JITSI Vol 1 No. 1 (2026)
Journal of Information Technology and System Integration
https://journal2.ummat.ac.id/index.php/jitsi

KOMPARASI METODE SVM DAN CNN UNTUK KLASIFIKASI EMOSI SUARA BERDASARKAN FITUR SPEKTRAL DAN PROSODI

Hamida[1*], Annisa Hardianti Faradela[2], Andriany[3], Fatiya Qurrotu'Aini[4]
[1,2]Sistem & Teknologi Informasi, Universitas Muhammadiyah Mataram
[3,4] Sistem & Teknologi Informasi, Universitas Muhammadiyah Mataram
email: xxx@xxxxx.com[1*]

Abstrak: Speech Emotion Recognition (SER) merupakan bidang penelitian yang bertujuan mengidentifikasi kondisi emosional pembicara berdasarkan karakteristik akustik suara. Penelitian ini membandingkan performa Support Vector Machine (SVM) dan Convolutional Neural Network (CNN) dalam klasifikasi emosi suara menggunakan kombinasi fitur prosodi dan fitur spektral pada dataset Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Tahapan penelitian meliputi preprocessing sinyal audio, ekstraksi fitur prosodi berupa pitch, RMS energy, dan zero crossing rate, serta ekstraksi fitur spektral menggunakan Mel-Frequency Cepstral Coefficients (MFCC). Pada model SVM, fitur direpresentasikan dalam bentuk vektor numerik yang diringkas menggunakan statistik agregat, sedangkan pada model CNN digunakan arsitektur dual-input yang memproses representasi spektral MFCC dan fitur prosodi secara terpisah. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model SVM memperoleh accuracy sebesar 76,21%, precision sebesar 0,7629, recall sebesar 0,7621, dan F1-score sebesar 0,7596. Sementara itu, model CNN memperoleh accuracy sebesar 73,61%, precision sebesar 0,7392, recall sebesar 0,7361, dan F1-score sebesar 0,7339. Hasil penelitian menunjukkan bahwa SVM memberikan performa klasifikasi yang lebih baik dibandingkan CNN pada kombinasi fitur yang digunakan. Temuan ini mengindikasikan bahwa representasi fitur prosodi dan MFCC dalam bentuk vektor statistik masih efektif untuk klasifikasi emosi suara pada dataset RAVDESS.

Kata Kunci : Speech Emotion Recognition, Support Vector Machine, Convolutional Neural Network, MFCC, Prosodi, RAVDESS

PENDAHULUAN

Suara manusia merupakan salah satu media komunikasi yang digunakan untuk menyampaikan kepada individu lain. Suara terbentuk dari sinyal yang dipengaruhi oleh waktu dan frekuensi dalam bentuk sinyal disk informasi rit serta memiliki karakteristik vokal yang berbeda pada setiap individu, seperti nada, warna suara, dan volume [1]. Selain sebagai media penyampaian informasi verbal, suara juga dapat merepresentasikan kondisi emosional seseorang saat berbicara. Emosi merupakan salah satu unsur penting dalam komunikasi dan interaksi sosial karena membantu individu menyampaikan kondisi afektif serta memahami respons lawan bicara [2]. Emosi dapat diekspresikan secara verbal maupun nonverbal, dan suara menjadi salah satu media yang paling umum digunakan dalam penyampaian informasi emosional. Ucapan yang mengandung emosi umumnya dapat dikenali melalui perubahan karakteristik akustik dibandingkan dengan ucapan dalam kondisi *netral* [3]. Kondisi tersebut menunjukkan bahwa suara memiliki peran penting dalam menyampaikan informasi emosional dalam proses komunikasi manusia [4].

