

PENGENALAN EMOSI MANUSIA MELALUI SUARA MENGUNAKAN METODE *CONVOLUTIONAL NEURAL NETWORK* DAN TRANSFORMER

**Ida Bagus Sebali Bayu Nandana^{1*}, Ni Putu Novita Puspa Dewi², I Nyoman Saputra
Wahyu Wijaya³**

^{1,2,3}Program Studi Ilmu Komputer, Fakultas Teknik dan Kejuruan, Universitas Pendidikan
Ganesha, Singaraja, Indonesia

bagus.sebali@student.undiksha.ac.id, novita.puspa.dewi@undiksha.ac.id
, wahyu.wijaya@undiksha.ac.id

ABSTRAK

Komunikasi antara manusia dan mesin saat ini masih minim komponen emosional karena sistem umumnya hanya memahami konten verbal tanpa menangkap muatan emosi dari nada dan timbre suara. Penelitian ini bertujuan merancang model pengenalan emosi berbasis suara (Speech Emotion Recognition) dengan mengintegrasikan arsitektur Convolutional Neural Network (CNN) dan Transformer guna menangkap fitur spasial sekaligus dependensi temporal jangka panjang pada sinyal wicara. Data penelitian merupakan gabungan empat korpus publik, yaitu RAVDESS, CREMA-D, SAVEE, dan TESS, dengan total 12.798 berkas audio yang mencakup tujuh kelas emosi (marah, jijik, takut, senang, netral, sedih, dan terkejut). Sinyal audio diproses melalui ekstraksi fitur Mel Spectrogram beserta koefisien delta dan delta-delta, kemudian diseimbangkan melalui augmentasi data sebelum dilatih menggunakan skema 5-Fold Stratified Cross Validation. Hasil pengujian pada data uji independen menunjukkan model hibrida mencapai rata-rata akurasi 5-fold sebesar 75,80%, dengan model performa terbaik (fold tunggal) mencapai akurasi 77,00% dan Macro Average F1-Score 78,00%, serta seluruh kelas emosi memperoleh nilai AUC di atas 0,90. Penerapan Ensemble Learning (Soft Voting) dari kelima model fold selanjutnya meningkatkan akurasi pengujian menjadi 81,00% dengan Macro Average F1-Score 0,82, yang menjadi capaian akhir tertinggi dalam penelitian ini. Hasil ini membuktikan bahwa integrasi CNN dan Transformer efektif serta mampu melakukan generalisasi dengan baik pada data suara yang heterogen.

Kata kunci: pengenalan emosi suara; convolutional neural network; transformer; mel spectrogram; deep learning

ABSTRACT

Communication between humans and machines currently lacks a natural emotional component, since systems generally understand verbal content without capturing the emotional meaning conveyed by tone and timbre. This study aims to design a Speech Emotion Recognition model by integrating Convolutional Neural Network (CNN) and Transformer architectures to capture spatial features as well as long-range temporal dependencies in speech signals. The dataset combines four public corpora, namely RAVDESS, CREMA-D, SAVEE, and TESS, comprising 12,798 audio files across seven emotion classes (angry, disgusted, fearful, happy, neutral, sad, and surprised). The audio signals were processed through Mel Spectrogram feature extraction along with delta and delta-delta coefficients, balanced through data augmentation, and trained using a 5-Fold Stratified Cross Validation scheme. Testing on an independent test set shows that the hybrid model achieves a 5-fold average accuracy of 75.80%, while the best-performing single-fold model reaches 77.00% accuracy with a Macro Average F1-Score of 78.00%, and all emotion

classes obtain an AUC value above 0.90. Applying Ensemble Learning (Soft Voting) across the five fold models further improves the testing accuracy to 81.00% with a Macro Average F1-Score of 0.82, representing the highest overall result achieved in this study. These results demonstrate that the integration of CNN and Transformer is effective and generalizes well on heterogeneous speech data.

Keywords: *Speech Emotion Recognition; convolutional neural network; transformer; mel spectrogram; deep learning*

1. PENDAHULUAN

Emosi merupakan aspek krusial dalam kehidupan manusia dan memiliki pengaruh signifikan terhadap perilaku. Kemampuan mengenali emosi membantu manusia untuk saling memahami, terutama dalam komunikasi. Konsep emosi berkaitan erat dengan affect dan mood yang membentuk spektrum pengalaman psikologis yang kompleks [1]. Namun, komunikasi antara manusia dan mesin saat ini dinilai masih kurang memiliki komponen emosional yang natural, karena mesin umumnya hanya mampu memahami konten verbal tanpa mampu menangkap makna emosional di balik nada dan timbre suara.

Emosi dapat diekspresikan melalui berbagai saluran, salah satunya adalah suara, khususnya dalam percakapan spontan. Pengenalan Emosi Suara (*Speech Emotion Recognition/SER*) telah menjadi bidang penelitian penting dalam interaksi manusia-komputer (*Human-Computer Interaction/HCI*), yang memanfaatkan fakta bahwa suara sering mencerminkan emosi melalui nada dan timbre. Berbeda dengan ekspresi wajah, pengenalan emosi dari suara terutama bergantung pada pengolahan sinyal berkelanjutan berdasarkan fitur suara dan fitur prosodi [2]. Meskipun kemajuan signifikan telah dicapai, identifikasi emosi secara akurat masih menjadi tantangan yang sulit karena sifat pembentukan dan ekspresi emosi yang kompleks serta bervariasi antarindividu.

Berbagai penelitian terdahulu telah mengeksplorasi permasalahan pengenalan emosi melalui suara dengan pendekatan yang beragam. Sejumlah studi memanfaatkan teknik ekstraksi fitur klasik yang dipadukan dengan model pembelajaran mesin konvensional, seperti *Support Vector Classifier* untuk otentikasi suara [3], *Support Vector Machine* dengan fitur *Discrete Wavelet Transform* [4], Hidden Markov Model yang dikombinasikan dengan *Bidirectional Associative Memory* [5], serta berbagai teknik machine learning lain untuk deteksi emosi dalam bahasa lisan [6]. Pendekatan lain memanfaatkan sinyal biologis seperti EEG melalui tinjauan komprehensif berbasis machine learning [7]. Sinyal EEG dinilai mampu merepresentasikan pengalaman afektif manusia sekaligus mendeteksi perubahan awal dalam reaksi emosional [8], serta unggul dari sisi portabilitas dan biaya rendah dalam menyediakan data spasial, temporal, dan spektral mengenai kondisi emosional secara langsung melalui aktivitas otak [9], [10]. Akurasi pengenalan emosi berbasis EEG turut dipengaruhi oleh faktor individu seperti kemampuan kognitif dan jenis kelamin, yang dapat memunculkan variasi pola sinyal antarsubjek [11].

Perkembangan deep learning selanjutnya mendorong penggunaan arsitektur *Convolutional Neural Network* (CNN) secara luas dalam SER, antara lain melalui kombinasi fitur hibrida dengan CNN [12], penggabungan CNN dengan *Extreme Learning Machine* untuk pengenalan emosi audiovisual [13], penerapan CNN murni untuk klasifikasi emosi ucapan [14], [15], [16], serta CNN yang dipadukan dengan LSTM [17]

maupun classifier gamma berbasis *Error Correcting Output Code* [18]. Studi lain menekankan pentingnya peningkatan akurasi SER dalam konteks Bahasa Indonesia [19], dan penerapan deep learning untuk peningkatan pengalaman pengguna secara umum [20]. Beberapa penelitian juga membandingkan performa berbagai arsitektur jaringan saraf tiruan untuk mendeteksi emosi melalui suara dan sinyal ucapan [21], [22]. Keunggulan utama CNN, yakni kemampuannya mendeteksi elemen visual penting secara otomatis serta efisiensinya dalam menangani variasi data tanpa memerlukan rekayasa fitur manual yang kompleks, turut menjadi salah satu faktor pendorong adopsi arsitektur ini secara luas [23].

