TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

Implementasi Multimodal Emotion Recognition Menggunakan CNN-BiLSTM-Attention pada Audio dan Mobilenetv2 pada Ekspresi Wajah

Ai Solihah*, Alun Sujjada, Ivana Lucia Kharisma

Fakultas Komputer Teknik dan Desain, Program Studi Teknik Informatika, Universitas Nusa Putra, Sukabumi, Indonesia
Email: [1,*]ai.solihah_ti22@nusaputra.ac.id, [2]alun.sujjada@nusaputra.ac.id, [3]ivana.lucia@nusaputra.ac.id
Email Penulis Korespondensi: ai.solihah_ti22@nusaputra.ac.id

Abstrak–Perkembangan teknologi kecerdasan buatan telah mendorong pengembangan sistem pengenalan emosi yang mampu memahami kondisi emosional manusia secara lebih akurat. Namun, pendekatan unimodal yang hanya memanfaatkan satu sumber data, seperti suara atau ekspresi wajah, masih memiliki keterbatasan dalam menghadapi kondisi lingkungan yang dinamis. Penelitian ini bertujuan untuk mengimplementasikan sistem *Multimodal Emotion Recognition* dengan menggabungkan modalitas audio dan visual guna meningkatkan akurasi deteksi emosi. Modalitas audio diproses menggunakan arsitektur CNN-BiLSTM-Attention dengan masukan berupa Mel-Spectrogram, sedangkan modalitas visual diproses menggunakan MobileNetV2 untuk mengenali ekspresi wajah. Integrasi kedua modalitas dilakukan menggunakan metode *Late Fusion* pada tingkat keputusan (*decision level*) dengan pendekatan *weighted fusion*, yaitu bobot 0,4 untuk audio dan 0,6 untuk visual. Metode penelitian yang digunakan adalah metode eksperimen dengan pendekatan kuantitatif. Tahapan penelitian meliputi studi literatur, pengumpulan data, *preprocessing* audio dan video, perancangan model, pelatihan model, integrasi multimodal, serta evaluasi sistem. Data audio diekstraksi menjadi fitur Mel-Spectrogram, sedangkan data video diproses menggunakan MediaPipe Face Detection untuk memperoleh area wajah sebelum dilakukan klasifikasi emosi. Sistem dirancang untuk mengenali tiga kelas emosi, yaitu netral, senang, dan marah. Evaluasi performa model dilakukan menggunakan *Confusion Matrix*, akurasi, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa integrasi multimodal menggunakan metode *Late Fusion* mampu memanfaatkan keunggulan masing-masing modalitas sehingga menghasilkan deteksi emosi yang lebih akurat dan stabil dibandingkan penggunaan satu modalitas saja. Sistem yang dikembangkan juga mampu melakukan klasifikasi emosi secara *real-time* melalui kamera dan mikrofon, sehingga berpotensi diterapkan pada berbagai bidang seperti pendidikan, layanan pelanggan, sistem interaksi manusia-komputer, dan pemantauan kondisi emosional pengguna. Kontribusi utama penelitian ini adalah pengembangan sistem *Multimodal Emotion Recognition* yang mengintegrasikan model CNN-BiLSTM-Attention pada modalitas audio dan MobileNetV2 pada modalitas visual menggunakan metode *Late Fusion*. Sistem yang diusulkan mampu memanfaatkan informasi emosi dari dua modalitas secara bersamaan untuk meningkatkan akurasi dan stabilitas deteksi emosi secara real-time, sehingga dapat menjadi alternatif solusi dalam pengembangan sistem interaksi manusia-komputer berbasis kecerdasan buatan.

Kata Kunci: Multimodal Emotion Recognition; CNN-BiLSTM-Attention; MobileNetV2; Late Fusion; Ekspresi Wajah; Audio; Deep Learning

Abstract–The advancement of artificial intelligence technology has significantly contributed to the development of emotion recognition systems capable of understanding human emotional states more accurately. However, unimodal approaches that rely on a single data source, such as speech or facial expressions, still face limitations when dealing with dynamic environmental conditions. This study aims to implement a Multimodal Emotion Recognition system by integrating audio and visual modalities to improve emotion detection accuracy. The audio modality is processed using a CNN-BiLSTM-Attention architecture with Mel-Spectrogram representations as input, while the visual modality is analyzed using MobileNetV2 to recognize facial expressions. The integration of both modalities is performed using the Late Fusion method at the decision level through a weighted fusion approach, assigning weights of 0.4 and 0.6 to the audio and visual modalities, respectively. This research employs an experimental method with a quantitative approach. The research stages include literature review, data collection, audio and video preprocessing, model design, model training, multimodal integration, and system evaluation. Audio data are transformed into Mel-Spectrogram features, while video data are processed using MediaPipe Face Detection to extract facial regions prior to emotion classification. The proposed system is designed to recognize three emotional classes: neutral, happy, and angry. Model performance is evaluated using a Confusion Matrix, accuracy, precision, recall, and F1-score metrics. The results indicate that multimodal integration using the Late Fusion method effectively leverages the strengths of each modality, resulting in more accurate and robust emotion detection compared to unimodal approaches. Furthermore, the developed system is capable of performing real-time emotion classification through a camera and microphone, making it potentially applicable in various domains, including education, customer service, human-computer interaction systems, and user emotional state monitoring. The primary contribution of this research is the development of a multimodal emotion recognition system that integrates a CNN-BiLSTM-Attention model for the audio modality and MobileNetV2 for the visual modality using a late fusion method. The proposed system leverages emotional information from both modalities simultaneously to enhance the accuracy and stability of real-time emotion detection, thereby offering an alternative solution for the development of AI-based human-computer interaction systems.

Keywords: Multimodal Emotion Recognition; CNN-BiLSTM-Attention; MobileNetV2; Late Fusion; Facial Expression Recognition; Audio Processing; Deep Learning

1. PENDAHULUAN

Emosi merupakan aspek *fundamental* dalam kehidupan manusia yang berperan penting dalam proses komunikasi, pengambilan keputusan, serta respons perilaku. Emosi tidak hanya mencerminkan kondisi psikologis seseorang, tetapi juga mempengaruhi cara individu berinteraksi dengan lingkungan dan individu lain(Ekspresi & Sosial, 2025). Pemahaman terhadap emosi menjadi komponen penting dalam berbagai bidang, seperti psikologi, pendidikan,

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

kesehatan, serta interaksi manusia dan komputer. Deteksi emosi, atau pengenalan emosi secara otomatis, memiliki kegunaan luas dalam aplikasi praktis. Misalnya, dalam bidang kesehatan mental, deteksi emosi dapat membantu mengidentifikasi gangguan seperti depresi atau kecemasan melalui analisis ekspresi wajah dan nada suara, memungkinkan intervensi dini (S Budiawan, ST Alvianus Dengen, 2025). Di bidang pendidikan, sistem deteksi emosi dapat menyesuaikan materi pembelajaran berdasarkan emosi siswa, seperti meningkatkan motivasi saat mendeteksi kebosanan. Dalam interaksi manusia dan komputer, deteksi emosi meningkatkan empati mesin, seperti asisten virtual yang merespons dengan lebih manusiawi terhadap emosi pengguna, mengurangi frustrasi dan meningkatkan kepuasan. Selain itu, dalam keamanan dan forensik, deteksi emosi dapat mendeteksi tanda-tanda stres atau kebohongan, berkontribusi pada pengambilan keputusan yang lebih akurat. Namun, mengenali emosi secara manual atau otomatis menghadapi berbagai permasalahan dan kendala (Februari et al., 2025).

Secara umum, emosi bersifat subjektif dan kontekstual, sehingga sulit untuk distandarisasi. Kendala utama dalam deteksi emosi meliputi variasi individu, seperti perbedaan budaya dalam ekspresi emosi misalnya, emosi marah mungkin diekspresikan berbeda antara budaya Timur dan Barat. Selain itu, faktor lingkungan seperti pencahayaan buruk, *noise* suara, serta sudut pengambilan gambar yang tidak ideal dapat menurunkan akurasi deteksi (Budiono & Masing, 2022). Dari sisi teknis, sistem unimodal tidak selalu menangkap nuansa emosi kompleks secara optimal, seperti ketika ekspresi wajah netral disertai nada suara marah, sehingga berpotensi menimbulkan kesalahan interpretasi. Permasalahan ini diperburuk oleh bias dalam dataset, di mana data pelatihan tidak selalu merepresentasikan keberagaman populasi, menyebabkan model kurang mampu melakukan generalisasi(Kamal, 2025). Secara etis, deteksi emosi menimbulkan risiko privasi, karena analisis wajah atau suara dapat dianggap invasif apabila dilakukan tanpa persetujuan yang jelas. Oleh karena itu, pengembangan sistem deteksi emosi perlu mempertimbangkan aspek etika, perlindungan data, serta dampak sosial dan psikologis bagi pengguna (Razaqa et al., 2024).

Sebelum berkembangnya teknologi dan kecerdasan buatan (*Artificial Intelligence*/AI), pengenalan emosi manusia dilakukan secara manual melalui observasi subjektif oleh ahli psikologi atau individu terlatih. Metode tradisional ini melibatkan analisis ekspresi wajah, nada suara, bahasa tubuh, serta konteks situasional, seperti yang diperkenalkan oleh Paul Ekman melalui teori ekspresi wajah universal pada tahun 1970-an. Dalam praktiknya, ahli psikologi menggunakan pendekatan seperti *Facial Action Coding System* (FACS) untuk mengodekan pergerakan otot wajah, sementara analisis suara dilakukan dengan mengamati parameter akustik seperti pitch dan intensitas (Guntoro et al., 2022). Namun, pendekatan ini sangat bergantung pada interpretasi manusia, sehingga rentan terhadap bias subjektif, kelelahan, serta inkonsistensi hasil. Sebagai contoh, emosi marah dapat salah diidentifikasi sebagai senang apabila konteks budaya atau situasional tidak diperhitungkan. Dalam bidang klinis, metode wawancara dan tes psikologis yang digunakan untuk menilai emosi juga memerlukan waktu yang lama dan sulit diterapkan secara masif (Li et al., 2022). Meskipun menekankan aspek empati manusia, pendekatan tradisional ini kurang efisien untuk kebutuhan sistem *real-time*. Oleh karena itu, perkembangan teknologi komputer dan AI mendorong otomatisasi pengenalan emosi berbasis data, yang menjadi dasar bagi pengembangan sistem *deteksi* emosi modern, termasuk pendekatan berbasis *deep learning* dan *multimodal* (Wu et al., n.d.).

Dalam beberapa tahun terakhir, penelitian *deteksi* emosi berbasis *deep learning* berkembang pesat, khususnya pada pendekatan *multimodal* yang mengintegrasikan informasi suara, ekspresi wajah, dan modalitas visual lainnya untuk meningkatkan akurasi dan ketahanan sistem dibandingkan pendekatan unimodal. Studi survei terkini menunjukkan bahwa penggabungan fitur audio dan visual mampu merepresentasikan emosi manusia secara lebih komprehensif serta menghasilkan performa yang lebih unggul dalam mengenali emosi kompleks dibandingkan penggunaan satu modalitas saja(Kamal, 2025). Berbagai penelitian terbaru telah mengusulkan strategi *fusion* dan arsitektur *deep learning* seperti CNN dan LSTM untuk sistem pengenalan emosi multimodal, yang terbukti meningkatkan akurasi pada dataset *benchmark* seperti RAVDESS, SAVEE, dan CREMA-D (Boitel et al., 2025,Islam et al., 2024, Almulla, 2025). Pendekatan ini juga menunjukkan ketahanan yang lebih baik terhadap gangguan lingkungan nyata, yang sering menjadi kelemahan pada sistem unimodal (Cai et al., 2021)(Mehendale, 2020). Namun demikian, penelitian mutakhir masih menghadapi tantangan berupa keterbatasan dataset multimodal, kesulitan sinkronisasi antar modalitas, serta kebutuhan akan model yang efisien untuk aplikasi *real-time* (Liu et al., 2025). Oleh karena itu, pengembangan teknik *fusion*, *attention mechanism*, dan arsitektur ringan terus dilakukan untuk mendukung penerapan sistem deteksi emosi di kondisi dunia nyata.