Perubahan kondisi emosional seseorang dapat memengaruhi karakteristik vokal yang dihasilkan, seperti intonasi, kecepatan bicara, dan tekanan suara [5]. Variasi karakteristik tersebut menyebabkan setiap emosi memiliki pola akustik yang khas sehingga dapat dianalisis melalui pendekatan pengolahan sinyal suara [6]. Dibandingkan dengan ekspresi wajah atau bahasa tubuh, suara dinilai lebih mudah diperoleh dan dapat digunakan dalam berbagai kondisi tanpa memerlukan kontak visual secara langsung [7]. Perkembangan teknologi pengolahan sinyal digital juga memungkinkan analisis suara dilakukan secara otomatis untuk mengenali emosi seseorang. Kemampuan sistem dalam mengidentifikasi emosi berdasarkan sinyal suara mendorong berkembangnya penelitian di bidang *Speech Emotion Recognition* (SER) [8]. Penelitian mengenai SER terus berkembang karena memiliki potensi penerapan pada berbagai bidang, seperti layanan pelanggan, pendidikan, kesehatan mental, dan interaksi manusia–komputer [9].

Seiring dengan perkembangan metode machine learning dan deep learning, sistem SER mengalami peningkatan performa dalam proses klasifikasi emosi suara. Penelitian terdahulu telah menggunakan berbagai metode untuk meningkatkan akurasi pengenalan emosi suara [10]. Penelitian yang dilakukan oleh [11] menggunakan metode *Support Vector Machine* (SVM) dan *Multi Layer Perceptron* (MLP) untuk klasifikasi emosi suara menggunakan fitur MFCC, MEL, Chroma, dan Tonnetz. Hasil penelitian menunjukkan bahwa metode tersebut mampu menghasilkan akurasi sebesar 86,5% dalam proses pengenalan emosi suara. Penelitian yang dilakukan oleh [12] menggunakan metode Support Vector Machine (SVM) untuk klasifikasi emosi suara dengan memanfaatkan fitur prosodik dan spektral, yaitu pitch, intensity, dan Mel-Frequency Cepstral Coefficients (MFCC). Hasil penelitian menunjukkan bahwa metode Support Vector Machine (SVM) menghasilkan akurasi sebesar 77,86% dalam proses pengenalan emosi suara. Selain itu, penelitian lain yang dilakukan oleh [13] membandingkan metode Support Vector Machine (SVM), Convolutional Neural Network (CNN), dan Long Short-Term Memory (LSTM). Hasil penelitian menunjukkan bahwa SVM memperoleh kinerja terbaik

dengan akurasi sebesar 75%, sedangkan CNN dan LSTM masing-masing mencapai 50% dan 25%. Penelitian tersebut menggunakan empat kategori emosi, yaitu anger, sadness, happiness, dan neutral.

Selain pendekatan berbasis machine learning klasik seperti SVM, penelitian dalam bidang SER juga mulai banyak mengadopsi pendekatan berbasis pembelajaran mendalam (deep learning), khususnya CNN, untuk meningkatkan performa klasifikasi emosi suara. Penelitian yang dilakukan oleh [14] menerapkan metode CNN pada proses klasifikasi emosi suara menggunakan dataset RAVDESS dan MELD. Hasil penelitian menunjukkan bahwa metode CNN mampu menghasilkan akurasi sebesar 94,0% pada dataset RAVDESS dan 91,9% pada dataset MELD. Selain itu, penelitian yang dilakukan oleh [15] menerapkan model *dual-stream hybrid deep learning* dengan kombinasi fitur prosodi, CNN, BiGRU, dan *Multi-Head Attention* untuk klasifikasi emosi suara. Hasil penelitian menunjukkan bahwa metode tersebut mampu menghasilkan akurasi sebesar 97,64% pada dataset RAVDESS.

Berdasarkan penelitian terdahulu, penelitian ini mengusulkan komparasi antara metode CNN dan SVM untuk klasifikasi emosi suara dengan menggunakan kombinasi fitur prosodi dan spektral. Kebaruan penelitian ini terletak pada komparasi sistematis antara CNN dan SVM menggunakan kombinasi fitur prosodi dan spektral secara terintegrasi untuk klasifikasi tujuh kelas emosi dalam satu kerangka eksperimen yang seragam, yang belum secara eksplisit dilakukan pada penelitian sebelumnya.

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja metode CNN dan SVM dalam klasifikasi emosi suara berdasarkan fitur prosodi dan spektral. Selain itu, penelitian ini bertujuan untuk mengimplementasikan dan menguji model pada tujuh kategori emosi, yaitu *calm, happy, sad, angry, fearful, disgust,* dan *surprise*, guna mengetahui performa masing-masing metode dalam proses pengenalan emosi suara.

