih baik dibandingkan model *Audio*. Selisih sebesar 24,44% pada model *Audio* dan 12,52% pada model *Visual* merupakan data yang masih mengalami kesalahan klasifikasi (*misclassification*) selama proses evaluasi model.

Selanjutnya, dilakukan pengujian sistem terhadap 10 responden, dimana setiap responden melakukan 60 kali pengujian yang terdiri atas empat kategori emosi, yaitu *Happy*, *Neutral*, *Angry*, dan *Sad*, sehingga total data pengujian pada setiap mode berjumlah 600 data. Berdasarkan hasil pengujian tersebut, sistem menunjukkan tingkat keberhasilan alat yang berbeda pada setiap mode pengujian.

Pada Mode *Visual*, sistem memiliki tingkat keberhasilan tertinggi pada emosi *Happy* sebesar 90,00% dan *Neutral* sebesar 86,00%, sedangkan emosi *Sad* dan *Angry* memperoleh tingkat keberhasilan sebesar 56,67% dan 24,67%.

Pada Mode *Audio*, sistem menunjukkan tingkat keberhasilan tertinggi pada emosi *Sad* sebesar 88,67%, sedangkan emosi *Happy*, *Neutral*, dan *Angry* memperoleh tingkat keberhasilan masing-masing sebesar 10,67%, 0,00%, dan 10,00%. Hasil tersebut menunjukkan bahwa sistem *Audio* masih dipengaruhi oleh karakteristik suara pengguna, kualitas mikrofon, serta kebisingan lingkungan pada saat proses perekaman.

Pada Mode *Fusion*, hasil penggabungan prediksi *Audio* dan *Visual* menggunakan metode *Weighted Late Fusion* menghasilkan tingkat keberhasilan tertinggi pada emosi *Happy* sebesar 80,67%, diikuti emosi *Neutral* sebesar 68,00%,

sedangkan emosi *Sad* dan *Angry* memperoleh tingkat keberhasilan sebesar 32,67% dan 13,33%. Hasil tersebut menunjukkan bahwa metode *Fusion* mampu memanfaatkan informasi dari kedua modalitas untuk menghasilkan sistem yang lebih stabil dibandingkan penggunaan modalitas *Audio* secara tunggal.

Secara keseluruhan, penelitian ini berhasil mencapai tujuan yang telah ditetapkan, yaitu membangun sistem deteksi emosi manusia berbasis *Audio-Visual* menggunakan kombinasi metode MFCC, *Random Forest*, *CNN*, dan Weighted Late *Fusion*, serta berhasil mengimplementasikan sistem dalam bentuk aplikasi GUI berbasis Python yang mampu melakukan deteksi emosi secara real-time dan telah diuji langsung terhadap 10 responden.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat digunakan untuk pengembangan sistem pada penelitian selanjutnya, yaitu:

1. Menambah jumlah dan variasi dataset *Audio* serta *Visual* sehingga model mampu mempelajari karakteristik emosi yang lebih beragam dan meningkatkan akurasi model maupun tingkat keberhasilan sistem pada pengujian real-time.
2. Menggunakan perangkat perekaman yang lebih baik, seperti mikrofon eksternal dan webcam beresolusi tinggi, untuk meningkatkan kualitas data masukan serta mengurangi pengaruh kebisingan (*noise*), pencahayaan, dan kualitas citra terhadap proses deteksi emosi.
3. Mengembangkan model *Audio* menggunakan metode Deep Learning, seperti Long Short-Term Memory (LSTM), Convolutional Neural Network (*CNN*), atau Transformer, sehingga sistem dapat mengenali karakteristik suara dengan lebih baik dan meningkatkan kemampuan klasifikasi emosi.
4. Melakukan pengujian terhadap jumlah responden yang lebih banyak dengan rentang usia, jenis kelamin, dan karakteristik suara maupun ekspresi wajah yang lebih beragam. Hal ini bertujuan untuk mengetahui kemampuan sistem dalam melakukan generalisasi pada kondisi penggunaan yang lebih luas.
5. Menambahkan kategori emosi lain, seperti Fear, Disgust, dan Surprise, sehingga sistem mampu mengenali emosi manusia secara lebih lengkap sesuai dengan kondisi nyata.
6. Mengembangkan metode *Audio-Visual Fusion* menggunakan pendekatan yang lebih adaptif, seperti Attention-Based *Fusion*, Decision-Level *Fusion*, atau metode pembelajaran multimodal lainnya, sehingga proses penggabungan informasi *Audio* dan *Visual* dapat menghasilkan keputusan klasifikasi yang lebih akurat dan konsisten.