Sebelum membahas pendekatan multimodal, penting untuk meninjau beberapa penelitian terkait deteksi emosi unimodal beserta keterbatasannya. Penelitian terdahulu mengusulkan pengenalan emosi berbasis suara menggunakan arsitektur BiLSTM dan menunjukkan bahwa sinyal audio mampu merepresentasikan dinamika emosi seperti marah dan senang dengan cukup baik(N.A., 2025). Namun, pendekatan berbasis suara sangat bergantung pada kualitas sinyal audio dan rentan terhadap gangguan *noise* lingkungan serta variasi karakteristik pembicara, sehingga performa model dapat menurun secara signifikan dalam kondisi nyata (Shane et al., 2025). Di sisi lain, penelitian sebelumnya memanfaatkan arsitektur MobileNetV2 untuk pengenalan emosi berbasis ekspresi wajah dan menunjukkan efisiensi komputasi yang tinggi(Dhimas et al., 2024). Meskipun demikian, metode berbasis wajah menghadapi kendala berupa sensitivitas terhadap pose non-frontal, pencahayaan yang tidak ideal, serta keterbatasan dataset yang umumnya dikumpulkan dalam kondisi terkontrol, sehingga mengurangi kemampuan generalisasi model. Keterbatasan pada masing-masing pendekatan unimodal ini menunjukkan bahwa emosi manusia tidak dapat direpresentasikan secara utuh melalui satu modalitas saja(Syaqialloh, 2025). Oleh karena itu, pendekatan deteksi emosi multimodal yang mengintegrasikan informasi audio dan visual menjadi alternatif yang lebih andal, karena memungkinkan setiap

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

modalitas saling melengkapi untuk menghasilkan prediksi emosi yang lebih akurat, stabil, dan mendekati kondisi nyata (Hadi & Sari, 2025).

Dalam penelitian ini, dikembangkan sistem deteksi emosi multimodal yang mengintegrasikan informasi dari suara dan ekspresi wajah untuk mengenali tiga kategori emosi, yaitu netral, senang, dan marah. Modalitas audio menggunakan dataset *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) yang telah banyak digunakan sebagai *benchmark* dalam penelitian pengenalan emosi(Khubrani, 2026). Dari dataset tersebut dipilih tiga emosi yang menjadi fokus penelitian, yaitu netral, senang dan marah. Dataset audio asli terdiri dari 480 data suara yang berasal dari 24 aktor. Untuk meningkatkan jumlah data pelatihan dan mengurangi resiko overfitting, dilakukan proses augmentasi pada data pelatihan sehingga jumlah data audio meningkat menjadi 1.120 data, yang terdiri atas 960 data pelatihan (train), 80 data validasi (validation), dan 80 data pengujian (test). Sementara itu, modalitas visual menggunakan dataset yang dikumpulkan secara mandiri melalui proses perekaman video ekspresi wajah menggunakan kamera telepon seluler. Dataset awal terdiri dari 80 video yang merepresentasikan tiga kategori emosi, yaitu netral, senang, dan marah. Selanjutnya, video diektraksi menjadi frame citra menggunakan *Roboflow* dan diberi label sesuai kategori emosi yang diteliti. Hasil proses ekstraksi menghasilakn citra, yang terdiri atas 5.037 citra pelatihan (train), 491 citra validasi (validation), dan 249 citra data pengujian (test).

Pada Penelitian terdahulu modalitas audio, sinyal suara terlebih dahulu dikonversi menjadi *representasi Mel-Spectogram*, kemudian diproses menggunakan kombinasi *Convolutional Neural Network* (CNN), *Bidirectional Long Short-Term Memory* (BiLSTM), dan *Attention Mechanism* (Wijaya, 2024). CNN digunakan untuk mengekstraksi pola spasial pada *Mel-Spectogram*, BiLSTM digunakan untuk mempelajari hubungan temporal antar fitur, sedangkan Attention Mechanism berfungsi untuk memberikan focus pada bagian sinyal yang paling relevan terhadap emosi. Pada modalitas visual, ekspresi wajah dianalisis menggunakan arsitektur *MobileNetV2* yang dikombinasikan dengan *MediaPipe Face Detection* untuk mendukung *deteksi* wajah secara efisien dan *real-time* (Ardiansyah et al., 2024). Selanjutnya, hasil prediksi dari dua modalitas diintegrasikan menggunakan pendekatan multimodal fusion untuk menghasilkan keputusan emosi yang lebih akurat dibandingkan penggunaan satu modalitas saja. Dengan memanfaatkan kombinasi CNN-BiLSTM-Attention pada audio dan *MobileNetV2* pada visual (Caldwell & Beaulieu, 2024). Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi ilmiah dalam pengembangan sistem deteksi emosi yang lebih akurat, inklusif, dan aplikatif pada bidang kesehatan dan pendidikan, serta menjadi dasar bagi pengembangan kecerdasan buatan yang lebih empatik dan bertanggung jawab di masa depan (Singla et al., 2024).

Berdasarkan penelitian terdahulu, sebagian besar penelitian deteksi emosi masih berfokus pada pendekatan unimodal, baik menggunakan ekspresi wajah maupun sinyal suara secara terpisah. Penelitian yang menggunakan CNN pada citra wajah menunjukkan hasil yang baik, namun performanya dapat menurun akibat perubahan pencahayaan dan posisi wajah. Sementara itu, penelitian berbasis audio menggunakan LSTM atau BiLSTM mampu mengenali pola emosi dari suara, tetapi rentan terhadap gangguan kebisingan (*noise*) (Nurjihan et al., 2024; Sutarti & Fariza Syaqialloh, 2025). Beberapa penelitian multimodal telah dilakukan dengan menggabungkan data audio dan visual, namun sebagian besar masih menggunakan arsitektur konvensional tanpa mekanisme *attention* serta belum mengoptimalkan proses penggabungan keputusan menggunakan metode *Late Fusion* (Riyandi & Sumarsono, 2025; Roihan et al., 2025). Oleh karena itu, penelitian ini mengusulkan implementasi sistem *Multimodal Emotion Recognition* yang mengintegrasikan CNN-BiLSTM-Attention pada modalitas audio dan MobileNetV2 pada modalitas visual dengan pendekatan *Late Fusion*. Pendekatan ini diharapkan mampu meningkatkan akurasi dan stabilitas deteksi emosi dibandingkan metode yang telah digunakan pada penelitian sebelumnya.

Berdasarkan uraian tersebut, penelitian ini memiliki kontribusi pada pengembangan bidang *Multimodal Emotion Recognition* melalui integrasi model CNN-BiLSTM-Attention untuk analisis emosi berbasis audio dan MobileNetV2 untuk analisis emosi berbasis ekspresi wajah. Selain itu, penelitian ini menerapkan metode *Late Fusion* dengan pembobotan yang disesuaikan berdasarkan performa masing-masing modalitas sehingga menghasilkan sistem deteksi emosi yang lebih akurat dan stabil. Kontribusi lainnya adalah implementasi sistem deteksi emosi secara real-time menggunakan kamera dan mikrofon yang berpotensi diterapkan pada bidang pendidikan, layanan pelanggan, dan sistem interaksi manusia-komputer.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian, Metode Penelitian, dan Studi Literatur

Berdasarkan metodologi yang telah dirancang, penelitian ini menggunakan metode eksperimen dengan pendekatan kuantitatif untuk mengembangkan sistem *Multimodal Emotion Recognition* berbasis *deep learning*. Tahapan penelitian dilakukan secara sistematis, dimulai dari studi literatur, pengumpulan data audio dan video, *preprocessing* data, perancangan model, pelatihan model, hingga evaluasi sistem. Studi literatur berperan sebagai landasan teoritis dalam memahami konsep emosi, metode deteksi emosi, serta pemanfaatan teknologi kecerdasan buatan, khususnya *Convolutional Neural Network* (CNN) dan *BiLSTM*, yang digunakan sebagai dasar dalam perancangan sistem.

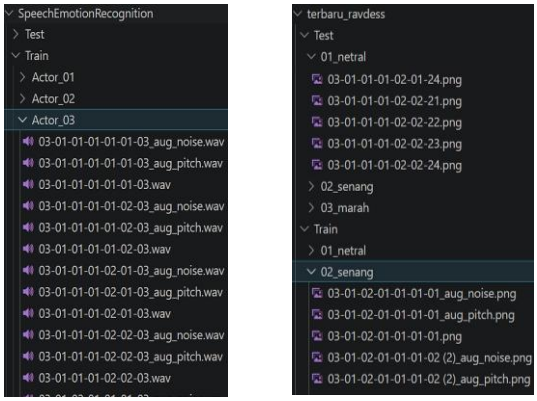
Penelitian ini mengintegrasikan model MobileNet untuk ekstraksi fitur visual dari ekspresi wajah dan BiLSTM untuk menganalisis pola temporal pada data audio. Kedua modalitas tersebut digabungkan melalui teknik *feature fusion* untuk meningkatkan akurasi klasifikasi emosi. Evaluasi sistem dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* guna mengukur performa model secara objektif. Dengan tahapan dan metode yang diterapkan,

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

penelitian ini diharapkan mampu menghasilkan sistem deteksi emosi multimodal yang lebih akurat, efektif, dan optimal dalam mengenali emosi berdasarkan kombinasi ekspresi wajah dan karakteristik suara.

2.2 Pengumpulan Dataset

Data yang di kumpulkan pada penelitian ini berupa data audio dan video, jumlah data yang dikumpulkan sebanyak 480 data kemudian di agumnetasi menjadi 1120 audio dan 80 data video mentah yang di ekstrak menjadi 5.777 frame data video. Proses pengumpulan data dilakukan menggunakan perangkat telepon genggam. Dataset penelitian ini dibuat untuk 3 emosi, yaitu “netral”, “senang”, “marah”.

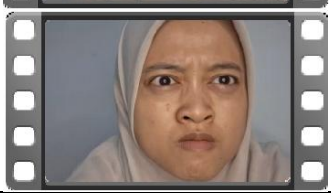


(a) (b)

Gambar 1. Dataset Audio (a) dataset awal, (b) dataset mel-spectogram

Gambar 1 merupakan dataset audio pada penelitian ini menggunakan dataset Ryerson Audio-Visual *Database of Emonional Speech and Song* (RAVDESS) yang telah banyak digunakan sebagai *benchmark* dalam penelitian pengenalan emosi. Dari dataset tersebut dipilih tiga emosi yang menjadi fokus penelitian, yaitu “netral”, “senang”, dan “marah”. Seluruh data audio disimpan dalam format WAV agar kompatibel dengan proses pengolahan sinyal audio. Dataset audio ini selanjutnya digunakan sebagai masukan pada tahap pra-pemrosesan dan pelatihan model deep learning untuk modalitas suara dalam sistem deteksi emosi multimodal.