Secara khusus, Mountzouris et al. [24] mengusulkan model CNN yang dipadukan dengan mekanisme *attention* untuk pengenalan emosi berbasis suara dan berhasil mencapai akurasi 77% pada dataset RAVDESS serta 74% pada dataset SAVEE, didukung oleh penggunaan *Mel-frequency cepstral coefficients* (MFCC) sebagai fitur masukan. Penelitian tersebut turut menyarankan eksplorasi arsitektur modern lain seperti Transformer, yang unggul dalam memahami hubungan jangka panjang antareleman sekuensial dalam data. Rekomendasi ini diperkuat oleh Ullah et al. [25] yang membuktikan bahwa kombinasi CNN dan *Multi-Head Convolutional Transformer* mampu meningkatkan akurasi pengenalan emosi suara pada dataset RAVDESS hingga 82,31%. Perkembangan lebih lanjut menunjukkan bahwa mekanisme attention berbasis Transformer murni turut terbukti efektif, sebagaimana ditunjukkan oleh Akilandasowmya dan Mayuranathan [26] yang menerapkan multi-head self-attention pada fitur gabungan chroma, Mel spectrogram, dan MFCC untuk mencapai akurasi 87,4% pada dataset CREMA-D. Pendekatan berbasis ensemble classifier juga menunjukkan hasil yang menjanjikan, seperti pada penelitian Barik dan Sridhar [27] yang mengombinasikan ekstraksi fitur Inception-v4 dengan Random Forest sebagai classifier, mencapai akurasi 97,57% pada dataset RAVDESS dan 90,62% pada dataset SAVEE. Selain itu, representasi berbasis Self-Supervised Learning seperti WavLM juga terbukti efektif ketika dipadukan dengan mekanisme kompresi fitur, sebagaimana ditunjukkan oleh Chakhtouna dkk. [28] yang mencapai akurasi 84,72% pada dataset RAVDESS dalam skenario speaker-dependent. Namun demikian, arsitektur CNN murni masih memiliki keterbatasan dalam menangkap dependensi temporal jangka panjang pada sinyal wicara yang dinamis, sedangkan mekanisme self-attention pada Transformer terbukti efektif dalam menangkap hubungan kontekstual antarfitur akustik secara simultan [25].

Berdasarkan uraian tersebut, penelitian ini mengusulkan arsitektur hibrida yang menggabungkan CNN dan Transformer untuk mengenali emosi manusia melalui suara. Kombinasi ini dipilih untuk menggabungkan keunggulan CNN dalam mengekstraksi fitur spasial secara hierarkis dari representasi citra Mel spectrogram, dengan kekuatan mekanisme self-attention pada Transformer dalam menangkap dependensi jangka panjang antarfitur akustik secara simultan. Penelitian ini juga menggabungkan empat dataset publik (RAVDESS, CREMA-D, SAVEE, dan TESS) sekaligus, berbeda dengan sebagian besar penelitian terdahulu yang hanya menggunakan satu korpus tunggal, sehingga model yang dihasilkan diharapkan memiliki kemampuan generalisasi yang lebih representatif terhadap variabilitas akustik dan demografi penutur di dunia nyata.

Adapun tujuan dari penelitian ini adalah: (1) merancang dan membangun pemodelan sistem pengenalan emosi berbasis suara dengan menerapkan arsitektur CNN dan Transformer secara terintegrasi guna meningkatkan akurasi dan efektivitas proses

identifikasi emosi; dan (2) melakukan evaluasi secara menyeluruh terhadap performa model yang dikembangkan menggunakan metrik akurasi, precision, recall, F1-Score, dan ROC/AUC, sehingga dapat menggambarkan kinerja sistem pada berbagai kondisi data dan konteks penggunaan.

2. METODE PENELITIAN

Penelitian ini mengadopsi pendekatan kuantitatif yang bertujuan untuk mengukur, menganalisis, dan menginterpretasikan data numerik guna menilai performa metode CNN dan Transformer dalam pengenalan emosi manusia berdasarkan suara. Variabel independen dalam penelitian ini mencakup metode klasifikasi yang diterapkan, teknik ekstraksi fitur, serta konfigurasi parameter model seperti jumlah lapisan, ukuran filter, dan learning rate. Variabel dependen merupakan kinerja model yang diukur melalui metrik evaluasi akurasi, precision, recall, F1-Score, dan AUC-ROC, sedangkan variabel kontrol meliputi format dataset, teknik preprocessing, dan algoritma pelatihan yang diterapkan secara konsisten pada setiap eksperimen.

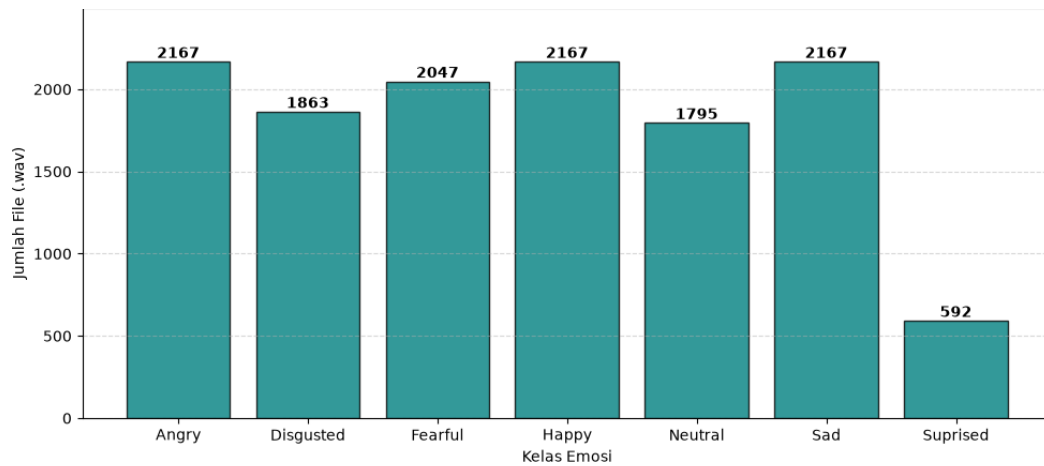
Dataset dan Pengumpulan Data

Data yang digunakan merupakan data sekunder yang bersumber dari kumpulan data RAVDESS, CREMA-D, SAVEE, dan TESS yang diakses melalui repositori publik Kaggle. Dataset tersebut secara keseluruhan mencakup 12.798 berkas audio rekaman suara berformat .wav yang terdistribusi ke dalam tujuh kategori kelas emosi, yaitu marah, jijik, takut, senang, netral, sedih, dan terkejut. Rincian profil dataset disajikan pada Tabel 1.