7. Melakukan optimasi proses deteksi secara real-time, terutama pada sistem *Audio*, agar lebih tahan terhadap pengaruh noise, perbedaan intonasi, dan kualitas perekaman suara. Optimasi ini diharapkan dapat meningkatkan tingkat keberhasilan sistem pada kondisi penggunaan secara langsung.
8. Mengembangkan aplikasi menjadi perangkat lunak yang lebih lengkap, baik dalam bentuk aplikasi desktop, mobile, maupun *embedded system*, sehingga sistem dapat diterapkan pada berbagai bidang, seperti Human Computer Interaction (HCI), pendidikan, kesehatan, layanan pelanggan, maupun sistem pemantauan kondisi emosional pengguna.

Daftar Pustaka

[1] K. A. Araño, P. Gloor, C. Orsenigo, and C. Vercellis, "When Old Meets New: Emotion Recognition from Speech Signals," *Cognit. Comput.*, 2021, doi: 10.1007/s12559-021-09865-2.

[2] Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," 2021, [Online]. Available: https://www.researchgate.net/publication/351476658_Facial_Emotion_Recognition_State_of_the_Art_Performance_on_FER2013

[3] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech Emotion Recognition Using Machine Learning Techniques: Feature Extraction and Comparison of CNN and Random Forest," *PLoS One*, 2023, doi: 10.1371/journal.pone.0291500.

[4] G. T. Waleed and S. H. Shaker, "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN," *Information*, pp. 1–18, 2025, doi: 10.3390/info16070518.

[5] A. Z. Fadhil, Muhammad Farhan, "Speech Emotion Recognition with Optimized Multi-feature Stack Using Deep CNN," *Biomed. Signal Process. Control*, 2024, [Online]. Available: https://www.researchgate.net/publication/386302099_Speech_emotion_recognition_with_optimized_multi-feature_stack_using_deep_convolutional_neural_networks

[6] H. Heidari, M. Murugappan, J. Shaikh Mohammed, and M. E. H. Chowdhury, "A Novel Partitioned Random Forest Method-Based Facial Emotion Recognition," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3560362.

[7] Q. Hu, Y. Peng, and Z. Zheng, "A Deep Learning Framework for Gender Sensitive Speech Emotion Recognition Based on MFCC Feature Selection

and SHAP Analysis," *Expert Syst. Appl.*, 2025, [Online]. Available: https://www.researchgate.net/publication/394317617_A_deep_learning_framework_for_gender_sensitive_speech_emotion_recognition_based_on_MFCC_feature_selection_and_SHAP_analysis

[8] M. Sharafi, "A Novel Spatio-Temporal Convolutional Neural Framework for Multimodal Emotion Recognition," *Expert Syst. Appl.*, 2022, [Online]. Available: https://www.researchgate.net/publication/383885857_Multimodal_Emotion_Recognition_Using_Visual_Vocal_and_Physiological_Signals_A_Review

[9] Y. Cai, X. Li, and J. Li, "Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review," *Electronics*, 2023, [Online]. Available: https://www.researchgate.net/publication/368769173_Emotion_Recognition_Using_Different_Sensors_Emotion_Models_Methods_and_Datasets_A_Comprehensive_Review/fulltext/63f8d341b1704f343f7c8e70/Emotion-Recognition-Using-Different-Sensors-Emotion-Models-Methods-

[10] Y. Wu, Q. Mi, and T. Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," *Artif. Intell. Rev.*, 2025, [Online]. Available: https://www.mdpi.com/2313-7673/10/7/418

[11] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal Emotion Recognition: A Comprehensive Review, Trends, and Challenges," *Multimed. Tools Appl.*, 2024, [Online]. Available: https://www.researchgate.net/publication/384746023_Multimodal_emotion_recognition_A_comprehensive_review_trends_and_challenges

[12] E. M. G. Younis, "Machine Learning for Human Emotion Recognition: A Comprehensive Review," *J. Electr. Syst.*, 2024, [Online]. Available:

https://link.springer.com/article/10.1007/s00521-024-09426-2

[13] S. Zhang, “Deep Learning-Based Multimodal Emotion Recognition from Audio, Visual, and Text Modalities,” *Appl. Sci.*, 2024.