Tabel 1. Kelas Emosi

Gambar	Kelas
	Netral
	Senang
	Marah

Beradasarkan Tabel 1 bahwa dataset video pada penelitian ini dikumpulkan secara mandiri oleh peneliti untuk merepresentasikan 3 kelas emosi, yaitu “netral”, “senang”, dan “marah”. Proses pengambilan data dilakukan dengan merekam video ekspresi wajah subjek menggunakan perangkat telepon genggam. Setiap video direkam dengan menampilkan ekspresi wajah yang sesuai dengan emosi yang diteliti, sehingga dapat menangkap karakteristik visual emosi secara alami. Video hasil perekaman selanjutnya diunggah ke platform Roboflow untuk dilakukan proses ekstraksi frame dan pelabelan data. Setiap satu video diekstraksi menjadi 5 frame untuk meningkatkan jumlah data citra wajah. Frame yang dihasilkan kemudian diberi label sesuai dengan kelas emosi yang direpresentasikan. Seluruh citra wajah hasil pelabelan digunakan sebagai dataset video yang selanjutnya diproses pada tahap pra-pemrosesan dan pelatihan model deep learning berbasis CNN dalam sistem deteksi emosi multimodal.

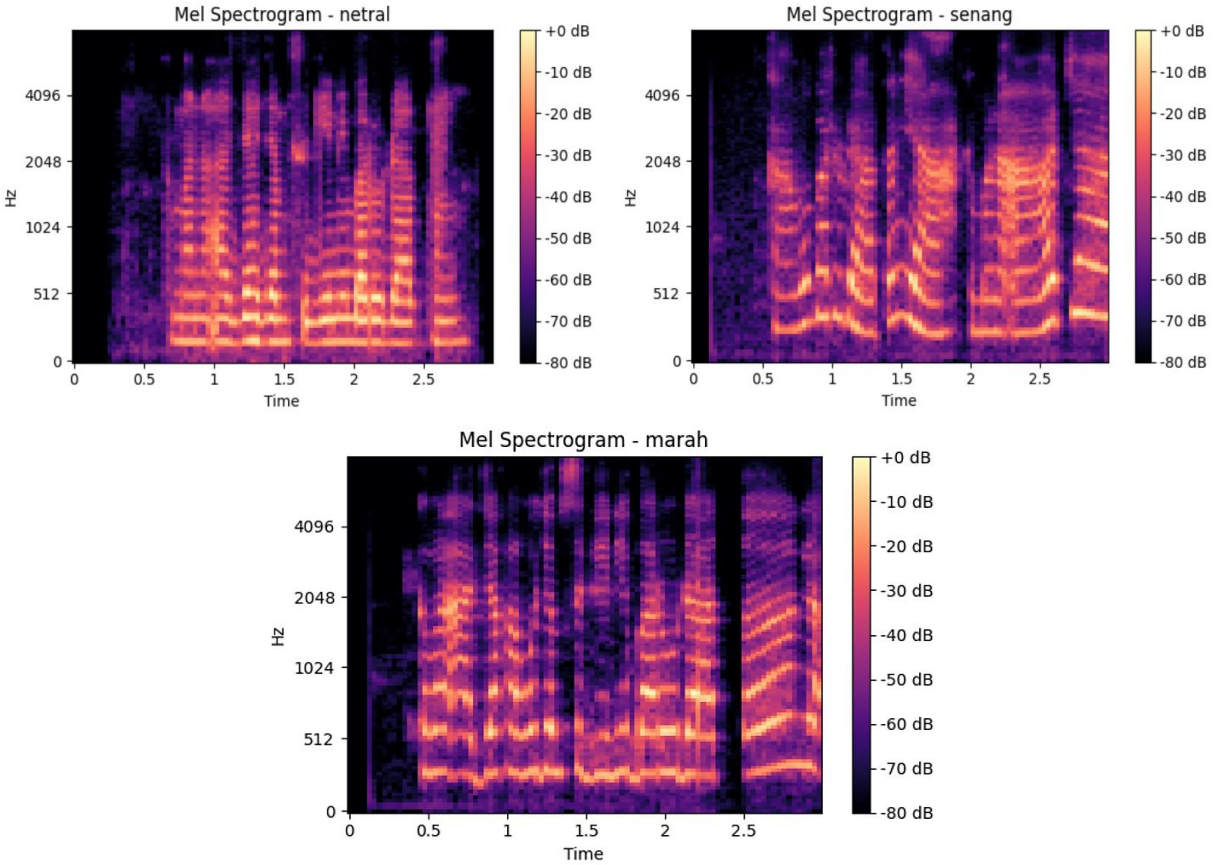
TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

2.3 Pra-Pemrosesan Data

Tahap pra-pemrosesan data merupakan tahapan penting dalam penelitian ini untuk memastikan data audio dan video yang digunakan memiliki kualitas yang baik serta sesuai dengan kebutuhan masukan model deep learning. Pra-pemrosesan dilakukan setelah proses pengumpulan dataset dan sebelum tahap pelatihan model, sebagaimana ditunjukkan pada flowchart tahapan penelitian. Proses ini bertujuan untuk mengurangi gangguan noise, menyeragamkan format data, serta mengekstraksi informasi yang relevan dari masing-masing modalitas.

2.3.1 Pra-Pemrosesan Data Audio

Dataset audio yang digunakan dalam penelitian ini berasal dari Ryerson Audio-Visual Database of Emonional Speech and Song (RAVDESS) yang merupakan salah satu dataset standar dalam penelitian pengenalan emosi berbasis suara. Penelitian ini memanfaatkan tiga kategori emosi, yaitu netral, senang, dan marah. Data audio dipilih berdasarkan kode emosi yang terdapat pada skema penamaan file RAVDESS, kemudian dilakukan proses pengelompokan ulang sehingga hanya data yang termasuk ke dalam tiga kategori emosi tersebut yang digunakan. Selanjutnya, penamaan label kelas disesuaikan dengan kategori emosi penelitian, yaitu (01) kode netral, (02) kode senang, dan (03) kode marah, untuk memudahkan proses preprocessing, pelatihan model, serta evaluasi sistem multimodal yang dikembangkan. Setelah itu, data audio diekstraksi menjadi fitur Mel-Spectrogram sebagai representasi karakteristik frekuensi dan temporal sinyal suara yang berkaitan dengan emosi. Adapun hasil dari ekstraksi fitur Mel-Spectrogram seperti gambar di bawah ini:



(c)

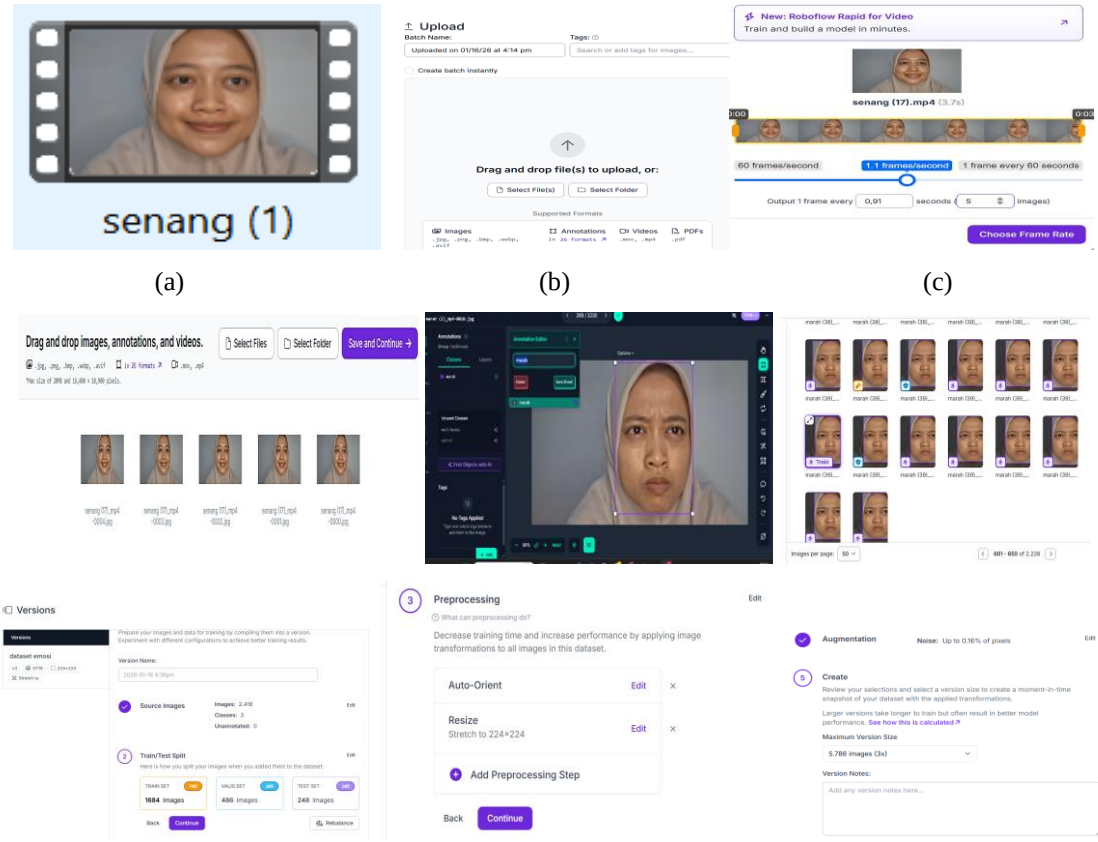
Gambar 2. Mel-Spectrogram 3 Emosi

Gambar 2 merupakan proses ekstraksi ini bertujuan untuk mengubah sinyal audio satu dimensi menjadi representasi dua dimensi yang dapat diproses oleh Convolutional Neural Network (CNN). Hasil fitur kemudian dinormalisasi untuk menyeragamkan rentang nilai data, sehingga dapat meningkatkan stabilitas dan efisiensi proses pelatihan model CNN–BiLSTM.

2.3.2 Pra-Pemrosesan Data Video

Data video hasil perekaman ekspresi wajah selanjutnya diproses menggunakan platform Roboflow. Pada tahap ini, setiap video diekstraksi menjadi beberapa frame secara periodik, di mana satu video diambil sebanyak lima frame untuk merepresentasikan ekspresi wajah subjek. Frame yang dihasilkan kemudian diberi label sesuai dengan kelas emosi, yaitu netral, senang, dan marah. Adapun proses pra-pemrosesan data video seperti pada Gambar 3. berikut ini:

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629



Gambar 3. pra-pemrosesan data video : (a) video, (b) Upload roboflow, (c) Ektraksi video ke Frame, (d) Hasil Frame, (e) Labeling, (f) Hasil Labeling, (g) Pembagian dataset, (h) Resize, (i) Augmentasi & Create Maximum Version Size.

Selain proses pelabelan, dilakukan pula augmentasi data pada citra wajah untuk meningkatkan variasi dataset dan mengurangi risiko overfitting. Selanjutnya, seluruh citra wajah diubah ukurannya menjadi 224 × 224 piksel agar sesuai dengan spesifikasi masukan model CNN berbasis MobileNet. Normalisasi nilai piksel juga diterapkan untuk menjaga konsistensi data dan meningkatkan performa model dalam mengenali pola ekspresi wajah.

2.3.3 Pembagian Dataset Audio

Dataset audio dari tiga kelas emosi ini berjumlah 480 file suara dari 24 aktor, setiap aktor memiliki 20 data audio yaitu terdiri dari 4 audio netral, 8 audio senang, dan 8 audio marah. Kemudian dataset dibagi menjadi tiga subset, yaitu data pelatihan sekitar 16 aktor (67%), data validasi 4 aktor (16.5%), dan data pengujian 4 aktor (16.5%). Pembagian dilakukan secara proporsional pada setiap kelas emosi untuk menjaga keseimbangan distribusi data. Dataset ini kemudian digunakan sebagai input pada tahap pelatihan dan evaluasi model deep learning. Pembagian Dataset dapat di lihat pada Tabel 2.