TINJAUAN PUSTAKA

Speech Emotion Recognition (SER) merupakan teknologi yang digunakan untuk mengenali kondisi emosional seseorang berdasarkan karakteristik akustik suara yang dihasilkan saat berbicara. Teknologi ini memanfaatkan berbagai fitur akustik untuk mengidentifikasi emosi, seperti senang, sedih, marah, takut, terkejut, dan netral. Perkembangan machine learning dan deep learning telah mendorong peningkatan performa sistem SER pada berbagai aplikasi, antara lain interaksi manusia–komputer, layanan pelanggan, serta pemantauan kesehatan mental berbasis suara. Penelitian oleh [16] menjelaskan bahwa analisis emosi melalui suara dapat memberikan informasi emosional secara efektif dan bersifat non-invasif tanpa memerlukan interaksi langsung dengan pengguna, sehingga berpotensi mendukung deteksi dini gangguan kesehatan mental.

Berbagai penelitian telah memanfaatkan dataset RAVDESS untuk menguji performa model klasifikasi emosi. Penelitian oleh [17] menunjukkan bahwa *Convolutional Neural Network* (CNN) mampu menghasilkan performa tinggi pada dataset tersebut, sedangkan Penelitian oleh [18] mengembangkan model CNN-BiLSTM yang memberikan hasil lebih baik melalui kombinasi ekstraksi fitur lokal dan pemodelan urutan data secara global. Penelitian lain oleh [19] menerapkan pendekatan CNN-LSTM, sementara [20] menggunakan arsitektur *Dual-Layer LSTM* untuk meningkatkan kemampuan model dalam menangkap pola temporal pada sinyal suara. Temuan-temuan tersebut secara konsisten menunjukkan bahwa RAVDESS merupakan dataset yang representatif untuk mengevaluasi berbagai metode klasifikasi emosi berbasis suara.

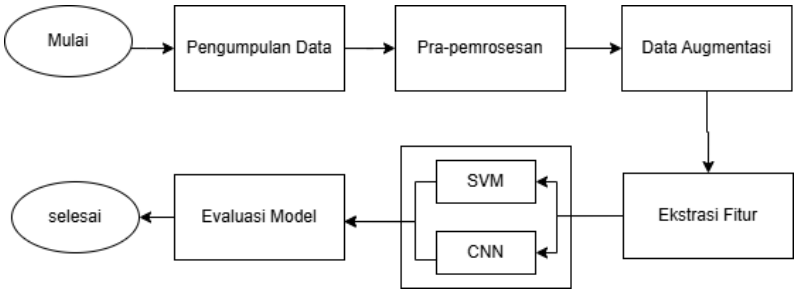
Tahap ekstraksi fitur memegang peranan penting dalam SER karena kualitas fitur yang diekstraksi sangat memengaruhi akurasi klasifikasi. Fitur akustik pada umumnya dikelompokkan menjadi dua kategori, yaitu fitur spektral, seperti *Mel Frequency Cepstral Coefficients* (MFCC), dan fitur prosodi, seperti *pitch*, *Root Mean Square* (RMS), dan *Zero Crossing Rate* (ZCR). Penelitian ini secara khusus menggunakan kombinasi MFCC sebagai representasi fitur spektral bersama *pitch*, RMS, dan ZCR sebagai representasi fitur prosodi. Penelitian oleh [21] menunjukkan bahwa *pitch* dan energi suara memiliki hubungan yang kuat dengan perubahan emosi pembicara, sedangkan Penelitian oleh [22] menemukan bahwa kombinasi fitur prosodi dan spektral secara umum mampu meningkatkan kemampuan model dalam membedakan berbagai kategori emosi dibandingkan dengan penggunaan satu jenis fitur saja. Temuan ini diperkuat oleh hasil penelitian [23], yang menunjukkan bahwa kombinasi MFCC, ZCR, RMSE, dan *pitch* pada model berbasis CNN mampu mencapai akurasi klasifikasi emosi suara sekitar 88% pada dataset berbahasa Indonesia, membuktikan bahwa kombinasi fitur yang relatif sederhana ini tetap kompetitif tanpa memerlukan fitur spektral tambahan seperti *Mel-Spectrogram* atau *Chroma STFT*. Berdasarkan temuan tersebut, penggunaan MFCC bersama fitur prosodi *pitch*, RMS, dan ZCR pada penelitian ini didasarkan pada pertimbangan efisiensi komputasi tanpa mengorbankan akurasi klasifikasi secara signifikan.