Tabel 1. Profil Dataset

Kelas	RAVDESS	CREMA-D	SAVEE	TESS	Jumlah
Marah	376	1.271	120	400	2.167
Jijik	192	1.271	-	400	1.863
Takut	376	1.271	-	400	2.047
Senang	376	1.271	120	400	2.167
Netral	188	1.087	120	400	1.795
Sedih	376	1.271	120	400	2.167
Terkejut	192	-	-	400	592
Jumlah	2.076	7.442	480	2.800	12.798

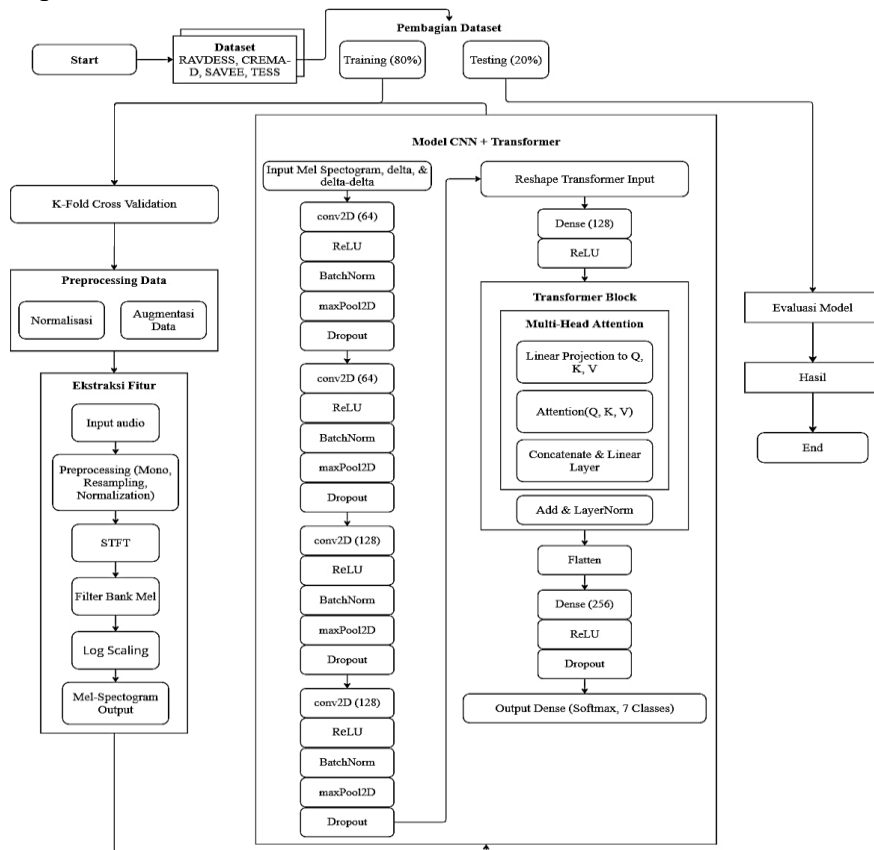
Dataset CREMA-D menyumbang porsi data terbesar dengan 7.442 berkas, sedangkan dataset SAVEE memiliki proporsi terkecil yakni 480 berkas. Seluruh berkas audio dipisahkan secara terprogram ke dalam kelompok data pelatihan dan validasi sebesar 80% (10.238 berkas) serta kelompok data pengujian murni (*testing set*) sebesar 20% (2.560 berkas) sebelum masuk ke tahap prapemrosesan. Distribusi jumlah berkas untuk setiap kelas emosi disajikan pada Gambar 1.



Gambar 1. Distribusi Jumlah Berkas Audio per Kelas Emosi

Tahapan Penelitian

Penelitian ini terdiri atas beberapa tahapan utama, yaitu pengumpulan data, *preprocessing* data, augmentasi data, ekstraksi fitur menjadi Mel Spectrogram, pelatihan model CNN dan Transformer menggunakan skema *5-Fold Stratified Cross Validation*, hingga evaluasi kinerja model pada data pengujian. Alur penelitian secara lengkap ditampilkan pada Gambar 2.



Gambar 2. Alur Penelitian

Preprocessing Data

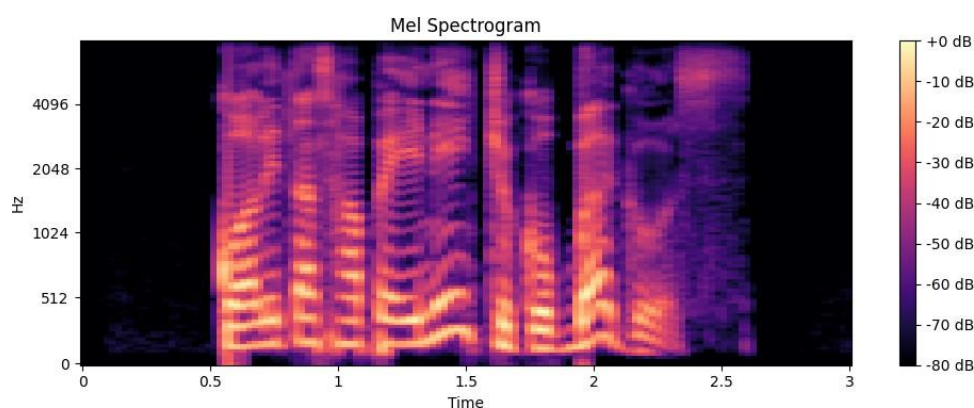
Tahap *preprocessing* meliputi: (a) normalisasi format audio ke ekstensi .wav dengan penyesuaian sample rate 16.000 Hz dan durasi tetap 4 detik agar seluruh berkas dapat diproses melalui pipeline yang sama; dan (b) augmentasi data yang diterapkan hanya pada data latih dan validasi untuk mengatasi ketidakseimbangan jumlah sampel antarkelas emosi, sementara data pengujian murni dipertahankan pada kondisi asli. Teknik augmentasi yang diterapkan berbasis pustaka Librosa meliputi *Additive White Gaussian Noise* (AWGN) dengan noise level 0,003, *Time Stretching* perlambatan dengan faktor 0,9, dan *Time Stretching* percepatan dengan faktor 1,1, yang diterapkan secara bergantian melalui algoritma *forced balancing* hingga jumlah berkas kelas minoritas setara dengan kelas mayoritas. Rekapitulasi jumlah data sebelum dan sesudah augmentasi disajikan pada Tabel 2.

Tabel 2. Data Sebelum dan Sesudah Augmentasi

Data	Sebelum	Sesudah
Training dan Validasi	10.238 (80%)	12.138 (82,58%)
Testing	2.560 (20%)	2.560 (17,42%)
Total	12.798	14.698

Ekstraksi Fitur

Sinyal audio diubah dari domain waktu ke domain frekuensi menggunakan *Short-Time Fourier Transform* (STFT), kemudian dikonversi ke skala Mel menggunakan filter bank Mel sehingga distribusi frekuensi menyesuaikan dengan cara manusia mendengar suara. Transformasi logaritma diterapkan pada amplitudo spektrum untuk menghasilkan Mel spectrogram sebagaimana ditunjukkan pada Gambar 3. Selain fitur statis Mel Spectrogram, diekstraksi pula koefisien Delta (turunan pertama) dan Delta-Delta (turunan kedua) untuk menangkap informasi transisi nada dan energi suara seiring waktu. Efektivitas representasi Mel Spectrogram beserta fitur turunannya juga telah dibuktikan pada berbagai tugas berbasis suara lain, seperti klasifikasi ucapan gagap (stuttering) menggunakan fitur MFCC dan Mel Spectrogram beserta turunan derajat keduanya [29], serta identifikasi patologi suara menggunakan fitur Mel spectrogram dengan pendekatan transfer learning [30]. Ketiga komponen fitur tersebut digabungkan menjadi satu matriks fitur berdimensi $800 \times 128 \times 3$ sebagai masukan model.