Tabel 2. Pembagian Dataset Audio

Dataset	Jumlah
Train	320
Validasi	80
Test	80
Total	480

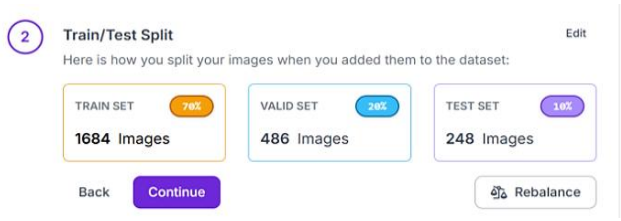
Setelah pembagian dataset, dilakukan augmentasi pada data train yang berjumlah 320 audio menjadi 960 audio train. Data augmentasi terdiri dari audio normal, noise, dan pitch, jadi jumlah keseluruhan dataset audio menjadi 1120 data audio.

2.3.4 Pembagian Dataset Video

Dataset video yang terdiri dari tiga kelas emosi ini berjumlah 2418 frame citra dan setaip kelasnya memiliki sampel 823 netral, 822 senang, dan 774 frame citra marah. Dataset ini telah melalui tahap pra-pemrosesan selanjutnya dibagi menjadi tiga bagian, yaitu data pelatihan (*training set*), data validasi (*validation set*), dan data pengujian (*testing set*). Pembagian dataset dilakukan secara proporsional dengan rasio 70% data pelatihan, 20% data validasi, dan 10% data

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

pengujian, sebagaimana ditetapkan pada proses pengelolaan dataset menggunakan Roboflow. Pembagian dataset bisa di lihat pada Gambar 5.



Gambar 4. Pembagian Dataset Video

Setelah pembagian dataset tersebut, data pelatihan ini kemudian di augmentasi dan jumlah data pelatihan bertambah menjadi 5037 frame citra, digunakan untuk melatih model MobileNetV2 dalam mempelajari pola ekspresi wajah pada masing-masing kelas emosi. Data validasi yang berjumlah 486 frame citra dimanfaatkan untuk memantau kinerja model selama proses pelatihan serta membantu dalam penyesuaian parameter guna mencegah terjadinya *overfitting*. Sementara itu, data pengujian yang berjumlah 248 frame citra digunakan untuk mengevaluasi kinerja akhir sistem secara objektif dalam mengenali emosi berdasarkan ekspresi wajah.

2.4 Merancang dan Melatih Model

Model yang digunakan dalam penelitian ini terdiri dari dua arsitektur utama, yaitu model video dan model audio. Model video memanfaatkan arsitektur Convolutional Neural Network (CNN) berbasis transfer learning menggunakan MobileNetV2, sedangkan model audio menggunakan kombinasi CNN dan Bidirectional LSTM untuk menangkap pola spasial dan temporal dari sinyal suara. Pendekatan multimodal ini dirancang untuk mengenali emosi berdasarkan ekspresi wajah (video) dan karakteristik suara (audio), dengan tiga kelas emosi yaitu Netral, Senang, dan Marah. Model dilatih menggunakan data yang telah melalui tahap preprocessing dan ekstraksi fitur. Tabel 2 berikut menunjukkan rincian parameter model yang digunakan.

Tabel 3. Rincian Model yang Digunakan

Parameter	Model Video	Model Audio
Hidden Layer	GlobalAveragePooling + Dense (128/256) + Dropout	Conv2D (32,64) + Bidirectional LSTM (64) + Attention + Dense (64)
Neorun	128–256, 3 (output)	64 (LSTM), 64 (Dense), 3 (output)
optimizer	Adam	Adam
loss	categorical_crossentropy	categorical_crossentropy
Metrics	accuracy	accuracy
Fungsi Aktivasi	ReLU dan Softmax	ReLU, Tanh, dan Softmax
Epoch	25	25

Berdasarkan Tabel 3, model video menggunakan arsitektur MobileNetV2 sebagai feature extractor untuk menangkap pola spasial pada citra wajah. Lapisan GlobalAveragePooling digunakan untuk mereduksi dimensi fitur sebelum diteruskan ke layer Dense. Fungsi aktivasi ReLU digunakan pada hidden layer untuk mempercepat proses konvergensi dan mengurangi risiko vanishing gradient, sedangkan Softmax pada layer output berfungsi menghasilkan probabilitas klasifikasi pada tiga kelas emosi.

Sementara itu, model audio mengombinasikan CNN untuk ekstraksi fitur spektral dan Bidirectional LSTM untuk memahami pola temporal dari sinyal suara. Fungsi aktivasi Tanh digunakan pada LSTM karena mampu memetakan nilai ke rentang −1 hingga 1, sehingga efektif dalam merepresentasikan dinamika emosi yang memiliki variasi positif dan negatif. Layer output menggunakan Softmax untuk menghasilkan probabilitas prediksi setiap kelas emosi. Optimizer Adam dipilih karena memiliki kemampuan adaptif dalam mengatur learning rate sehingga mempercepat proses pelatihan dan menghasilkan konvergensi yang stabil. Loss function categorical_crossentropy digunakan karena penelitian ini merupakan masalah klasifikasi multi-kelas dengan representasi output berbentuk probabilitas. Pendekatan multimodal dipilih karena kombinasi informasi visual dan audio dapat meningkatkan akurasi sistem dibandingkan penggunaan satu modalitas saja. Dengan arsitektur yang dirancang secara spesifik untuk masing-masing jenis data, model diharapkan mampu mengenali emosi secara lebih robust dan akurat dalam skenario real-time.

2.5 Integrasi Multimodal Fusion

Integrasi multimodal pada penelitian ini dilakukan dengan menggabungkan informasi emosi dari dua sumber data, yaitu audio dan visual, agar sistem dapat mengenali emosi dengan lebih akurat dibandingkan jika hanya menggunakan satu modalitas. Modalitas audio diproses menggunakan model CNN-BiLSTM-Attention dengan masukan berupa Mel-Spectrogram, sedangkan modalitas visual diproses menggunakan MobileNetV2 berdasarkan ekspresi wajah yang terdeteksi. Kedua model menghasilkan nilai probabilitas untuk tiga kelas emosi, yaitu netral, senang, dan marah. Probabilitas tersebut kemudian digunakan sebagai masukan dalam proses integrasi menggunakan metode *Late Fusion*.

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

Metode *Late Fusion* dilakukan pada tingkat keputusan (*decision level*) dengan menggabungkan hasil prediksi dari model audio dan visual menggunakan pembobotan. Dalam penelitian ini, bobot sebesar 0,4 diberikan pada modalitas audio dan 0,6 pada modalitas visual karena model visual menunjukkan performa yang lebih stabil. Hasil penggabungan probabilitas dari kedua modalitas menghasilkan probabilitas akhir untuk setiap kelas emosi, kemudian sistem memilih kelas dengan nilai probabilitas tertinggi sebagai prediksi akhir. Melalui mekanisme ini, sistem mampu memanfaatkan keunggulan masing-masing modalitas sehingga menghasilkan deteksi emosi yang lebih akurat, konsisten, dan tetap efektif dalam kondisi real-time meskipun terdapat gangguan seperti *noise* pada audio atau perubahan pencahayaan pada wajah.

2.6 Evaluasi Model dan Perangkat Penelitian

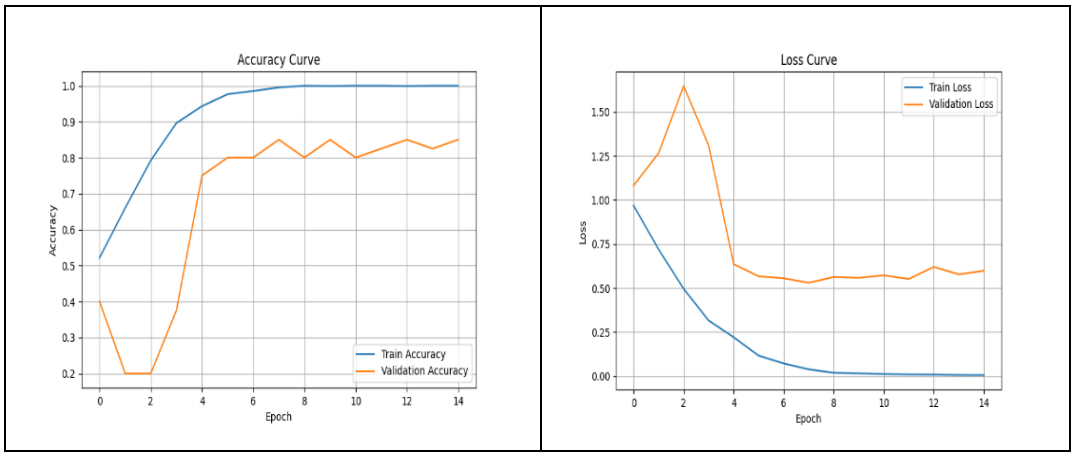
Evaluasi model dilakukan untuk mengukur performa sistem dalam mengklasifikasikan emosi ke dalam tiga kelas, yaitu Netral, Senang, dan Marah. Proses evaluasi dilakukan baik pada model video, model audio, maupun hasil fusion multimodal. Metode evaluasi yang digunakan meliputi Confusion Matrix dan perhitungan F1-Score. Perangkat penelitian merupakan sarana pendukung yang digunakan dalam proses pengembangan dan pengujian sistem klasifikasi emosi berbasis multimodal (video dan audio). Perangkat penelitian pada penelitian ini terdiri dari perangkat keras dan perangkat lunak yang mendukung proses pengumpulan data, preprocessing, pelatihan model, hingga implementasi sistem berbasis web secara real-time.

3. HASIL DAN PEMBAHASAN

3.1 Training Model

3.1.1 Training Model Audio

Model audio dilatih menggunakan dataset RAVDESS yang terdiri dari tiga kelas emosi yaitu netral, senang, dan marah dengan 25 *epoch* menggunakan CNN, Bilstm dan *Attention*. Data latih yang digunakan sebanyak 960 *mel-spectogram* yang di pilih secara acak, sedangkan untuk 80 *mel-spectogram* digunakan sebagai data validasi untuk mengevaluasi kinerja model selama proses berlangsungnya training. Adapun hasil training ditampilkan pada Gambar 5. yang menunjukkan nilai akurasi pada data training dan data validasi.



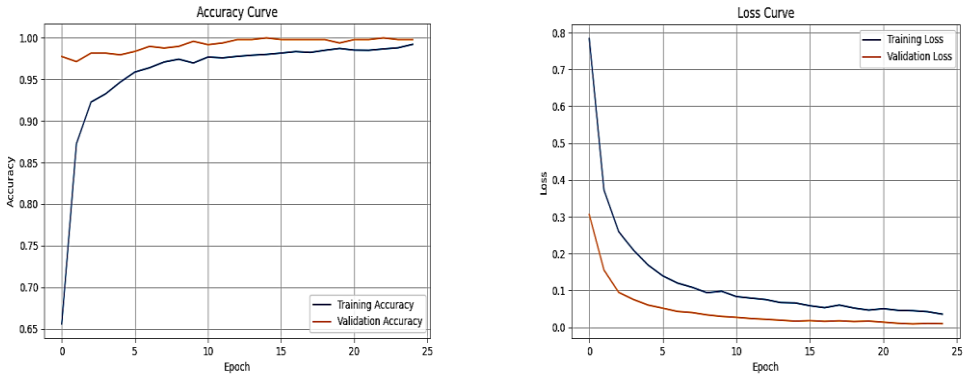
Gambar 5. Grafik akurasi audio

Berdasarkan Gambar 5, proses pelatihan dilakukan selama 25 *epoch* dengan menggunakan *optimizer Adam* dan *learning rate* sebesar 0,0001. Berdasarkan hasil training, nilai akurasi *training* mengalami peningkatan secara bertahap dari 68,27% pada *epoch* pertama menjadi 98,71% pada *epoch* ke-25. Pada saat yang sama, nilai *loss* mengalami penurunan dari 0,7358 menjadi 0,0412. Selain itu, akurasi validasi meningkat secara signifikan dari 97,35% pada *epoch* pertama hingga mencapai nilai tertinggi sebesar 100% pada *epoch* ke-11 dan *epoch* ke-23. Nilai validation loss juga terus menurun hingga mencapai 0,0115 pada akhir pelatihan. Hasil ini menunjukkan bahwa model mampu mempelajari pola emosi dari data suara dengan baik dan memiliki kemampuan generalisasi yang tinggi terhadap data validasi.