[14] P. Thakur, N. Kaur, and N. Aggarwal, “A Comprehensive Review of Unimodal and Multimodal Emotion Detection,” *Arch. Comput. Methods Eng.*, 2025, [Online]. Available: https://www.semanticscholar.org/paper/A-Comprehensive-Review-of-Unimodal-and-Multimodal-Thakur-Kaur/86df1a19ac1c8a7c51796c2e04e37e79d96fc617

[15] M. J. Dileep Kumar, “Multimodal Emotion Recognition: A Comprehensive Survey of Datasets, Methods, and Applications,” *IEEE Access*, 2025, [Online]. Available: https://ui.adsabs.harvard.edu/abs/2025IEEEA..1301067D/abstract

[16] C. Wu, “Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects,” *Artif. Intell. Rev.*, 2025, [Online]. Available: https://aclanthology.org/2025.findings-emnlp.332/

[17] R. Mobbs, D. Makris, and V. Argyriou, “Emotion Recognition and Generation: A Comprehensive Review of Face, Speech, and Text Modalities,” *Inf. Fusion*, 2025, [Online]. Available: https://www.researchgate.net/publication/388920796_Emotion_Recognition_and_Generation_A_Comprehensive_Review_of_Face_Speech_and_Text_Modalities

[18] Q. Ouyang, “Speech Emotion Detection Based on MFCC and CNN-LSTM Architecture,” *J. Intell. Syst.*, 2025, [Online]. Available: https://arxiv.org/abs/2501.10666

[19] B. C. Gül, “FedMultiEmo: Real-Time Emotion Recognition via Multimodal Federated Learning,” *IEEE Access*, 2025, [Online]. Available: https://arxiv.org/abs/2507.15470

[20] T. S. Gunawan, “Speech Emotion Recognition Using Deep Learning and MFCC Features,” *Int. J. Adv. Comput. Sci. Appl.*, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s00521-026-12186-w

[21] R. Donatus, “Interpretable Speech Emotion Recognition: A Comparative Study of BiLSTM Temporal Attention and Transformer-Based Multi-Head Self-Attention,” *Asian J. Electr. Sci.*, vol. 14, pp. 21–27, Oct. 2025, doi: 10.70112/ajes-2025.14.2.4286.

[22] R. E. Donatus, B. L. Pal, I. H. Donatus, and U. O. Chiedu, “Comparative Analysis of Spectrogram and MFCC Representations for Speech Emotion Recognition Using Machine Learning,” *African J. Comput. Sci. Technol.*, 2024.

[23] I. N. Afifah, T. Dutono, and T. B. Santoso, “Speech Emotional Recognition of Telephone Conversation by Using Deep Learning,” *Int. J. Commun. Networks Inf. Secur.*, 2024, [Online]. Available: https://www.ijcs.net/ijcs/index.php/ijcs/article/view/4442

[24] A. Lamichhane and G. Karn, “CNN-BiLSTM Based Facial Emotion Recognition,” *Int. J. Eng. Trends Technol.*, 2024, [Online]. Available: https://www.nepjol.info/index.php/injet/article/view/72579

[25] T. M. Little Flower, “Data Augmentation Using a 1D-CNN Model with MFCC/MFMC Features for Speech Emotion Recognition,” *J. Inst. Eng. Ser. B*, 2024, [Online]. Available: https://hrcak.srce.hr/en/326329

[26] T. M. Little Flower, “Speech Emotion Recognition via CNN-Transformer and Multidimensional Attention Mechanism,” *arXiv*, 2024, [Online]. Available: https://arxiv.org/abs/2403.04743

[27] J. Yu, W. Zhu, and J. Zhu, “Efficient Feature Extraction and Late Fusion Strategy for Audiovisual Emotional Mimicry Intensity Estimation,” *Pattern Recognit. Lett.*, 2024, [Online]. Available: https://arxiv.org/abs/2403.11757

[28] W. Dai, D. Zheng, F. Yu, Y. Zhang, and Y. Hou, “A Novel Approach for Multimodal Emotion Recognition: Multimodal Semantic Information Fusion,” *Knowledge-Based Syst.*, 2025, [Online]. Available: https://arxiv.org/abs/2502.08573

[29] Y. Wu, Q. Mi, and T. Gao, “A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions,” *Artif. Intell. Rev.*, 2025, [Online]. Available: file:///C:/Users/LENOVO/Downloads/biomimetics-10-00418-v2.pdf

[30] X. Tang, Y. Lin, T. Dang, Y. Zhang, and J. Cheng, “Speech Emotion Recognition Using CNN and Transformer,” 2024.