Dalam proses klasifikasi, *Support Vector Machine* (SVM) dan CNN merupakan dua metode yang paling banyak digunakan. SVM memiliki kemampuan yang baik dalam menangani data berdimensi tinggi sehingga tetap relevan digunakan dalam penelitian SER hingga saat ini, sebagaimana ditunjukkan oleh penelitian [24],[25], serta Penelitian oleh [26], yang memperoleh performa klasifikasi cukup baik melalui berbagai kombinasi fitur akustik pada dataset RAVDESS. Di sisi lain, CNN menjadi metode deep learning yang populer karena kemampuannya mengekstraksi fitur secara otomatis dari data suara tanpa memerlukan rekayasa fitur yang kompleks. Penelitian oleh [27], [28], [29] menunjukkan bahwa CNN dan berbagai pengembangannya mampu menghasilkan performa tinggi pada klasifikasi emosi suara. Berdasarkan

berbagai temuan tersebut, penggunaan dataset RAVDESS dengan kombinasi fitur MFCC, *pitch*, RMS, dan ZCR, serta perbandingan metode CNN dan SVM, masih relevan untuk dilakukan guna mengidentifikasi metode yang lebih efektif dalam klasifikasi emosi suara.

METODE

Tahapan penelitian terdiri atas proses pengumpulan data, pra-pemrosesan, data augmntasi, ekstraksi fitur, pelatihan model, dan evaluasi model. Adapun alur penelitian ditunjukkan pada gambar 1.



Gambar 1. Alur Penelitian

3.1 Pengumpulan Data

Penelitian ini menerapkan pendekatan kuantitatif eksperimental dengan tujuan membandingkan kinerja metode *Support Vector Machine* (SVM) dan *Convolutional Neural Network* (CNN) dalam melakukan klasifikasi emosi suara berdasarkan kombinasi fitur spektral dan prosodi. Dataset yang digunakan diperoleh dari platform Kaggle, yaitu *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS). Berdasarkan spesifikasi aslinya, dataset tersebut terdiri atas 1.440 berkas audio berformat .WAV 16-bit dengan frekuensi pengambilan sampel (*sampling rate*) sebesar 48.000 Hz, yang dihasilkan dari 60 sesi perekaman oleh 24 aktor profesional terdiri atas 12 aktor pria dan 12 aktor wanita. Setiap aktor mengucapkan dua pernyataan yang memiliki kesesuaian leksikal menggunakan aksen Amerika Utara yang netral. Penelitian ini memanfaatkan tujuh kategori emosi sebagai target klasifikasi, yaitu *calm*, *happy*, *sad*, *angry*, *fearful*, *disgust*, dan *surprise*.

3.2 Preprocessing Audio

Tahap pra-pemrosesan bertujuan meningkatkan kualitas sinyal audio dan menghasilkan data yang konsisten sebelum proses ekstraksi fitur dilakukan. Tahap ini diterapkan pada seluruh rekaman audio dataset RAVDESS guna memastikan setiap sampel memiliki kualitas dan dimensi yang seragam sebagai masukan pada tahap selanjutnya.

Tahap pertama adalah *trimming silence*, yang dilakukan untuk menghilangkan bagian awal dan akhir audio yang tidak mengandung informasi emosi, menggunakan ambang batas amplitudo sebesar 20 dB. Penentuan durasi target sinyal dilakukan berdasarkan persentil ke-95 dari distribusi durasi audio pascatrimming, bukan nilai rata-rata, karena rata-rata dinilai tidak mampu mengakomodasi variasi panjang sinyal secara memadai sehingga berisiko menghilangkan informasi penting akibat pemotongan berlebih pada sampel berdurasi panjang. Pendekatan ini menghasilkan durasi target sebesar 2,44 detik, atau setara 53.802 sampel pada frekuensi sampling 22.050 Hz. Sinyal yang lebih pendek dari durasi target tersebut diberi *zero-padding* pada bagian akhir, sedangkan sinyal yang melebihi durasi tersebut dipotong dari ujungnya, sehingga seluruh sampel memiliki dimensi input yang seragam.