Gambar 3. Mel Spectrogram

Arsitektur Model Hibrida CNN-Transformer

Arsitektur model yang diusulkan terdiri atas empat blok konvolusi (Conv2D-BatchNorm-MaxPool2D-Dropout) dengan jumlah filter 64, 64, 128, dan 128, masing-masing menggunakan kernel 3×3 dan fungsi aktivasi ReLU untuk mengekstraksi fitur spasial secara hierarkis. Keluaran blok konvolusi terakhir kemudian di-reshape menjadi representasi sekuensial dan diproyeksikan melalui lapisan Dense 128 unit sebelum dimasukkan ke dalam Transformer Block yang memanfaatkan *Multi-Head Attention* dengan 8 heads. Representasi hasil Transformer Block dirangkum melalui *Global Average Pooling* 1D menjadi vektor berukuran 128, kemudian diproses oleh lapisan Dense 256 unit dengan regulasi *Dropout* (0,2) sebelum dipetakan oleh lapisan Dense Softmax menuju probabilitas prediksi tujuh kelas emosi. Rincian perubahan dimensi pada setiap lapisan model disajikan pada Tabel 3.

Tabel 3. Alur Perubahan Dimensi Lapisan Model Hibrida

Lapisan	Konfigurasi	Output Shape	Keterangan
Input Layer	-	(None, 800, 128, 3)	Matriks Mel Spectrogram, delta, delta-delta
Conv2D Block 1	64 filter, 3x3, ReLU	(None, 800, 128, 64)	Ekstraksi fitur spasial awal
MaxPool2D Block 1	Pool 2x2	(None, 400, 64, 64)	Reduksi dimensi pertama
Conv2D Block 2	64 filter, 3x3, ReLU	(None, 400, 64, 64)	Ekstraksi fitur tingkat menengah
MaxPool2D Block 2	Pool 2x2	(None, 200, 32, 64)	Reduksi dimensi kedua
Conv2D Block 3	128 filter, 3x3, ReLU	(None, 200, 32, 128)	Ekstraksi fitur tingkat tinggi
MaxPool2D Block 3	Pool 2x2	(None, 100, 16, 128)	Reduksi dimensi ketiga
Conv2D Block 4	128 filter, 3x3, ReLU	(None, 100, 16, 128)	Ekstraksi fitur tingkat tinggi
MaxPool2D Block 4	Pool 2x2	(None, 50, 8, 128)	Reduksi dimensi akhir CNN
Reshape Layer	Height x Width	(None, 50, 1024)	Transformasi data 3D ke sekuensial
Projection Dense	128 unit, ReLU	(None, 50, 128)	Penyelarasan dimensi Transformer
Transformer Block	8 heads MHA, FFN 256	(None, 50, 128)	Pemodelan konteks waktu
GlobalAvgPooling1D	Rata-rata sumbu waktu	(None, 128)	Flatten representasi sekuensial
Classifier Dense	256 unit, ReLU	(None, 256)	Interpretasi fitur akhir
Output Dense	7 units, Softmax	(None, 7)	Probabilitas kelas emosi

Model diimplementasikan menggunakan pustaka TensorFlow dan Keras. Proses pelatihan dijalankan maksimal 200 epoch dengan ukuran batch 64/128, optimasi AdamW pada *learning rate* 0,0001, fungsi kerugian categorical crossentropy, penjadwalan laju pembelajaran berbasis Cosine Decay, serta mekanisme Early Stopping untuk menghentikan pelatihan secara otomatis ketika model mencapai titik konvergensi optimal. Evaluasi kestabilan model diuji melalui skema *5-Fold Stratified Cross Validation*, dengan arsitektur yang dilatih ulang pada setiap fold secara independen. Teknik k-fold cross validation dipilih karena mampu membagi data ke dalam beberapa

lipatan sehingga proses pelatihan dan validasi dapat dilakukan secara bergantian, sementara sebagian kecil data tetap dipertahankan sebagai data uji akhir yang independen [31], [32].

Tahap Pengujian

Pengujian performa model dilakukan pada kelompok data uji sebesar 20% dari keseluruhan dataset asli (2.560 berkas audio) yang tidak pernah dilibatkan dalam proses pelatihan, validasi, maupun augmentasi, sehingga mampu mencerminkan tingkat akurasi evaluasi yang objektif. Evaluasi keberhasilan model dalam mengklasifikasikan tujuh kelas emosi diukur menggunakan pendekatan *Confusion Matrix*, dengan parameter kuantitatif *Accuracy*, *Precision*, *Recall*, dan *F1-Score* yang dihitung menggunakan pustaka Scikit-Learn. Performa diskriminasi model turut diperkuat melalui visualisasi grafik *Receiver Operating Characteristic* (ROC) dan nilai *Area Under the Curve* (AUC) untuk kasus klasifikasi multikelas.

3. HASIL DAN PEMBAHASAN

Perbandingan dengan Penelitian Terdahulu

Untuk mengonfirmasi validitas performa arsitektur yang diusulkan, model hibrida turut diujicobakan secara terisolasi pada dataset tunggal RAVDESS dengan modifikasi susunan filter konvolusi (32, 64, 128, 128) serta augmentasi ekstensif. Pada skema tersebut, model mencapai akurasi validasi sebesar 94,77% dan akurasi pengujian sebesar 85,40%, dengan kesenjangan generalisasi yang sangat tipis (2,8%). Pencapaian ini melampaui hasil penelitian Mountzouris et al. [24] yang memperoleh akurasi 77% pada RAVDESS, serta mengungguli Ullah et al. [25] yang mencapai 82,31% pada dataset yang sama. Perbandingan lengkap disajikan pada Tabel 4.

Tabel 4. Perbandingan Kinerja Model Usulan dengan Penelitian Terdahulu

Peneliti (Tahun)	Arsitektur Model	Dataset	Akurasi	Akurasi Ensemble (Soft Voting)
Mountzouris et al. [24] (2023)	CNN + Attention	RAVDESS (tunggal)	77,00%	-
Ullah et al. [25] (2023)	CNN + Multi-Head Transformer	RAVDESS (tunggal)	82,31%	-
Zielonka et al. [33] (2022)	CNN (ResNet18)	Gabungan (RAVDESS, CREMA-D, SAVEE, TESS)	57,42%	-
Penelitian Usulan (2026)	CNN + Transformer Block	RAVDESS (tunggal)	85,40% (rata-rata 5-fold)	- (tidak diterapkan)

Penelitian Usulan (2026)	CNN + Transformer Block	Gabungan (RAVDESS, CREMA-D, SAVEE, TESS)	75,80% (rata- rata 5-fold)	81,00%
Peneliti (Tahun)	Arsitektur Model	Dataset	Akurasi	Akurasi Ensemble (Soft Voting)

Ketika diuji pada dataset gabungan yang jauh lebih heterogen dari sisi usia, aksen, dan kondisi perekaman, model tetap menunjukkan ketangguhan yang baik dengan akurasi validasi 79,65% dan akurasi pengujian 75,80%, jauh melampaui hasil Zielonka et al. [33] yang hanya mencapai 57,42% pada gabungan dataset yang identik. Penerapan tambahan *Ensemble Learning (Soft Voting)* dari lima model fold berhasil meningkatkan akurasi pengujian menjadi 81,00%, menegaskan bahwa arsitektur hibrida yang diusulkan memiliki fondasi generalisasi yang jauh lebih representatif terhadap kondisi data suara yang bervariasi.