3.1.2 Training Model Visual

Model visual dilatih menggunakan arsitektur *MobileNetV2* untuk melakukan klasifikasi wajah dengan 25 *epoch.* Data latih yang digunakan sebanyak 5037 *frame* citra yang dipilih secara acak. Sedangkan untuk 486 *frame* citra digunakan sebagai data validasi untuk mengevaluasi kinerja model selama proses *training*. Hasil *training* dan validasi ditampilkan pada Gambar 6 yang menunjukan nilai akurasi visual.

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629



Gambar 6. Grafik akurasi visual

Gambar 6 menunjukkan hasil proses pelatihan model *MobileNetV2* untuk mengklasifikasikan emosi wajah dalam tiga kelas; netral, bahagia, dan marah. Melihat kurva akurasi ini, kita dapat melihat bahwa nilai akurasi pelatihan melonjak dari kurang dari 65% pada *epoch* pertama menjadi hampir 99% pada *epoch* ke-25. Di sisi lain, akurasi validasi tetap berada pada nilai yang tinggi sejak sangat awal pelatihan dan seiring berjalannya pelatihan, akurasi validasi terus meningkat hingga mencapai hampir 100%. Kondisi ini menunjukkan bahwa model mampu mempelajari fitur-fitur penting dari citra ekspresi wajah yang menghasilkan kinerja kuat pada data validasi. Pada *kurva loss*, terlihat bahwa nilai loss pelatihan secara bertahap menurun dari sekitar 0,79 menjadi sekitar 0,04 pada akhir pelatihan. Hal yang sama terjadi pada nilai *loss validasi*, yang menurun dari sekitar 0,31 menjadi hampir 0,01. Penurunan nilai *loss* menandakan bahwa kesalahan prediksi model semakin kecil selama proses pembelajaran. Menurut teori *deep learning*, ketika sebuah model menunjukkan peningkatan akurasi yang stabil dan penurunan nilai kerugian, hal ini menandakan bahwa proses optimasi berjalan dengan baik, sehingga parameter model dapat menemukan representasi fitur yang semakin baik.

Selain itu, jarak antara kurva pelatihan dan kurva validasi relatif kecil, dan tidak terdapat pola divergensi yang signifikan. Bahkan pada beberapa *epoch*, akurasi validasi sedikit lebih tinggi daripada akurasi pelatihan. Situasi ini menunjukkan bahwa model tidak mengalami *overfitting* dan dapat berkinerja baik pada data yang belum pernah dilihat sebelumnya. Hasil penelitian menunjukkan bahwa penggunaan *MobileNetV2* untuk pembelajaran transfer efektif dalam mengekstraksi fitur ekspresi wajah dari dataset penelitian, sehingga memungkinkan pembuatan model klasifikasi emosi visual dengan akurasi yang sangat tinggi dan stabil.

Berdasarkan hasil pelatihan yang ditunjukkan pada Gambar 4.2, model *MobileNetV2* mencapai akurasi validasi mendekati 100% dengan nilai loss yang sangat rendah. Hasil ini menunjukkan bahwa model tersebut sangat baik dalam mengenali ekspresi wajah netral, bahagia, dan marah, sehingga menjadikannya komponen visual yang sesuai untuk sistem deteksi emosi *multimodal* yang dikembangkan dalam penelitian ini.

3.2 Pengujian Model

Pengujian model pada penelitian ini dilakukan dengan dua cara, yaitu menguji model menggunakan data uji (Testing) dan menguji model secara realtime yang sudah digabungkan menggunakan late fusion.

3.2.1 Pengujian dengan Data *Testing* Audio

Pengujian model audio dijalankan untuk mengevaluasi kemampuan model *CNN-BiLSTM-Attention* dalam mengklasifikasikan emosi berdasarkan data suara yang telah direpresentasikan dalam bentuk *mel-spectrogram*. Pada riset ini dipakai tiga kategori emosi dari dataset RAVDESS, yaitu netral, senang, dan marah. Penilaian dilakukan memakai metrik *precision*, *recall*, *F1-score*, dan *accuracy* untuk mengetahui tingkat keberhasilan model dalam mengenali masing-masing kelas emosi. Temuan pengujian model audio ditampilkan pada Tabel 4.

Tabel 4. Klasifikasi Report Audio

keterangan	Precision	Recall	F1-Score	Support
Netral	0.67	0.88	0.76	16
Senang	0.62	0.66	0.64	32
Marah	0.64	0.50	0.56	32
Accuracy	-	-	0.64	80
Macro avg	0.64	0.68	0.65	80
Weighted avg	0.64	0.64	0.63	80

Berdasarkan Tabel 4, model *CNN-BiLSTM-Attention* memperlihatkan kemampuan yang cukup baik dalam mengenali emosi dari data suara. Kelas netral memperoleh nilai *precision* sebesar 0,67, *recall* sebesar 0,88, dan *F1-score* sebesar 0,76. Nilai *recall* yang tinggi menunjukkan bahwa sebagian besar data suara dengan emosi netral berhasil dikenali dengan benar oleh model. Pada kelas senang, model memperoleh nilai *precision* sebesar 0,62, *recall* sebesar

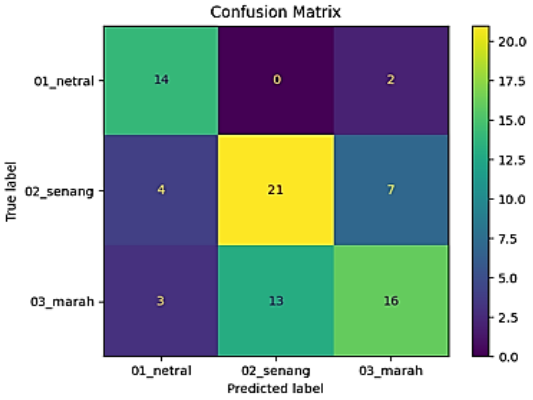
TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

0,66, dan *F1-score* sebesar 0,64. Temuan tersebut menunjukkan bahwa model mampu mengenali emosi senang dengan cukup baik, meskipun masih dijumpai beberapa kesalahan klasifikasi pada kelas tersebut.

Sementara itu, kelas marah memperoleh nilai *precision* sebesar 0,64, *recall* sebesar 0,50, dan *F1-score* sebesar 0,56. Nilai *recall* yang lebih rendah dibandingkan kelas lainnya memperlihatkan bahwa sebagian data suara dengan emosi marah masih sulit dibedakan oleh model dan cenderung diprediksi sebagai kelas emosi lain. Secara keseluruhan, model memperoleh nilai *accuracy* sebesar 63,75% pada data pengujian. Nilai *macro average* sebesar 0,65 dan *weighted average* sebesar 0,63 menunjukkan bahwa performa model relatif seimbang pada ketiga kelas emosi yang dipakai. Temuan ini menunjukkan bahwa model telah mampu mempelajari pola emosi dari sinyal suara, meskipun masih dijumpai kesalahan klasifikasi yang memengaruhi performa model pada data pengujian.

Selain temuan pengujian pada data *testing*, tahapan pelatihan model juga memperlihatkan nilai *train accuracy* sebesar 91,56%, *validation accuracy* sebesar 60,00%, dan *test accuracy* sebesar 63,75%. Perbedaan yang cukup besar antara akurasi pelatihan dan akurasi validasi menunjukkan bahwa model cenderung mempelajari pola pada data pelatihan dengan sangat baik, akan tetapi kemampuan generalisasinya terhadap data yang belum pernah dilihat masih terbatas. Kondisi ini mengindikasikan adanya kecenderungan *overfitting*, sehingga dibutuhkan analisis selanjutnya melalui *confusion matrix* untuk mengetahui distribusi kesalahan prediksi pada masing-masing kelas emosi.

Untuk memperoleh gambaran yang lebih rinci mengenai temuan klasifikasi pada setiap kategori emosi, dijalankan kajian memakai *confusion matrix*. Analisis ini dipakai untuk memperlihatkan jumlah data yang berhasil diprediksi dengan benar maupun data yang mengalami kesalahan klasifikasi pada masing-masing kelas emosi. Hasil *confusion matrix* model audio ditampilkan pada Gambar 7.



Gambar 7. *Confucion matrix* model audio

Berdasarkan Gambar 8, *confusion matrix* memperlihatkan temuan klasifikasi model *CNN-BiLSTM-Attention* pada tiga kelas emosi, yaitu netral, senang, dan marah. Nilai pada diagonal utama *confusion matrix* menunjukkan jumlah data yang berhasil diprediksi dengan benar, sedangkan nilai di luar diagonal menunjukkan jumlah data yang mengalami kesalahan klasifikasi. Secara keseluruhan, model menghasilkan 51 prediksi yang benar dari total 80 data uji, akibatnya memperoleh akurasi sebesar 63,75%.

Pada kelas netral, dari 16 data uji dijumpai 14 data yang berhasil diklasifikasikan dengan benar sebagai netral, sedangkan 2 data lainnya salah diprediksi sebagai marah. Temuan ini memperlihatkan bahwa model mempunyai kemampuan yang cukup baik dalam mengenali emosi netral dengan tingkat pengenalan yang relatif tinggi. Tidak terdapat data netral yang diprediksi sebagai senang, yang menunjukkan bahwa karakteristik emosi netral relatif mudah dibedakan dari emosi senang pada dataset yang dipakai.

Pada kelas senang, dari total 32 data uji dijumpai 21 data yang berhasil diprediksi dengan benar sebagai senang. Akan tetapi demikian, terdapat 4 data yang salah diklasifikasikan sebagai netral dan 7 data yang salah diklasifikasikan sebagai marah. Jumlah kesalahan yang cukup besar pada kelas ini memperlihatkan bahwa model masih mengalami kesulitan dalam membedakan karakteristik suara emosi senang dengan dua kelas lainnya. Kondisi ini mengakibatkan nilai *precision* dan *recall* pada kelas senang tidak setinggi kelas netral.

Sementara itu, pada kelas marah, dari 32 data uji dijumpai 16 data yang berhasil dikenali dengan benar sebagai marah. Sebanyak 13 data salah diprediksi sebagai senang dan 3 data lainnya diprediksi sebagai netral. Tingginya jumlah data marah yang diklasifikasikan sebagai senang memperlihatkan adanya kemiripan pola akustik tertentu antara kedua emosi tersebut pada dataset RAVDESS yang digunakan. Kondisi ini juga menguraikan mengapa nilai recall pada kelas marah relatif lebih rendah dibandingkan kelas lainnya.