[31] A. of AVaTER, “AVaTER: Fusing Audio, Visual, and Textual Modalities Using Cross-Modal Attention for Emotion Recognition,” *arXiv*, 2024, [Online]. Available: https://www.mdpi.com/1424-8220/24/18/5862

[32] T. Fawcett, “An Introduction to ROC Analysis,” 2006, [Online]. Available: https://people.inf.elte.hu/kiss/13dwhdm/roc.pdf

Biodata

Nama	: M .Afif Fuadie Yusuf
TTL	: Batam, 15 Maret 1998
Agama	: Islam
Alamat	: Perumahan, Sarmen Raya Blok J.01 Bengkong, Batam, Kepulauan Riau
Email	: afifyusuf055@gmail.com
Riwayat Pendidikan	SMA/SMK : SMKN 1 Batam
	SMP : SMPN 30 Batam

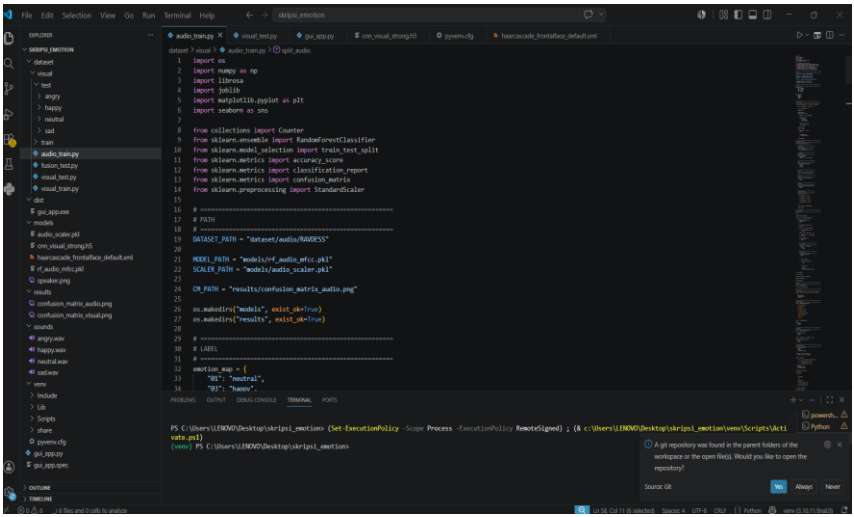
Lampiran

Berikut adalah tautan Source code Aplikasi Sistem Deteksi Emosi manusia dapat diunduh pada tautan berikut ini:

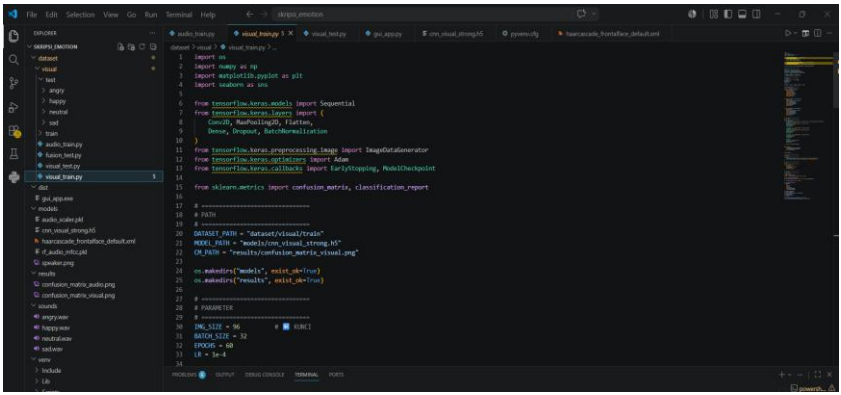
https://github.com/afifyusuf055/Sistem-Deteksi-Emosi-Manusia-menggunakan-metode-*Audio-Visual*-berbasis-MFCC-Random-Forest-dan-*CNN*.git

dan berikut merupakan link untuk aplikasi :

https://drive.google.com/drive/folders/1aur7kSmSEHq-pC0-vRwG7ckBNHmNScp_?usp=drive_link