Perlu ditegaskan bahwa penelitian ini melaporkan hasil pada tiga tingkat yang berbeda agar tidak menimbulkan kerancuan interpretasi: (1) rata-rata akurasi 5-fold cross validation, yang merepresentasikan performa generalisasi model secara keseluruhan dan menjadi acuan utama pada Tabel 4; (2) akurasi model performa terbaik, yaitu model dari satu fold dengan performa tertinggi, yang digunakan untuk analisis mendalam per kelas emosi pada Tabel 6 dan Gambar 4; serta (3) akurasi hasil Ensemble Learning (Soft Voting), yang menggabungkan probabilitas prediksi dari kelima model fold dan menjadi capaian akhir tertinggi pada penelitian ini. Pada dataset gabungan, ketiga tingkat tersebut secara berurutan mencapai 75,80% (rata-rata 5-fold), 77,00% (model terbaik, fold 1), dan 81,00% (ensemble, Tabel 7).

Evaluasi Kinerja Model pada Dataset Gabungan

Data eksperimen terbagi menjadi data pelatihan dan validasi menggunakan skema *5-Fold Stratified Cross Validation* (12.138 sampel setelah augmentasi) serta data pengujian murni sebesar 20% dari dataset dasar (2.560 sampel). Sebaran akurasi dan loss pada setiap fold disajikan pada Tabel 5.

Tabel 5. Distribusi Nilai Evaluasi Metrik Makro Fase Pelatihan dan Pengujian Pada Dataset Gabungan

Fold	Training Acc.	Training Loss	Validation Acc.	Validation Loss	Testing Acc.
1	0,9206	0,2167	0,8085	0,6569	0,7700
2	0,8878	0,2991	0,7879	0,6923	0,7600
3	0,9235	0,2137	0,8097	0,6582	0,7600
4	0,8300	0,4549	0,7697	0,6591	0,7400
5	0,8782	0,3255	0,8068	0,6098	0,7600
Mean	0,8880	0,3020	0,7965	0,6553	0,7580

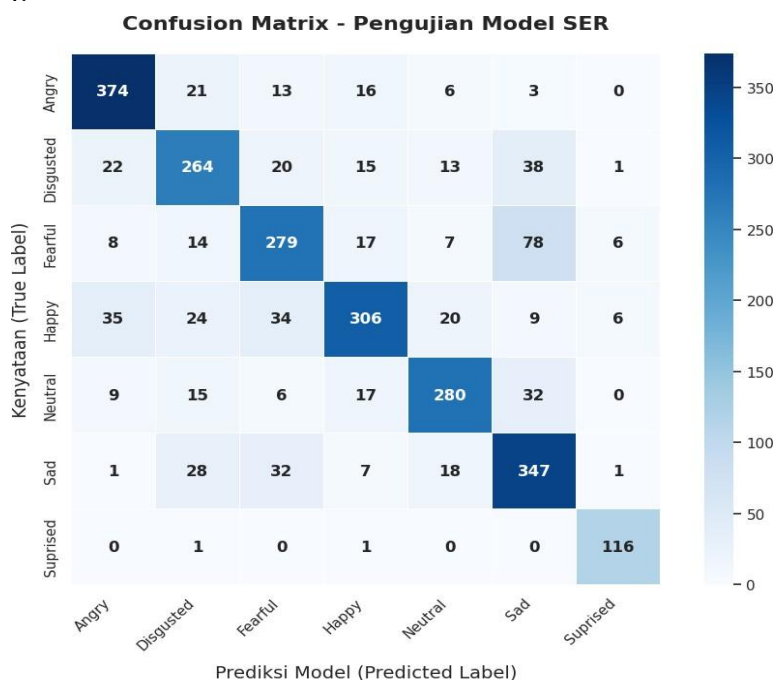
Tabel 5 menunjukkan stabilitas kinerja model yang tinggi, dengan rata-rata akurasi validasi 79,65% dan rata-rata akurasi testing 75,80%. Kesenjangan performa

(generalization gap) antara validasi dan pengujian murni tercatat sangat minim, yakni hanya 3,85%, yang mengindikasikan model tidak mengalami overfitting yang signifikan. Evaluasi performa secara mikro pada setiap kelas emosi untuk model fold 1 (model paling optimal) disajikan pada Tabel 6.

Tabel 6. Laporan Matriks Evaluasi Mikro Kelas Emosi (Dataset Gabungan) (Model Terbaik - Fold 1)

Kategori Emosi	Precision	Recall	F1-Score	Support
Marah	0,83	0,86	0,85	433
Jijik	0,72	0,71	0,71	373
Takut	0,73	0,68	0,70	409
Senang	0,81	0,71	0,75	434
Netral	0,81	0,78	0,80	359
Sedih	0,68	0,80	0,74	434
Terkejut	0,89	0,98	0,94	118
Accuracy			0,77	2.560
Macro avg	0,78	0,79	0,78	2.560
Weighted avg	0,77	0,77	0,77	2.560

Tabel 6 mengonfirmasi bahwa model memiliki sensitivitas identifikasi paling tinggi pada kategori emosi terkejut (F1-Score 0,94) dan marah (F1-Score 0,85), yang memiliki karakteristik akustik cukup khas dan mudah dibedakan. Performa pengenalan terendah berada pada kategori jijik dan takut (F1-Score masing-masing 0,71 dan 0,70), yang disebabkan oleh tingginya variabilitas kesamaan fitur akustik antara kedua kelas tersebut. Visualisasi Confusion Matrix hasil pengujian pada dataset gabungan disajikan pada Gambar 4.



Gambar 4. Confusion Matrix Hasil Pengujian pada Dataset Gabungan

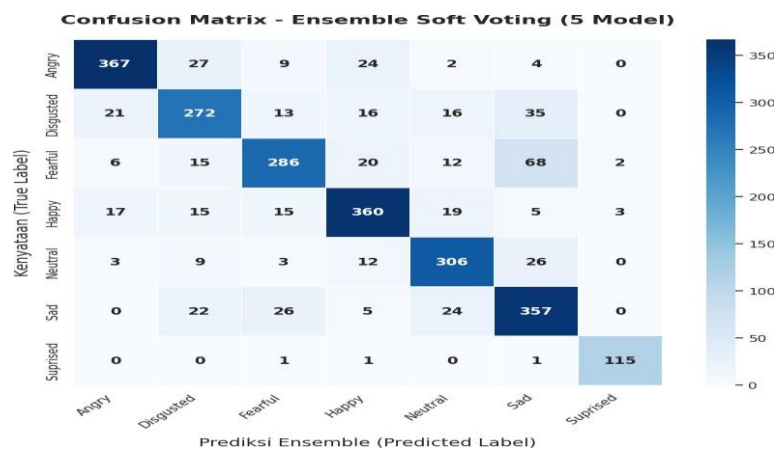
Gambar 4 memperlihatkan bahwa kesalahan klasifikasi umumnya terjadi antara kelas jijik dan takut, serta antara kelas sedih dan jijik, yang secara akustik memiliki pola energi dan intonasi yang relatif mirip. Sebaliknya, kelas terkejut memiliki prediksi yang sangat presisi dengan kesalahan klasifikasi yang minimal terhadap kelas lain.

Untuk mengonfirmasi peningkatan performa akibat penerapan Ensemble Learning (Soft Voting) terhadap kelima model fold pada dataset gabungan, Tabel 7 menyajikan laporan matriks evaluasi mikro tiap kelas emosi dari hasil pengujian ensemble.