Secara keseluruhan, *confusion matrix* memperlihatkan bahwa model mampu mengenali emosi netral dengan cukup baik, akan tetapi masih mengalami kesulitan dalam membedakan emosi senang dan marah. Kesalahan klasifikasi yang cukup sering terjadi antara kedua kelas tersebut mengindikasikan bahwa fitur *mel-spectrogram* yang dipakai belum sepenuhnya mampu merepresentasikan perbedaan karakteristik akustik kedua emosi secara terbaik. Meskipun demikian, model *CNN-BiLSTM-Attention* tetap mampu mempelajari pola emosi dari data suara dan membuahkan performa yang cukup baik sebagai modalitas audio dalam sistem *Multimodal Emotion Recognition* yang dikembangkan pada riset ini.

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

Perbedaan antara nilai train accuracy sebesar 91,56% dengan *test accuracy* sebesar 63,75% menunjukkan bahwa model masih memiliki kecenderungan *overfitting*. Kondisi ini mengindikasikan bahwa model mampu mempelajari pola pada data pelatihan dengan sangat baik, namun kemampuan generalisasinya terhadap data baru masih perlu ditingkatkan. Kondisi ini dapat terjadi karena jumlah data pelatihan yang *relative* terbatas serta kompleksitas model *CNN-BiLSTM-Attention* yang cukup tinggi dibandingkan jumlah data yang digunakan.

3.2.2 Pengujian dengan Data *Testing Visual*

Pengujian model visual dilakukan untuk mengevaluasi kemampuan *MobileNetV2* dalam mengklasifikasikan ekspresi wajah ke dalam tiga kategori emosi, yaitu netral, senang, marah. Evaluasi dilakukan menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*. Hasil pengujian model visual ditampilkan pada Tabel 5 berikut.

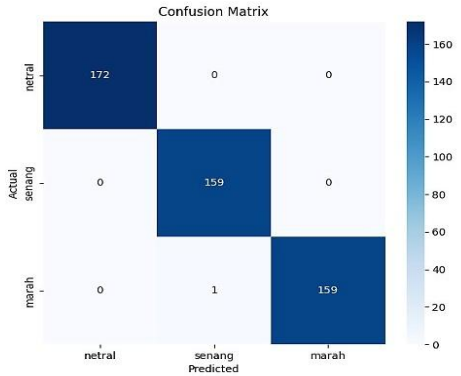
Tabel 5. Klasifikasi Report Visual

Keterangan	Precision	Recall	F1-Score	Suport
Netral	1.00	1.00	1.00	172
Senang	0.99	1.00	1.00	159
Marah	1.00	0.99	1.00	160
Accuracy	-	-	1.00	491
Macro avg	1.00	1.00	1.00	491
Weighted avg	1.00	1.00	1.00	491

Berdasarkan Tabel 5, model *MobileNetV2* memperlihatkan *performa* yang sangat baik pada seluruh kelas emosi yang diuji. Kelas netral memperoleh nilai *precision*, *recall*, dan *F1-score* sebesar 1,00, yang menunjukkan bahwa seluruh data wajah dengan ekspresi netral berhasil diklasifikasikan dengan benar tanpa kesalahan. Pada kelas senang, model memperoleh nilai *precision* sebesar 0,99, *recall* sebesar 1,00, dan *F1-score* sebesar 1,00. Temuan tersebut menunjukkan bahwa hampir seluruh data pada kelas senang dapat dikenali dengan sangat baik oleh model. Sementara itu, kelas marah memperoleh nilai *precision* sebesar 1,00, *recall* sebesar 0,99, dan *F1-score* sebesar 1,00. Nilai *recall* yang sedikit lebih rendah mengindikasikan adanya satu data marah yang diprediksi sebagai kelas lain, akan tetapi jumlah kesalahan tersebut sangat kecil dibandingkan keseluruhan data yang diuji akibatnya tidak menyajikan pengaruh yang bermakna terhadap performa model secara keseluruhan.

Secara umum, model membuahkan akurasi validasi sebesar 99,80% dengan *validation loss* sebesar 0,0086. Tingginya nilai akurasi memperlihatkan bahwa *MobileNetV2* mampu mengekstraksi fitur visual wajah secara berdaya guna untuk membedakan emosi netral, senang, dan marah. Di samping itu, nilai validation loss yang sangat rendah menunjukkan bahwa kesalahan prediksi model selama tahapan validasi relatif kecil. Nilai *macro average* dan *weighted average* yang sama-sama meraih 1,00 juga menunjukkan bahwa performa model konsisten pada seluruh kelas emosi dan tidak dipengaruhi oleh perbedaan jumlah data pada masing-masing kelas. Oleh sebab itu, dapat disimpulkan bahwa model *MobileNetV2* mempunyai kemampuan yang sangat baik dalam menjalankan klasifikasi ekspresi wajah dan layak dipakai sebagai modalitas visual dalam sistem *Multimodal Emotion Recognition* yang dikembangkan pada riset ini. Untuk melihat distribusi prediksi yang dihasilkan model secara lebih rinci, dijalankan kajian memakai *confusion matrix*. Temuan *confusion matrix* model visual ditunjukkan pada Gambar 8.

Berdasarkan Gambar 8, *confusion matrix* memperlihatkan bahwa model *MobileNetV2* mampu mengklasifikasikan sebagian besar data uji dengan sangat baik. Seluruh 172 data pada kelas netral berhasil diprediksi dengan benar sebagai netral, sedangkan seluruh 159 data pada kelas senang juga berhasil diklasifikasikan secara tepat. Pada kelas marah, dari total 160 data dijumpai 159 data yang berhasil dikenali dengan benar dan hanya 1 data yang salah diklasifikasikan sebagai kelas senang. Temuan tersebut menunjukkan bahwa model mempunyai kemampuan yang sangat tinggi dalam membedakan ketiga kategori emosi yang dipakai pada riset ini.



Gambar 8. *Confusion Matrix Visual*

Dominasi nilai pada diagonal utama *confusion matrix* memperlihatkan tingginya jumlah prediksi yang benar pada setiap kelas emosi. Sebaliknya, nilai di luar diagonal hampir seluruhnya bernilai nol yang mengindikasikan

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

rendahnya tingkat kesalahan klasifikasi. Secara keseluruhan, model membuahkan 490 prediksi benar dari total 491 data uji dengan akurasi sebesar 99,80%. Temuan ini memperlihatkan bahwa fitur-fitur ekspresi wajah yang diekstraksi oleh *MobileNetV2* mampu merepresentasikan karakteristik masing-masing emosi dengan baik akibatnya tahapan klasifikasi dapat dijalankan secara akurat.

Tingginya performa model tidak terlepas dari kemampuan *MobileNetV2* dalam mengekstraksi fitur visual secara efisien melalui cara kerja *Depthwise Separable Convolution* yang membuahkan jumlah tolok ukur lebih sedikit dibandingkan CNN konvensional. Di samping itu, penggunaan *transfer learning* dengan bobot awal dari *ImageNet* membantu model memperoleh representasi fitur visual yang lebih baik akibatnya mampu membedakan ekspresi netral, senang, dan marah secara berdaya guna. Meskipun demikian, temuan yang sangat tinggi ini tetap perlu dianalisis secara kritis karena dapat dipengaruhi oleh karakteristik dataset yang relatif terkontrol, jumlah kelas yang terbatas, serta kemiripan distribusi data pelatihan dan data validasi. Maka dari itu, pengujian pada data baru dengan kondisi lingkungan yang lebih beragam dibutuhkan untuk mengevaluasi kemampuan generalisasi model pada situasi nyata.

Dari hasil pengujian *classification report* dan *confusion matrix*, dapat disimpulkan bahwa model *MobileNetV2* mampu menjalankan klasifikasi ekspresi wajah dengan tingkat akurasi yang sangat tinggi. Maka dari itu, model visual yang dikembangkan layak dipakai sebagai salah satu modalitas utama dalam sistem *Multimodal Emotion Recognition* pada riset ini.

Berdasarkan hasil pengujian masing-masing modalitas, model visual memperlihatkan performa yang sangat tinggi dengan akurasi sebesar 99,80%, sedangkan model audio memperoleh akurasi sebesar 63,75%. Kedua model tersebut kemudian diintegrasikan memakai cara *late fusion* untuk membuahkan sistem *Multimodal Emotion Recognition* yang mampu memanfaatkan informasi dari suara dan ekspresi wajah secara bersamaan. Pengujian terhadap sistem *multimodal* dijalankan secara *real-time* untuk mengetahui kemampuan sistem dalam mendeteksi emosi pada kondisi penggunaan nyata.

3.2.3 Pengujian Multimodal *Real-Time*

Pengujian multimodal *real-time* dilakukan untuk mengevaluasi kemampuan sistem dalam mengenali emosi berdasarkan kombinasi informasi suara dan ekspresi wajah secara bersamaan. Pada tahap ini, model CNN-BiLSTM-Attention dan model visual MobileNetV2 diintegrasikan menggunakan metode *late fusion*. Pengujian dilakukan secara *real-time* dengan memanfaatkan mikrofon sebagai sumber data audio dan kamera sebagai sumber data visual. Hasil prediksi dari masing-masing modalitas kemudian digabungkan untuk menghasilkan keputusan emosi akhir yang ditampilkan oleh system.

Berdasarkan hasil pengujian multimodal secara *real-time* pada Tabel 6, system yang dikembangkan mampu mengenali tiga kategori emosi, yaitu netral, senang, dan marah, dengan memanfaatkan kombinasi informasi dari modalitas audio dan visual. Hasil pengujian menunjukkan bahwa Ketika kedua modalitas menghasilkan prediksi yang konsisten, system mampu memberikan tingkat keyakinan yang tinggi terhadap emosi yang terdeteksi. Selain itu, pada kondisi Ketika prediksi audio dan visual berbeda, metode *late fusion* tetap mampu menghasilkan keputusan akhir dengan mempertimbangkan kontribusi masing-masing modalitas melalui mekanisme pembobotan. Temuan ini menunjukkan bahwa integrasi CNN-BiLSTM-Attention pada audio dan MobileNetV2 pada ekspresi wajah dapat meningkatkan keandalan deteksi emosi dibandingkan penggunaan satu modalitas saja serta mendukung implementasi sistem deteksi emosi secara *real-time*.

Tabel 6. Hasil pengujian multimodal *real-time.*

No.	Hasil pengujian	Keterangan
1.		Sistem berhasil melakukan deteksi emosi secara *real-time* dengan memanfaatkan dua modalitas, yaitu audio dan visual. Pada modalitas visual, model MobileNetV2 memprediksi emosi netral dengan tingkat keyakinan sebesar 75,1%, sedangkan pada modalitas audio model CNN-BiLSTM-Attention memprediksi emosi netral dengan tingkat keyakinan sebesar 87,8%. Hasil prediksi dari dua modalitas tersebut kemudian digabungkan menggunakan metode *late fusion* untuk menghasilkan keputusan akhir sistem.
2.		Sistem berhasil mendeteksi emosi senang secara real-time melalui kombinasi modalitas visual dan audio. Modalitas visual memprediksi emosi senang dengan tingkat keyakinan sebesar 81,3%, sedangkan modalitas audio menghasilkan prediksi dengan tingkat keyakinan sebesar 79,6%. Hasil prediksi dari kedua modalitas digabungkan menggunakan metode *late fusion* sehingga menghasilkan prediksi akhir emosi senang dengan tingkat keyakinan sebesar 80,6%. Hasil prediksi menunjukkan bahwa sistem mampu mengenali pola emosi yang konsisten dari ekspresi wajah maupun karakteristik suara.