Selain penyeragaman format dan kualitas sinyal, kelas emosi *neutral* juga tidak diikutsertakan dalam proses klasifikasi. Kelas tersebut dihapus karena secara akustik sulit dibedakan dari kelas *calm*, sehingga kehadirannya berpotensi meningkatkan ambiguitas dalam proses pengenalan emosi [17].

3.3 Data Augementasi

Sebelum proses augmentasi, dataset yang telah melalui tahap seleksi dan pembersihan kelas dibagi menggunakan rasio 80% untuk data latih dan 20% untuk data uji secara *stratified* berdasarkan kelas emosi, guna memastikan proporsi setiap kelas tetap seimbang pada kedua subset.

Untuk meningkatkan jumlah data pelatihan serta memperkaya variasi karakteristik sinyal suara, diterapkan lima teknik augmentasi data pada level sinyal audio mentah, yaitu *noise injection* dengan faktor 0,005, *pitch shift* naik sebesar +2 *semitone*, *pitch shift* turun sebesar −2 *semitone*, *time stretch* dipercepat dengan faktor 1,15, dan *time stretch* diperlambat dengan faktor 0,85. Augmentasi diterapkan pada level sinyal mentah, sebelum proses ekstraksi fitur dilakukan, agar variasi akustik yang dihasilkan tetap konsisten dengan karakteristik alami suara manusia, berbeda dengan augmentasi pada level fitur yang berpotensi menghasilkan representasi tidak realistis.

Penerapan kelima teknik tersebut menghasilkan lima sampel baru dari setiap data audio asli, sehingga setiap data menghasilkan enam representasi yang berbeda secara akustik namun tetap mempertahankan label emosi yang sama. Augmentasi dilakukan secara proporsional pada seluruh kelas emosi untuk mempertahankan keseimbangan distribusi data, dan hanya diterapkan pada data latih agar evaluasi pada data uji tetap mencerminkan data asli tanpa rekayasa tambahan. Seluruh hasil augmentasi diseragamkan pada durasi target yang telah ditetapkan pada tahap *preprocessing*, melalui mekanisme *zero-padding* atau pemotongan yang sama seperti pada sampel *original*.

3.4 Ekstraksi Fitur

Tahap berikutnya adalah ekstraksi fitur, yang bertujuan mengonversi sinyal suara menjadi representasi numerik yang dapat digunakan pada tahap klasifikasi. Fitur yang digunakan dalam penelitian ini terdiri atas dua kelompok, yaitu fitur prosodi dan fitur spektral. Fitur prosodi digunakan untuk merepresentasikan karakteristik intonasi, energi, dan pola pengucapan, sedangkan fitur spektral digunakan untuk merepresentasikan karakteristik frekuensi dari sinyal suara.

Fitur prosodi diekstraksi dari tiga parameter akustik, yaitu frekuensi dasar (*pitch*) yang dihitung menggunakan algoritma PYIN, energi sinyal yang direpresentasikan melalui *Root Mean Square* (RMS), serta *Zero Crossing Rate* (ZCR). Ketiga parameter tersebut diringkas menjadi 11 fitur statistik, terdiri atas nilai rata-rata, standar deviasi, minimum, maksimum, dan *jitter* untuk *pitch*; nilai rata-rata, standar deviasi, maksimum, dan *shimmer* untuk RMS; serta nilai rata-rata dan standar deviasi untuk ZCR. Kesebelas fitur prosodi ini bersifat identik dan digunakan secara setara pada kedua model, baik SVM maupun CNN, setelah dinormalisasi menggunakan *StandardScaler*.