Tabel 7. Laporan Matriks Evaluasi Mikro Kelas Emosi pada Model Dataset Gabungan dengan Ensemble Learning (Soft Voting)

Kategori Emosi	Precision	Recall	F1-Score	Support
Marah	0,89	0,85	0,87	433
Jijik	0,76	0,73	0,74	373
Takut	0,81	0,70	0,75	409
Senang	0,82	0,83	0,83	434
Netral	0,81	0,85	0,83	359
Sedih	0,72	0,82	0,77	434
Terkejut	0,96	0,97	0,97	118
Accuracy			0,81	2.560
Macro avg	0,82	0,82	0,82	2.560
Weighted avg	0,81	0,81	0,81	2.560

Tabel 7 menunjukkan bahwa Ensemble Learning (Soft Voting) meningkatkan akurasi pengujian dari 77,00% (model terbaik, fold 1) menjadi 81,00%, serta menaikkan Macro Average F1-Score dari 0,78 menjadi 0,82. Peningkatan F1-Score paling signifikan terlihat pada kelas takut (dari 0,70 menjadi 0,75), senang (dari 0,75 menjadi 0,83), dan sedih (dari 0,74 menjadi 0,77), sementara kelas terkejut tetap mempertahankan performa terbaik dengan F1-Score 0,97. Hal ini menegaskan bahwa penggabungan probabilitas prediksi dari kelima model fold mampu menstabilkan keputusan klasifikasi lintas kelas emosi.



Gambar 5. Confusion Matrix Hasil Pengujian pada Dataset Gabungan dengan Ensemble Learning (Soft Voting)

Gambar 5 memperlihatkan bahwa konsentrasi nilai pada garis diagonal utama meningkat dibandingkan Gambar 4, dengan kesalahan klasifikasi antara kelas jijik dan takut yang berkurang. Sensitivitas kelas terkejut tetap sangat tinggi (115 dari 118 sampel diklasifikasikan dengan tepat), sementara kesalahan klasifikasi antara kelas sedih dan takut, serta antara sedih dan jijik, masih menjadi tantangan utama meskipun jumlahnya berkurang dibandingkan model tunggal.

Evaluasi Kinerja Model pada Dataset Tunggal

Sebagai pembanding, model hibrida turut dievaluasi secara terisolasi pada dataset tunggal RAVDESS dengan modifikasi susunan filter konvolusi (32, 64, 128, 128) serta augmentasi ekstensif sebagaimana dijelaskan pada Tabel 4. Secara makro, sebaran akurasi dan loss model pada setiap fold dataset tunggal disajikan pada Tabel 8.

Tabel 8. Distribusi Nilai Evaluasi Metrik Makro Fase Pelatihan dan Pengujian Pada Dataset Tunggal

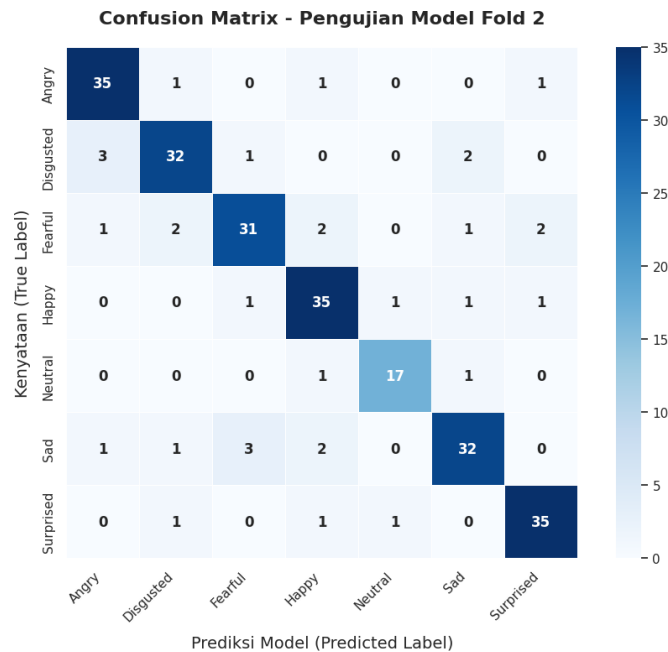
Fold	Training Acc.	Training Loss	Validation Acc.	Validation Loss	Testing Acc.
1	0,9532	0,1315	0,9352	0,2198	0,8600
2	0,9908	0,0261	0,9633	0,1235	0,8700
3	0,9965	0,0109	0,9483	0,2413	0,8500
4	0,9643	0,1006	0,9454	0,1981	0,8700
5	0,9739	0,0775	0,9464	0,2114	0,8200
Mean	0,9757	0,0693	0,9477	0,1988	0,8540

Tabel 8 menunjukkan stabilitas kinerja model yang lebih tinggi dibandingkan dataset gabungan, dengan rata-rata akurasi validasi 94,77% dan rata-rata akurasi pengujian 5-fold sebesar 85,40%. Kesenjangan performa (generalization gap) antara validasi dan pengujian murni tercatat sebesar 0,0937 (9,37%), sedikit lebih tinggi dibandingkan dataset gabungan namun masih tergolong wajar untuk ukuran dataset tunggal. Evaluasi performa secara mikro pada setiap kelas emosi untuk model fold 2 (model performa terbaik pada dataset tunggal) disajikan pada Tabel 9.

Tabel 9. Laporan Matriks Evaluasi Mikro Kelas Emosi pada Model Dataset Tunggal (Model Terbaik - Fold 2)

Kategori Emosi	Precision	Recall	F1-Score	Support
Marah	0,88	0,92	0,90	38
Jijik	0,86	0,84	0,85	38
Takut	0,86	0,79	0,83	39
Senang	0,83	0,90	0,86	39
Netral	0,89	0,89	0,89	19
Sedih	0,86	0,82	0,84	39
Terkejut	0,90	0,92	0,91	38
Accuracy			0,87	250
Macro avg	0,87	0,87	0,87	250
Weighted avg	0,87	0,87	0,87	250

Tabel 9 mengonfirmasi bahwa model pada dataset tunggal memiliki sensitivitas identifikasi yang tinggi dan merata pada seluruh kelas emosi, dengan F1-Score tertinggi pada kelas marah (0,90) dan terkejut (0,91), serta F1-Score terendah pada kelas takut (0,83). Nilai Macro Average F1-Score sebesar 0,87 ini lebih tinggi dibandingkan performa pada dataset gabungan (0,78 untuk model terbaik, 0,82 untuk ensemble), yang mengindikasikan bahwa heterogenitas dataset gabungan tetap menjadi tantangan meskipun model telah mencapai generalisasi yang baik.