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

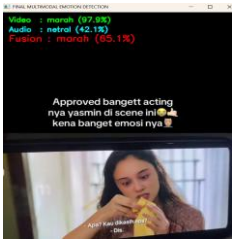
No.	Hasil pengujian	Keterangan
3.		Pada pengujian ini, model visual mengklasifikasikan emosi sebagai marah dengan tingkat keyakinan 77,8%, sedangkan modalitas audio mengklasifikasikan sebagai emosi senang dengan tingkat keyakinan sebesar 60,2%. Perbedaan prediksi antara kedua modalitas menunjukkan bahwa ekspresi wajah dan suara dapat memberikan informasi yang berbeda. Selanjutnya, sistem melakukan proses *late fusion* dengan pembobotan 0,6 untuk visual dan 0,4 untuk audio sehingga menghasilkan prediksi akhir marah dengan probabilitas 50,3%. Hasil ini menunjukkan bahwa keputusan akhir system dipengaruhi oleh kontribusi kedua modalitas, sedangkan modalitas visual memberikan pengaruh yang lebih besar karena memiliki bobot yang lebih tinggi dalam proses integrasi multimodal.

3.2.4 Pengujian Multimodal Video

Pengujian multimodal menggunakan video dilakukan untuk mengevaluasi kemampuan sistem *multimodal emotion recognition* dalam mendeteksi emosi berdasarkan kombinasi informasi audio dan visual yang diperoleh dari video masukan. Pada pengujian ini, sistem memproses ekspresi wajah menggunakan model MobileNetV2 dan karakteristik suara menggunakan model CNN-BiLSTM-Attention, kemudian menggabungkan hasil prediksi kedua modalitas menggunakan metode *late fusion* untuk menghasilkan keputusan emosi akhir. Hasil pengujian ditampilkan pada beberapa sampel video yang mewakili masing-masing kategori emosi, yaitu netral, senang, dan marah.

Setiap hasil menampilkan nilai probabilitas prediksi dari modalitas visual, modalitas audio, dan hasil *fusion* sebagai keputusan akhir sistem. Dengan demikian, pengujian ini dapat digunakan untuk mengamati tingkat konsistensi prediksi antar modalitas serta efektivitas metode *late fusion* dalam meningkatkan keandalan deteksi emosi pada data video. Dibawah ini terdapat Tabel 7, pengujian multimodal menggunakan masukan video.

Tabel 7. Hasil pengujian multimodal video

No.	Hasil Pengujian	Keterangan
1.		Berdasarkan hasil pengujian multimodal pada klip video tersebut, model video berhasil mendeteksi emosi marah dengan tingkat keyakinan yang sangat tinggi, yaitu 97,9%, sedangkan model audio mendeteksi emosi tersebut sebagai netral dengan tingkat keyakinan sebesar 42,1%. Perbedaan hasil ini menunjukkan bahwa ekspresi wajah dan audio yang ditampilkan dalam adegan tersebut lebih jelas mengekspresikan emosi marah. Setelah proses fusi, yang menggabungkan prediksi audio dan video, sistem menghasilkan prediksi akhir berupa emosi marah dengan tingkat kepercayaan sebesar 65,1%. Hasil ini menunjukkan bahwa model multimodal mampu menggabungkan informasi dari kedua modalitas tersebut sekaligus tetap mempertahankan emosi utama yang terdeteksi dari bagian visual, sehingga menghasilkan klasifikasi yang lebih akurat terhadap keadaan emosional yang ditampilkan dalam adegan tersebut.
2.		Berdasarkan hasil pengujian multimodal dari gambar, model video mendeteksi emosi kebahagiaan dengan tingkat keyakinan 76,9%, sedangkan model audio juga mendeteksi emosi kebahagiaan dengan tingkat keyakinan 69,4%. Hasil prediksi yang serupa dari kedua metode tersebut menunjukkan bahwa ekspresi wajah dan karakteristik suara sama-sama mengindikasikan adanya emosi positif. Setelah proses fusi, yang menggabungkan informasi audio dan video, sistem menghasilkan prediksi akhir bahwa emosi yang dimaksud adalah bahagia dengan tingkat kepercayaan sebesar 73,9%. Hasil ini menunjukkan bahwa model multimodal dapat secara konsisten menggabungkan informasi visual dan audio untuk menghasilkan klasifikasi emosi yang lebih stabil dan akurat. Tingkat kepercayaan yang tinggi pada kedua modalitas tersebut menunjukkan bahwa ekspresi wajah dan intonasi suara dalam adegan tersebut saling mendukung untuk menunjukkan emosi bahagia yang diekspresikan.
3.		Berdasarkan hasil uji multimodal, model video mendeteksi emosi kemarahan dengan tingkat keyakinan 91,8%, sedangkan model audio mendeteksi emosi kebahagiaan dengan tingkat keyakinan 82,8%. Perbedaan prediksi ini menunjukkan adanya ketidaksesuaian antara informasi visual dan audio dalam adegan tersebut. Setelah proses fusi, sistem menghasilkan prediksi akhir berupa emosi 'marah' dengan tingkat kepercayaan sebesar 57,5%, yang menunjukkan bahwa informasi visual memiliki pengaruh yang lebih dominan dalam menentukan emosi akhir. Hasil ini menunjukkan kemampuan model untuk menggabungkan dua sumber informasi yang berbeda guna menghasilkan keputusan emosional yang lebih akurat.

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

Berdasarkan pengujian beberapa klip video pada Tabel 7, model multimodal mampu mendeteksi emosi dengan menggabungkan informasi dari modalitas visual dan audio secara efektif. Ketika kedua modalitas tersebut memberikan prediksi yang sama, seperti pada kasus emosi senang, hasil penggabungan tersebut menghasilkan tingkat kepercayaan yang tinggi dan klasifikasi yang lebih stabil. Sementara itu, ketika prediksi untuk video dan audio berbeda, model tetap dapat menentukan emosi yang dominan melalui proses penggabungan, dengan informasi visual biasanya memiliki pengaruh yang lebih besar terhadap keputusan akhir. Hal ini terlihat pada beberapa pengujian di mana prediksi akhir menunjukkan emosi marah, meskipun modalitas audio mendeteksi emosi yang berbeda. Secara keseluruhan, hasil pengujian menunjukkan bahwa penggunaan pendekatan multimodal dapat meningkatkan keandalan deteksi emosi dibandingkan dengan hanya menggunakan satu jenis data, karena pendekatan ini dapat memanfaatkan dan menggabungkan informasi dari ekspresi wajah dan karakteristik suara untuk menghasilkan klasifikasi emosi yang lebih akurat yang mencerminkan keadaan emosional yang ditampilkan dalam video

3.3 Pembahasan

Penelitian tentang pengenalan emosi telah mengalami perkembangan yang pesat dengan memanfaatkan berbagai jenis modalitas, seperti audio, visual, atau kombinasi dari keduanya (multimodal). Berbagai studi sebelumnya menunjukkan bahwa penggunaan lebih dari satu sumber informasi dapat meningkatkan kemampuan sistem dalam mengenali emosi manusia. Namun, tingkat keberhasilan yang dicapai sangat tergantung pada jenis dataset, jumlah kategori emosi, arsitektur model yang digunakan, serta metode fusi yang diterapkan.

Penelitian yang dilakukan oleh Long Liu, Qingquan Luo, Wenbo Zhang, Mengxuan Zhang, dan Bowen Zhai menerapkan metode D-SCFN dan Stacked LSTM pada dataset Aff-Wild2 untuk mengidentifikasi delapan kategori emosi dasar. Temuan dari penelitian ini menunjukkan nilai F1-Score sebesar 44,5% serta performa dalam mendeteksi Action Unit (AU) sebesar 48,3%. Hasil ini menunjukkan bahwa pengenalan emosi dengan jumlah kategori yang cukup banyak dan variasi data yang beragam masih menghadapi tantangan yang cukup rumit.

Penelitian yang dilakukan oleh Long Liu, Qingquan Luo, Wenbo Zhang, Mengxuan Zhang, dan Bowen Zhai menerapkan metode D-SCFN dan Stacked LSTM pada dataset Aff-Wild2 untuk mengidentifikasi delapan kategori emosi dasar. Temuan dari penelitian ini menunjukkan nilai F1-Score sebesar 44,5% serta performa dalam mendeteksi Action Unit (AU) sebesar 48,3%. Hasil ini menunjukkan bahwa pengenalan emosi dengan jumlah kategori yang cukup banyak dan variasi data yang beragam masih menghadapi tantangan yang cukup rumit.

Penelitian lain oleh Mohammed A. Almulla menerapkan Deep Convolutional Neural Networks (DCNN) dan decision-level fusion pada dataset FER-2013, RAVDESS, dan ETB. Penelitian tersebut menghasilkan akurasi sebesar 100% pada data audio, 69% pada data visual, dan 64% pada data teks, sedangkan akurasi multimodal mencapai 80%. Temuan tersebut memperlihatkan bahwa kombinasi beberapa modalitas dapat menghasilkan performa yang lebih stabil dibandingkan penggunaan satu jenis data secara terpisah.

Md. Milon Islam, Sheikh Nooruddin, Fakhri Karray, dan Ghulam Muhammad mengembangkan sistem untuk mengenali emosi dengan pendekatan fusi tingkat model, menggabungkan Bi-LSTM, DSCNN, ResNet-34, dan detektor wajah HOG pada dataset BioVid Emo DB yang mencakup lima kategori emosi. Penelitian ini menghasilkan akurasi sebesar 97,34% saat menggunakan detektor wajah HOG, yang kemudian meningkat menjadi 97,92% dengan penerapan ResNet-34. Tingginya tingkat akurasi ini menunjukkan bahwa informasi visual memiliki peran yang sangat signifikan dalam proses pengenalan emosi.

Sementara itu, Naveed Ahmed, Zaher Al Aghbari, dan Shini Girija melakukan penelitian mengenai berbagai teknik machine learning dan deep learning menggunakan dataset DEAP dan SEED, yang mencakup jumlah kategori emosi yang lebih banyak, yaitu antara 14 hingga 25 jenis emosi. Dalam pendekatan machine learning, algoritma SVM mencatatkan akurasi tertinggi sebesar 91,3% pada dataset DEAP. Sedangkan dalam pendekatan deep learning, kombinasi CNN-RNN-AE berhasil mencapai akurasi 95%, dan kombinasi Bimodal-LSTM meraih akurasi 93,97%. Temuan ini menunjukkan bahwa penerapan arsitektur deep learning dapat meningkatkan kemampuan sistem dalam melakukan klasifikasi emosi yang lebih rumit.

Studi terbaru yang dilakukan oleh Jinhang Liu, Ziyuan Chen, Zhiwei Ye, dan timnya memperkenalkan Adaptive External Attention-enhanced Fusion Network (AEFNet) untuk pengenalan emosi multimodal pada dataset IEMOCAP dan MELD. Penelitian ini mencapai akurasi 74,86% pada IEMOCAP dan 67,36% pada MELD. Hasil ini menunjukkan bahwa penggunaan mekanisme attention dan strategi fusi yang adaptif dapat meningkatkan kemampuan model dalam menggabungkan informasi dari berbagai modalitas.

Berbeda dengan penelitian sebelumnya, penelitian ini menggunakan pendekatan multimodal yang menggabungkan CNN-BiLSTM-Attention untuk modalitas audio dan MobileNetV2 untuk modalitas visual. Kumpulan data yang digunakan mencakup dataset RAVDESS untuk data audio dan dataset yang dikumpulkan sendiri untuk data visual, dengan tiga kategori emosi: netral, senang, dan marah. Hasil pengujian menunjukkan bahwa model audio mencapai akurasi sebesar 63,75%, sedangkan model visual mencapai akurasi sebesar 99,80%.