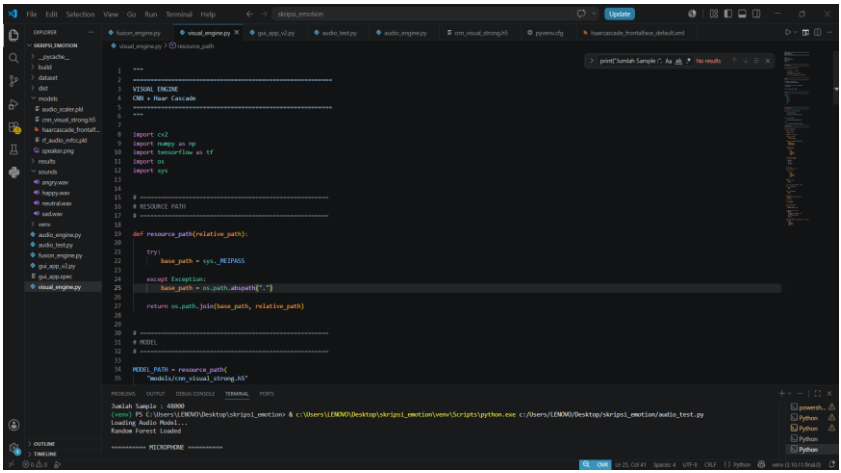


Gambar 7. 1 Potongan Baris Program *AudioTrain*



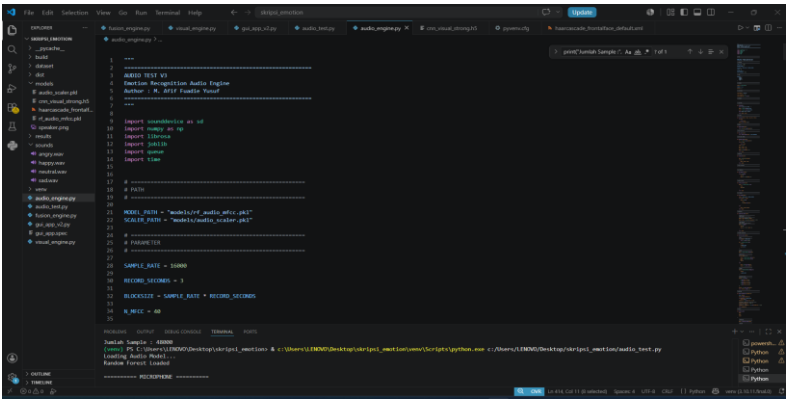
```
import os
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import (
    Conv2D, MaxPooling2D, Flatten,
    Dense, Dropout, BatchNormalization
)
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping, ModelCheckpoint

from sklearn.metrics import confusion_matrix, classification_report

# ==========================
# PATH
# ==========================
DATASET_PATH = "dataset/visual/train"
MODEL_PATH = "models/cnn_visual_strong.h5"
CM_PATH = "results/confusion_matrix_visual.png"

os.makedirs("models", exist_ok=True)
os.makedirs("results", exist_ok=True)

# ==========================
# PARAMETER
# ==========================
IMG_SIZE = 96
BATCH_SIZE = 32
EPOCHS = 60
LR = 1e-4
```

Gambar 7. 2 Potongan Baris Program *Visual Train*

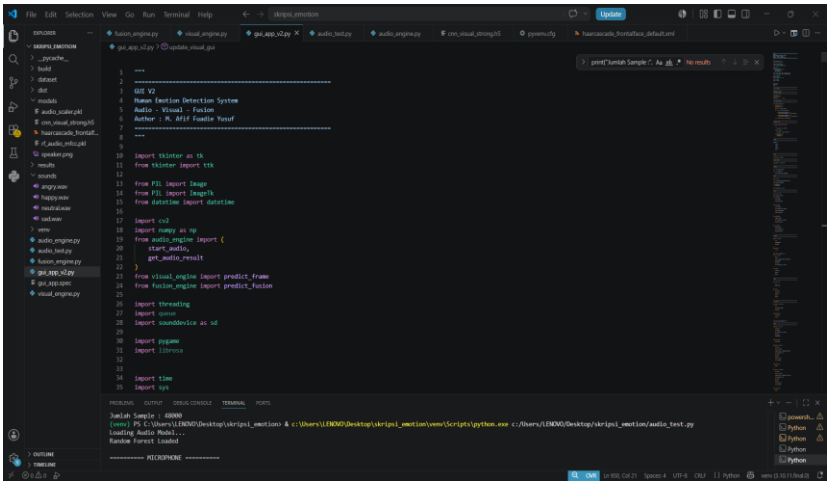


```
"""
==========================================================
VISUAL ENGINE
CNN + Haar Cascade
==========================================================
"""

import cv2
import numpy as np
import tensorflow as tf
import os
import sys

# ==========================================================
# RESOURCE PATH
# ==========================================================
def resource_path(relative_path):

    try:
        base_path = sys._MEIPASS

    except Exception:
        base_path = os.path.abspath(".")

    return os.path.join(base_path, relative_path)


# ==========================================================
# MODEL
# ==========================================================

MODEL_PATH = resource_path(
    "models/cnn_visual_strong.h5"
)
```

Gambar 7. 3 Potongan Program *Visual* Test



Gambar 7. 4 Potongan Program *Audio* Test



Gambar 7. 5 Potongan Baris Program Gui