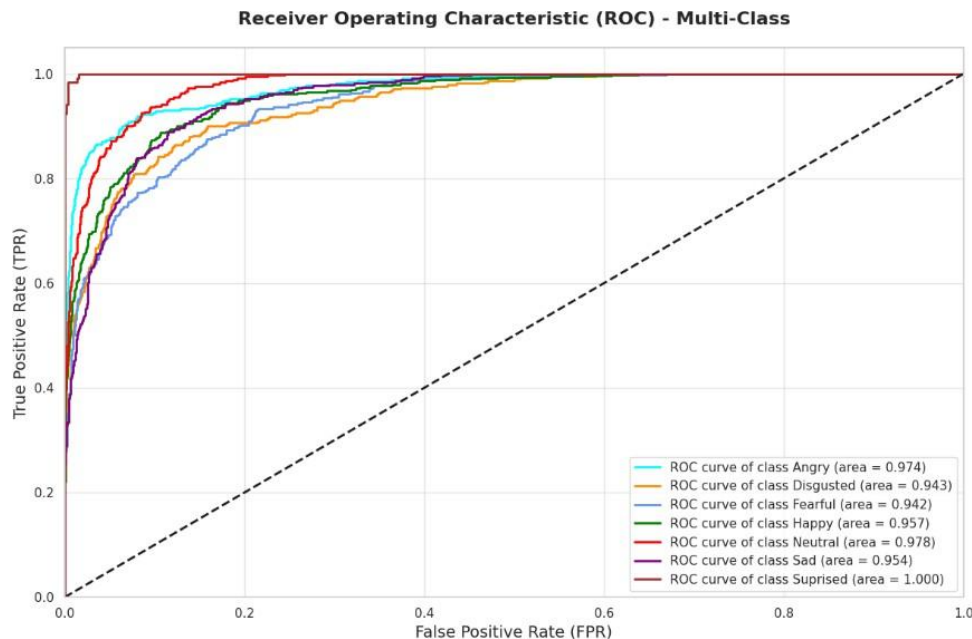


Gambar 6. Confusion Matrix Hasil Pengujian pada Dataset Tunggal (RAVDESS, Fold 2)

Gambar 6 menunjukkan pemusatan nilai yang sangat padat pada garis diagonal utama untuk seluruh kelas emosi pada dataset tunggal. Kesalahan klasifikasi yang tersisa berskala jauh lebih kecil dibandingkan dataset gabungan, dengan pola kebingungan yang masih berkisar pada kelas emosi bernuansa negatif seperti sedih, takut, dan jijik, meskipun dengan kuantitas kesalahan yang jauh lebih sedikit dibandingkan pengujian pada dataset gabungan.

Análisis Curva ROC/AUC

Performa diskriminasi multi-kelas model turut dievaluasi melalui kurva ROC dan nilai AUC sebagaimana ditampilkan pada Gambar 7. Seluruh kategori emosi berhasil melampaui ambang batas nilai AUC 0,90, dengan rincian: emosi terkejut mencatatkan skor AUC sempurna (1,000), diikuti netral (0,978), marah (0,974), senang (0,957), sedih (0,954), jijik (0,943), dan takut (0,942).



Gambar 7. Grafik ROC/AUC Multi-Kelas

Hasil ROC/AUC ini menunjukkan bahwa model hibrida CNN-Transformer memiliki kemampuan diskriminasi antarkelas yang sangat andal meskipun dilatih pada dataset gabungan yang heterogen. Secara keseluruhan, hasil penelitian ini membuktikan bahwa integrasi CNN dan Transformer dengan mekanisme attention efektif dalam menangkap informasi spasial sekaligus dependensi temporal jangka panjang pada sinyal suara, sehingga mampu meningkatkan akurasi dan keandalan sistem pengenalan emosi berbasis suara dibandingkan dengan pendekatan CNN murni pada penelitian-penelitian sebelumnya [24], [33], [14], [15], [16], [17].

4. KESIMPULAN

Berdasarkan hasil analisis data, pengujian, dan pembahasan yang telah diuraikan, penelitian ini menarik beberapa simpulan penting sebagai berikut:

1. Pemodelan sistem pengenalan emosi berbasis suara berhasil dirancang dan dibangun secara sistematis melalui penggabungan blok konvolusi CNN 2D (dengan fungsi aktivasi ReLU, Batch Normalization, Max Pooling, dan Dropout) yang diintegrasikan dengan blok Transformer Encoder (8 heads attention dan dimensi kunci penyelarasan 128).
2. Pengujian model hibrida usulan terhadap kelompok data testing membuktikan tingkat keberhasilan klasifikasi yang tangguh dan optimal. Pada dataset gabungan, rata-rata akurasi pengujian 5-fold cross validation mencapai 75,80%, sedangkan model performa terbaik (fold tunggal) mencapai akurasi pengujian 77,00% dengan Macro Average F1-Score sebesar 78,00%.
3. Model hibrida menunjukkan performa diskriminasi multi-kelas yang sangat andal, dibuktikan oleh visualisasi kurva ROC-AUC di mana seluruh kategori emosi

berhasil melampaui ambang batas nilai 0,90, dengan emosi terkejut mencatatkan skor AUC sempurna (1,000).

4. Kestabilan arsitektur dibuktikan oleh kesenjangan generalisasi antara rata-rata akurasi validasi silang (79,65%) dan rata-rata akurasi pengujian murni (75,80%) yang sangat tipis, yakni hanya sebesar 3,85%, sehingga membuktikan bahwa model memiliki kemampuan generalisasi yang kuat dan terbebas dari gejala overfitting akut.
5. Penerapan Ensemble Learning (Soft Voting) dari kelima model fold mampu meningkatkan akurasi pengujian menjadi 81,00% dengan Macro Average F1-Score sebesar 0,82, menegaskan bahwa hasil ensemble ini merupakan capaian akhir tertinggi penelitian sekaligus potensi peningkatan performa lebih lanjut melalui strategi penggabungan model.

Untuk pengembangan penelitian selanjutnya, disarankan eksplorasi dataset audio wicara spontan dunia nyata dengan kondisi lingkungan berderau heterogen, penerapan arsitektur berbasis Self-Supervised Learning seperti Wav2Vec 2.0 atau HuBERT sebagai backbone ekstraktor fitur, pengembangan ke arah sistem pengenalan emosi multimodal yang mengintegrasikan fitur visual dan teks, serta optimalisasi kecepatan komputasi melalui kuantisasi model (TensorFlow Lite) agar dapat diimplementasikan pada perangkat dengan sumber daya terbatas.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Ilmu Komputer, Fakultas Teknik dan Kejuruan, Universitas Pendidikan Ganesha atas dukungan fasilitas dan bimbingan akademik yang telah diberikan selama penelitian ini berlangsung.

DAFTAR PUSTAKA

- [1] H. A. Elfenbein, "Emotion in Organizations: Theory and Research," *Annual Review of Psychology*, vol. 74, pp. 489–517, 2022, doi: 10.1146/annurev-psych-032720.
- [2] U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte, "Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets," *Engineering Applications of Artificial Intelligence*, vol. 130, 2024, doi: 10.1016/j.engappai.2023.107708.
- [3] F. P. Anindya, D. E. Herwindiati, and N. J. Perdana, "Pengenalan Suara Manusia Menggunakan Support Vector Classifier (SVC) untuk Proses Otentikasi," *Computatio: Journal of Computer Science and Information Systems*, vol. 7, no. 1, 2023.
- [4] T. B. Putri, S. Saidah, B. Hidayat, F. Qothrunnada, and D. Darwindra, "Deteksi Emosi Berdasarkan Sinyal Suara Manusia Menggunakan Discrete Wavelet Transform (DWT) dengan Klasifikasi *Support Vector Machine* (SVM)," *Jurnal Ilmu Komputer dan Informatika*, vol. 3, no. 1, pp. 1–10, 2023, doi: 10.54082/jiki.45.