Berdasarkan hasil penelitian, dapat disimpulkan bahwa modalitas visual memberikan kontribusi yang jauh lebih tinggi dibandingkan dengan modalitas audio dalam mengenali emosi. Tingginya akurasi model visual menunjukkan bahwa arsitektur MobileNetV2 berhasil mengekstraksi karakteristik ekspresi wajah secara efektif, sehingga menghasilkan performa klasifikasi yang sangat baik. Sebaliknya, akurasi model audio yang lebih rendah mengindikasikan bahwa variasi dalam karakteristik suara, intonasi, dan kualitas rekaman masih menjadi tantangan dalam proses pengenalan emosi berbasis audio.


TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi *CNN-BiLSTM-Attention* dan *MobileNetV2* mampu memberikan performa yang baik dalam sistem pengenalan emosi. Meskipun penelitian ini hanya menggunakan tiga kelas emosi, tingkat akurasi yang mencapai 99,80% menunjukkan bahwa pendekatan yang diusulkan memiliki kemampuan yang sangat baik dalam mengenali ekspresi wajah. Hasil penelitian ini juga mendukung kebaruan yang disuslkan pada penelitian in., yaitu penggunaan kombinasi *CNN-BiLSTM-Attention* pada modalitas audio dan *MobileNetV2* pada modalitas visual dalam sistem pengenalan emosi multimodal. Berbeda dengan Sebagian besar penelitian terdahulu yang menggunakan arsitektur multimodal yang kompleks dan melibatkan banyak modalitas, pendekatam yang diusulkan pada penelitian ini hanya memanfaatkan dua modalitas utama, yaitu audio dan visual. Meskipun demikian, model visual mampu mencapai akurasi sebesar 99,80%, yang menunjukkan bahwa arsitektur *MobileNetV2* efektif dalam mengekstraksi fitur ekspresi wajah dengan kebutuhan komputasi yang relative rendah. Temuan ini menunjukkan bahwa system multimodal yang lebih sederhana dan efisien tetap mampu menghasilkan performa yang sangat baik untuk klasifikasi emosi diskrit. Dengan demikian, penelitian ini memberikan kontribusi dalam pengembangan sistem multimodal emotin recognition yang ringan, efektif, dan berpotensi untuk diterapkan pada aplikasi *real-time*.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, proses ekstraksi fitur pada sistem deteksi emosi multimodal berhasil diterapkan pada dua modalitas, yaitu audio dan video. Pada modalitas audio, sinyal suara diolah menjadi representasi Mel-Spectrogram kemudian diproses menggunakan arsitektur CNN-BiLSTM-Attention untuk mempelajari karakteristik emosi, sedangkan pada modalitas video dilakukan ekstraksi fitur ekspresi wajah menggunakan MobileNetV2. Kedua modalitas tersebut selanjutnya diintegrasikan melalui metode *late fusion* (*decision-level fusion*) sehingga berhasil membentuk sistem deteksi emosi multimodal yang mampu mengklasifikasikan tiga kategori emosi, yaitu netral, senang, dan marah. Dengan demikian, tujuan penelitian dalam mengolah fitur audio dan video serta membangun sistem deteksi emosi multimodal berbasis *deep learning* telah tercapai. Berdasarkan hasil evaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*, model audio CNN-BiLSTM-Attention memperoleh akurasi sebesar 63,75%, sedangkan model visual MobileNetV2 memperoleh akurasi sebesar 99,80% pada data pengujian. Hasil evaluasi tersebut menunjukkan bahwa model MobileNetV2 mampu mengenali ekspresi wajah dengan sangat baik, sementara model CNN-BiLSTM-Attention masih memiliki keterbatasan dalam mengenali emosi dari sinyal suara. Secara keseluruhan, penelitian ini berhasil mengimplementasikan sistem deteksi emosi multimodal berbasis suara dan video menggunakan kombinasi CNN-BiLSTM-Attention dan MobileNetV2 serta menjawab rumusan masalah dan mencapai tujuan penelitian yang telah ditetapkan. Penelitian ini berkontribusi dalam pengembangan sistem deteksi emosi multimodal dengan mengintegrasikan CNN-BiLSTM-Attention pada data audio dan MobileNetV2 pada data visual menggunakan metode *Late Fusion*. Integrasi kedua modalitas tersebut memungkinkan sistem menghasilkan prediksi emosi yang lebih akurat dan konsisten dibandingkan pendekatan unimodal. Selain memberikan kontribusi secara teknis pada pengembangan model *deep learning* untuk pengenalan emosi, penelitian ini juga menyediakan dasar implementasi sistem deteksi emosi real-time yang dapat dimanfaatkan pada berbagai aplikasi berbasis kecerdasan buatan.

REFERENCES

Almulla, M. A. (2025). A multimodal emotion recognition system using deep convolution neural networks. *Journal of Engineering Research*, *13*(2), 721–729. https://doi.org/10.1016/j.jer.2024.03.021

Ardiansyah, I., Huda, F. Al, & Yudistira, N. (2024). *Pengembangan Aplikasi Mobile Klasifikasi Ekspresi Wajah Manusia pada Platform Android Menggunakan Arsitektur MobileNet*. *1*(1).

Boitel, E., Mohasseb, A., & Haig, E. (2025). MIST : Multimodal emotion recognition using DeBERTa for text , Semi-CNN for speech , ResNet-50 for facial , and 3D-CNN for motion analysis. *Expert Systems With Applications*, *270*(April 2024), 126236. https://doi.org/10.1016/j.eswa.2024.126236

Budiono, L. A., & Masing, M. (2022). *INNOVATIVE : Volume 2 Nomor 1 Tahun 2022 Research & Learning in Primary Education Emosi Dalam Perspektif Lintas Budaya*. *2*(1995), 579–584.

Cai, X., Yuan, J., Zheng, R., Huang, L., & Church, K. (2021). *Speech Emotion Recognition with Multi-task Learning*. 4508–4512.

Caldwell, J. E., & Beaulieu, M. (2024). *Cross-Modal Knowledge Transfer for Handling Missing Modalities in Affective Computing*.

Dhimas, D., Putra, P., Anaga, G. K., Fitriyana, W. T., & Informatika, T. (2024). *Implementasi Algoritma Convolutional Neural Network Arsitektur Mobilenetv2 Untuk Klasifikasi Ekspresi Wajah Pada Dataset FER*. *3*, 291–297.

Ekspresi, D. A. N., & Sosial, D. A. N. R. (2025). *Kecerdasan Emosional Perempuan :*

Februari, V. N., Hidayat, M. M., Darell, E., Genio, G., Astutik, A. D., & Taufik, A. (2025). *Dike Jurnal Ilmu Multidisiplin Analisis Ekspresi Wajah Untuk Deteksi Emosi Dan Menerapkannya Dalam Game Dan Survei Kepuasan Pengguna Dike Jurnal Ilmu Multidisiplin*. *3*, 31–35.

Guntoro, A. L. S., Julianto, E., & Budiyanto, D. (2022). *Pengenalan Ekspresi Wajah Menggunakan Convolutional Neural Network*. 155–160.

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1065-1080
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10629

Hadi, M. N., & Sari, R. W. (2025). *Multi-Label Emotion Detection for Mental Health Monitoring Using Deep CNN and Visual Attention*. *5*(2), 597–605. https://doi.org/10.30811/jaise.v5i2.6961

Islam, M., Nooruddin, S., Karray, F., & Muhammad, G. (2024). Enhanced multimodal emotion recognition in healthcare analytics : A deep learning based model-level fusion approach. *Biomedical Signal Processing and Control*, *94*(November 2023), 106241. https://doi.org/10.1016/j.bspc.2024.106241

Kamal, A. (2025). *Pengenalan Emosi Oleh AI dalam komunikasi dan pengambilan keputusan . Kemampuan untuk mengenali sosial dan profesional . Dengan perkembangan teknologi kecerdasan buatan menarik dan memiliki berbagai aplikasi potensial 1 . Pengenalan emosi oleh AI melibatk*. *1*(1), 1–9.

Khubrani, M. (2026). *Article in Press Multimodal Stress-Aware Ensemble Learning fusion for real-time proxy-based stress recognition integrating typing , facial , and voice IN IN*.

Li, Y., Wei, J., Liu, Y., Kauttonen, J., & Zhao, G. (2022). Deep Learning for Micro-Expression Recognition : A Survey. *IEEE Transactions on Affective Computing*, *13*(4), 2028–2046. https://doi.org/10.1109/TAFFC.2022.3205170

Liu, L., Luo, Q., Zhang, W., Zhang, M., & Zhai, B. (2025). Multimodal emotion recognition method in complex dynamic scenes. *Journal of Information and Intelligence*, *3*(3), 257–274. https://doi.org/10.1016/j.jiixd.2025.02.004

Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences*, *2*(3), 1–8. https://doi.org/10.1007/s42452-020-2234-1

N.A., M. I. (2025). *Pengembangan Speech Emotion Recognition*. 1–2.

Nurjihan, S. W., Nurbadillah, N., Faturrahman, N., Wiguna, I. M., Lasardi, E. M., Giri, E. P., & Mindara, G. P. (2024). indonesia Pengenalan Pola Ekspresi Wajah Untuk Pengolahan Citra Menggunakan Metode Convolutional Neural Network. *JATISI*, *11*(4).

Razaqa, D., Rezka, M., Maghribi, A., Gunasti, N. D., & Wati, T. (2024). *Analisis Etika dan Dampak Penggunaan Sistem Pengenalan Wajah untuk Manajemen Kehadiran di Lingkungan Sekolah*. *4221*, 50–57.

Riyandi, R., & Sumarsono, S. (2025). Analisis Efektivitas Fusi Fitur Multimodal dalam Klasifikasi Citra Daun Herbal. *Jurnal Teknik Informatika Dan Sistem Informasi*, *11*(3), 463–474. https://doi.org/10.28932/jutisi.v11i3.12262

Roihan, A., Dwi, R., & Astriyani, E. (2025). Analisis Konseptual Fusi Multimodal Wajah dan Visual Speech untuk Autentikasi Biometrik Non-Vokal terhadap Deepfake. *Journal of Big Data Analytic and Artificial Intelligence*, *8*(2), 54–61. https://doi.org/10.71302/jbidai.v8i2.84

S Budiawan, ST Alvianus Dengen, R. A. (2025). *Dialog Digital: Memahami Komunikasi Manusia Dan Mesin Dalam Era Interkoneksi*.

Shane, K., Prakoso, B., & Prasetio, B. H. (2025). *Implementasi Speech Recognition berbasis Raspberry Pi 5 pada Ekosistem Smart-Home menggunakan Algoritma Gated Recurrent Unit (GRU)*. *9*(3), 1–10.

Singla, C., Singh, S., Sharma, P., Mittal, N., & Gared, F. (2024). Emotion recognition for human – computer interaction using high - level descriptors. *Scientific Reports*, 1–12. https://doi.org/10.1038/s41598-024-59294-y

Sutarti, & Fariza Syaqialloh. (2025). Klasifikasi dan Pengenalan Emosi dari Ekspresi Wajah Menggunakan CNN-BiLSTM dengan Teknik Data Augmentation. *Decode: Jurnal Pendidikan Teknologi Informasi*, *5*(1), 79–91. https://doi.org/10.51454/decode.v5i1.1038

Syaqialloh, F. (2025). *Klasifikasi dan Pengenalan Emosi dari Ekspresi Wajah Menggunakan CNN-BiLSTM dengan Teknik Data Augmentation*. *5*(1), 79–91.

Wijaya, H. (2024). *Teknologi Pengenalan Suara tentang Metode , Bahasa dan Tantangan : Systematic Literature Review*. *7*(2). https://doi.org/10.32877/bt.v7i2.1888

Wu, R., Wang, H., Chen, H., & Carneiro, G. (n.d.). *Deep Multimodal Learning with Missing Modality: A Survey*. *1*(1).