Fitur spektral diekstraksi menggunakan metode *Mel Frequency Cepstral Coefficients* (MFCC) dengan parameter N_MFCC sebanyak 13 koefisien, N_FFT sebesar 2.048, dan *hop length* sebesar 512 sampel. Representasi fitur spektral ini selanjutnya diproses secara berbeda untuk masing-masing model, mengikuti kebutuhan struktural algoritmanya. Pada model SVM, setiap koefisien MFCC diringkas menggunakan nilai rata-rata dan standar deviasi, sehingga menghasilkan 26 fitur spektral yang kemudian digabungkan dengan 11 fitur prosodi, membentuk satu vektor fitur berdimensi tetap sebanyak 37 dimensi sebagai masukan model SVM.

Pada model CNN, fitur spektral dipertahankan dalam bentuk matriks dua dimensi berukuran (13, 107, 1) tanpa diringkas menjadi statistik agregat, dengan nilai 13 menyatakan jumlah koefisien MFCC dan 107 menyatakan jumlah *frame* temporal hasil penyeragaman durasi pada tahap *preprocessing*, sehingga informasi temporal sinyal tetap terjaga. Fitur prosodi pada model CNN diproses melalui cabang *dense* terpisah yang kemudian digabungkan dengan keluaran cabang konvolusi sebelum lapisan klasifikasi akhir, sehingga model menerima dua jenis masukan secara simultan, yaitu representasi spektral berbentuk matriks dan vektor prosodi.

Perbedaan mendasar antara representasi fitur untuk SVM dan CNN terletak pada cara informasi temporal dikelola, bukan pada sumber data yang digunakan, karena kedua model menggunakan parameter ekstraksi MFCC yang identik. Pada SVM, informasi temporal diringkas menjadi statistik agregat yang sesuai dengan kebutuhan struktural algoritma berbasis *hyperplane*, sedangkan pada CNN, informasi temporal dipertahankan secara eksplisit dalam struktur matriks sehingga model dapat mempelajari pola dinamis yang terkandung dalam perubahan karakteristik akustik sepanjang durasi ucapan, sesuai dengan keunggulan struktural CNN dalam menangkap pola spasial dan temporal.

3.5 SVM

Pada model SVM, data fitur terlebih dahulu dinormalisasi menggunakan StandardScaler agar setiap variabel berada pada skala yang seragam. Normalisasi ini dilakukan untuk mengurangi pengaruh perbedaan rentang nilai antarfitur, sehingga proses pembelajaran model dapat berlangsung lebih optimal. Secara matematis, normalisasi StandardScaler dinyatakan sesuai dengan persamaan berikut [30].

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Keterangan: Pada Persamaan (1), z merupakan nilai fitur setelah dinormalisasi dan x merupakan nilai fitur asli. Parameter μ menyatakan nilai rata-rata dari fitur, sedangkan σ menyatakan nilai standar deviasi dari fitur yang digunakan dalam proses normalisasi StandardScaler.

Setelah proses normalisasi dilakukan, data fitur yang telah berada pada skala seragam digunakan sebagai masukan pada model Support Vector Machine (SVM). SVM bekerja dengan mencari hyperplane terbaik yang mampu memisahkan data ke dalam kelas emosi yang berbeda secara optimal. Pemisahan tersebut dilakukan dengan memaksimalkan margin antara hyperplane dan data terdekat dari setiap kelas [31]. Secara matematis, persamaan SVM dapat dinyatakan melalui persamaan berikut [32].

$$f(x) = \sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b \quad (2)$$

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (3)$$

Keterangan: Pada persamaan (2), $f(x)$adalah fungsi keputusan SVM, α_i adalah koefisien Lagrange, y_i adalah label kelas data latih, $K(x_i, x)$adalah fungsi kernel, x_i adalah vektor fitur data latih, x adalah vektor fitur data masukan, dan ba dalah bias. Pada persamaan (3), $K(x_i, x_j)$ menyatakan tingkat kemiripan antara dua vektor fitur, sedangkan γ mengatur pengaruh jarak antar data terhadap nilai kernel.