-
- [5] R. Meiyanti, "Pendeteksi Pengenalan Emosi pada Manusia Menggunakan Hidden Markov Model dan Bidirectional Associative Memory dengan Suara," *Jurnal TIKA*, vol. 6, pp. 231–237, 2021, doi: 10.51179/tika.v6i03.756.
- [6] M. Hosain, M. Yeasmin Arafat, G. Zahirul Islam, J. Uddin, M. Mobarak Hossain, and F. Alam, "Emotional Expression Detection in Spoken Language Employing Machine Learning Algorithms," n.d.
- [7] E. H. Houssein, A. Hammad, and A. A. Ali, "Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review," *Neural Computing and Applications*, vol. 34, no. 15, pp. 12527–12557, 2022, doi: 10.1007/s00521-022-07292-4.
- [8] I. M. A. Wirawan, K. Y. E. Aryanto, I. N. Sukajaya, N. N. M. Agustini, and D. A. W. M. Putri, "Hybrid method for optimizing emotion recognition models on electroencephalogram signals," *IAES International Journal of Artificial Intelligence*, vol. 14, no. 3, pp. 2302–2314, 2025, doi: 10.11591/ijai.v14.i3.pp2302-2314.
- [9] I. M. A. Wirawan, R. Wardoyo, and D. Lelono, "The challenges of emotion recognition methods based on electroencephalogram signals: A literature review," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1508–1519, 2022, doi: 10.11591/ijece.v12i2.pp1508-1519.
- [10] I. M. A. Wirawan and I. M. A. Indrawan, "Subject Independent Emotion Recognition Using Electroencephalogram Signals with Continuous Capsule Network Method," *JUITA*, vol. 13, no. 1, pp. 37–46, 2025.
- [11] I. M. A. Wirawan, I. G. M. Darmawiguna, and I. B. N. Pascima, "Improving baseline reduction for emotion recognition based on electroencephalogram signals," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 4, pp. 4263–4272, 2024, doi: 10.11591/ijai.v13.i4.pp4263-4272.
- [12] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "*Speech Emotion Recognition* through Hybrid Features and Convolutional Neural Network," *Applied Sciences (Switzerland)*, vol. 13, no. 8, 2023, doi: 10.3390/app13084750.
- [13] A. Ashraf, T. S. Gunawan, F. Arifin, M. Kartiwi, A. Sophian, and M. H. Habaebi, "On the Audio-Visual Emotion Recognition using Convolutional Neural Networks and Extreme Learning Machine," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 684–697, 2022, doi: 10.52549/ijeei.v10i3.3879.
- [14] K. Chauhan, K. K. Sharma, and T. Varma, "*Speech Emotion Recognition* Using Convolution Neural Networks," in *Proc. Int. Conf. Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 1176–1181, doi: 10.1109/ICAIS50930.2021.9395844.
- [15] F. Jonatan Tanudjaja, E. Y. Puspaningrum, and V. Via, "Klasifikasi Jenis Emosi Melalui Ucapan Menggunakan Metode Convolutional Neural Network," *Teknologi: Jurnal Ilmiah Sistem Informasi*, vol. 13, no. 2, pp. 1–11, 2023, doi: 10.26594/teknologi.v13i1.3740.
- [16] R. Nurcahyo and M. Iqbal, "Speaker Emotion Recognition Using Convolutional Neural Networks," *Jurnal RESTI*, vol. 6, no. 1, pp. 115–122, 2022, doi: 10.29207/resti.v6i1.3726.

- [17] T. Praveen Blessington, D. Hadke, S. Chavan, P. Chavan, and P. Ingulkar, "Recognition of Emotion through Speech Using Deep Convolution and LSTM," *International Journal of Creative Research Thoughts*, vol. 11, 2023.
- [18] Y. Zhao and X. Shu, "Speech emotion analysis using *Convolutional Neural Network* (CNN) and gamma classifier-based error correcting output codes (ECOC)," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-47118-4.
- [19] F. Kasyidi, R. Ilyas, and N. M. Annisa, "Peningkatan Kemampuan Pengenalan Emosi Melalui Suara dalam Bahasa Indonesia," *MIND Journal*, vol. 6, no. 2, pp. 194–204, 2021, doi: 10.26760/mindjournal.v6i2.194-204.
- [20] C. N. Govardhini, A. Venu, K. Lakshminarayana, J. S. M. Kumar, and A. K. Rout, "Emotion Recognition Using Deep Learning Techniques for User Experience," *International Journal of Research Publication and Reviews*, vol. 4, no. 10, pp. 2644–2653, 2023, doi: 10.55248/gengpi.4.1023.102819.
- [21] R. Rastogi, T. Anand, S. K. Sharma, and S. Panwar, "Emotion Detection via Voice and Speech Recognition," *International Journal of Cyber Behavior, Psychology and Learning*, vol. 13, no. 1, 2023, doi: 10.4018/IJCBPL.333473.
- [22] A. Yazdani, H. Simchi, and Y. Shekofteh, "*Speech Emotion Recognition* Based on Deep Learning," in *Proc. 2021 11th Int. Conf. Computer Engineering and Knowledge (ICCKE)*, IEEE, 2021.
- [23] D. K. R. P. Wisnu, I. K. Purnamawan, and L. J. E. Dewi, "Analisis Komparasi Performa Arsitektur VGG16 dan CNN Kustom pada Klasifikasi Citra Buah Melon Berdasarkan Gambar Kulit," *Jurnal Informatika Teknologi dan Sains (JINTEKS)*, vol. 8, no. 1, pp. 438–448, 2026, doi: 10.51401/jinteks.v8i1.
- [24] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "*Speech Emotion Recognition* Using Convolutional Neural Networks with Attention Mechanism," *Electronics (Switzerland)*, vol. 12, no. 20, 2023, doi: 10.3390/electronics12204376.
- [25] R. Ullah et al., "*Speech Emotion Recognition* Using Convolution Neural Networks and Multi-Head Convolutional Transformer," *Sensors*, vol. 23, no. 13, 2023, doi: 10.3390/s23136212.
- [26] G. Akilandasowmya and M. Mayuranathan, "Enhancing speech-based emotion recognition with transformer models and attention mechanisms," *International Journal of Speech Technology*, vol. 29, no. 3, p. 55, 2026, doi: 10.1007/s10772-026-10280-x.
- [27] H. Barik and R. Sridhar, "Inception-based speech emotion recognition," *Multimedia Tools and Applications*, vol. 85, no. 7, p. 603, 2026, doi: 10.1007/s11042-026-21760-2.
- [28] A. Chakhtouna, S. Sekkate, and A. Adib, "Enhancing speech emotion recognition using statistical self-supervised embeddings," *Multimedia Tools and Applications*, vol. 85, no. 8, p. 634, 2026, doi: 10.1007/s11042-026-21812-7.
- [29] N. Banerjee, N. Sethi, and S. Borah, "Deep analysis of MFCC and MEL spectrogram features to recognize and classify stuttered speech," *Multimedia Tools and*

-
- Applications, vol. 84, no. 35, pp. 43827–43846, 2025, doi: 10.1007/s11042-025-20881-4.
- [30] R. N. Bashir, M. A. Shahid, T. Rashid, M. Faheem, T. Saidani, O. Saidani, and A. R. Khan, "Voice pathology identification using mel spectrogram features and deep learning," *Signal, Image and Video Processing*, vol. 19, no. 11, p. 909, 2025, doi: 10.1007/s11760-025-04527-4.
- [31] N. K. A. Artalia, I. M. D. Maysanjaya, and P. H. Suputra, "Sentiment Analysis of Bali's Rapid Tourism Growth Using IndoBERT with INSET Lexicon Automatic Labelling," vol. 15, no. 1, 2026.
- [32] N. L. P. R. Dewi, I. N. S. W. Wijaya, I. K. Purnamawan, and N. W. Marti, "Model Classifier Judul Berita Pariwisata Indonesia Berdasarkan Sentimen," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 117–124, 2024, doi: 10.25126/jtiik.20241117617.
- [33] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets," *Electronics (Switzerland)*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223831.