3.6 CNN

Convolutional Neural Network (CNN) digunakan sebagai model klasifikasi untuk mengenali emosi berdasarkan representasi spektral sinyal suara. CNN dipilih karena mampu mengekstraksi fitur secara otomatis dari representasi dua dimensi, seperti MFCC dan spektrogram, sehingga banyak diterapkan pada sistem *Speech Emotion Recognition* [33]. Model yang digunakan menerapkan arsitektur dual-input, yang terdiri atas citra fitur spektral (MFCC, Delta MFCC, dan Delta-Delta MFCC) sebagai masukan utama serta vektor fitur prosodi sebagai masukan pendukung. Cabang citra spektral terdiri atas empat lapisan Conv2D dengan jumlah filter 32, 64, 128, dan 256 menggunakan kernel berukuran 3×3, yang masing-masing diikuti oleh Batch Normalization, MaxPooling2D, dan Dropout. Keluaran lapisan konvolusi diringkas menggunakan Global Average Pooling dan diproses oleh lapisan Dense berukuran 128 neuron. Sementara itu, cabang prosodi memproses 11 fitur menggunakan dua lapisan Dense berukuran 32 dan 16 neuron dengan fungsi aktivasi ReLU.

Representasi fitur dari kedua cabang kemudian digabungkan menggunakan operasi concatenate dan diproses melalui lapisan Dense berukuran 128 neuron, Dropout sebesar 0,30, serta lapisan keluaran Softmax untuk menghasilkan probabilitas tujuh kelas emosi. Model dilatih menggunakan optimizer Adam dengan *learning rate* sebesar 0,001 dan fungsi kerugian *categorical cross-entropy*. Proses pelatihan dilakukan selama maksimum 100 *epoch* dengan *batch size* 32 dan *validation split* sebesar 15%, serta menerapkan Early Stopping dan ReduceLROnPlateau untuk meningkatkan stabilitas pelatihan. Arsitektur ini diterapkan pada tiga konfigurasi masukan, yaitu CNN Tingkat 1 (MFCC), CNN Tingkat 2 (MFCC + Delta MFCC), dan CNN Tingkat 3 (MFCC + Delta MFCC + Delta-Delta MFCC), dengan arsitektur jaringan dan parameter pelatihan yang sama. Perbedaan ketiga konfigurasi hanya terletak pada jumlah kanal fitur spektral yang digunakan sebagai masukan.

3.7 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja CNN dan SVM dalam mengklasifikasikan emosi suara. Pada penelitian ini, metrik evaluasi yang digunakan meliputi *accuracy, macro-F1,* dan *weighted-F1*. Ketiga metrik tersebut digunakan sebagai dasar untuk membandingkan performa kedua model secara objektif. Secara umum, metrik evaluasi tersebut dinyatakan melalui persamaan berikut.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$F_1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Keterangan: Pada persamaan (5), (6), (7), (8), TP adalah jumlah prediksi benar pada kelas positif, TN adalah jumlah prediksi benar pada kelas negatif, FP adalah jumlah prediksi salah sebagai kelas positif, dan FN adalah jumlah prediksi salah sebagai kelas negatif.

HASIL DAN PEMBAHASAN

4.1 Dataset dan Persiapan Data

Proses seleksi data menghasilkan 1.344 berkas audio yang digunakan dalam penelitian ini, setelah 96 berkas berlabel *neutral* dikeluarkan dari 1.440 berkas RAVDESS original. Ketujuh kategori emosi yang digunakan terdistribusi secara seimbang, dengan masing-masing kelas memuat 192 berkas. Keseimbangan distribusi tersebut menjadi prasyarat metodologis bagi objektivitas metrik evaluasi *accuracy* dan *macro-F1* pada tahap pengujian model, karena meniadakan dominasi kelas mayoritas yang dapat mendistorsi interpretasi performa klasifikasi. Penghapusan kelas *neutral* dilakukan

karena pada dataset RAVDESS emosi tersebut hanya direkam pada satu tingkat intensitas (*normal intensity*), sedangkan seluruh emosi lainnya direkam pada dua tingkat intensitas (*normal* dan *strong*), sehingga jumlah sampelnya hanya 96 berkas atau setengah dari jumlah sampel pada kelas emosi lainnya (192 berkas), yang berpotensi menyebabkan ketidakseimbangan distribusi data dan memengaruhi proses pembelajaran model klasifikasi. Distribusi jumlah sampel pada setiap kategori emosi setelah proses seleksi data dapat dilihat pada Tabel 2.

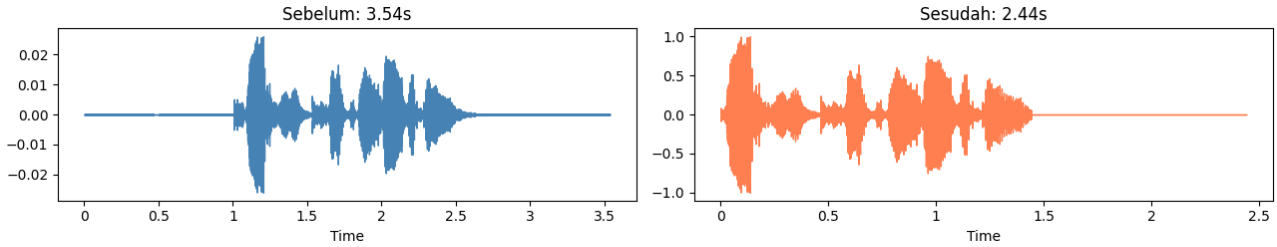
Tabel 1 Distribusi jumlah sampel per kelas emosi

Parameter	Jumlah File
calm	96
happy	192
sad	192
angry	192
fearful	192
disgust	192
surprised	192
Total	**1.344**

Karakteristik sinyal audio juga dianalisis untuk memperoleh gambaran umum mengenai dataset yang digunakan. Seluruh file audio memiliki format .WAV 16-bit dengan frekuensi pengambilan sampel (sampling rate) sebesar 48.000 Hz sesuai spesifikasi asli dataset RAVDESS. Berdasarkan hasil analisis terhadap 1.344 file audio, diperoleh durasi rata-rata sinyal sebesar 3,715 detik dengan panjang sinyal rata-rata sebesar 81.911 sampel. Informasi ini menunjukkan bahwa rekaman suara dalam dataset memiliki durasi yang relatif seragam sehingga sesuai digunakan dalam proses ekstraksi fitur dan pelatihan model klasifikasi emosi.

4.2. Hasil Preprocessing Audio

Tahap preprocessing diterapkan terhadap seluruh 1.344 file audio guna menghasilkan representasi sinyal yang konsisten sebelum proses ekstraksi fitur dilakukan. Analisis durasi terhadap file audio asli menunjukkan rentang antara 2,936 hingga 5,272 detik dengan rata-rata 3,715 detik. Setelah proses trimming silence diterapkan menggunakan ambang batas 20 dB, durasi aktif sinyal mengalami penurunan signifikan menjadi rata-rata 1,719 detik dengan rentang 0,836 hingga 3,390 detik. Penurunan rata-rata durasi sebesar 53,8% tersebut mengindikasikan bahwa proporsi segmen senyap pada dataset RAVDESS cukup besar, sehingga proses trimming berperan penting dalam menghilangkan informasi yang tidak relevan secara emosional sebelum tahap ekstraksi fitur. Perbandingan sinyal audio sebelum dan sesudah proses *trimming silence* ditampilkan pada Gambar 2.



Gambar 2. sinyal audio sebelum dan sesudah trimming silence

Penentuan durasi target tidak menggunakan nilai rata-rata, melainkan persentil ke-95 dari distribusi durasi pascatrimming, yaitu 2,44 detik atau setara dengan 53.802 sampel pada frekuensi sampling 22.050 Hz. Pendekatan ini dipilih karena nilai rata-rata pascatrimming tidak mampu mengakomodasi sebagian besar variasi panjang sinyal secara memadai, sehingga penggunaan persentil ke-95 menjamin bahwa hampir seluruh sampel dapat direpresentasikan tanpa kehilangan informasi yang berarti akibat pemotongan. Sinyal yang lebih pendek dari durasi target diberi zero-padding pada bagian akhir, sedangkan sinyal yang melebihi durasi tersebut dipotong dari ujungnya, sehingga seluruh sampel memiliki dimensi input yang seragam sebagai prasyarat tahap ekstraksi fitur prosodi dan spektral.

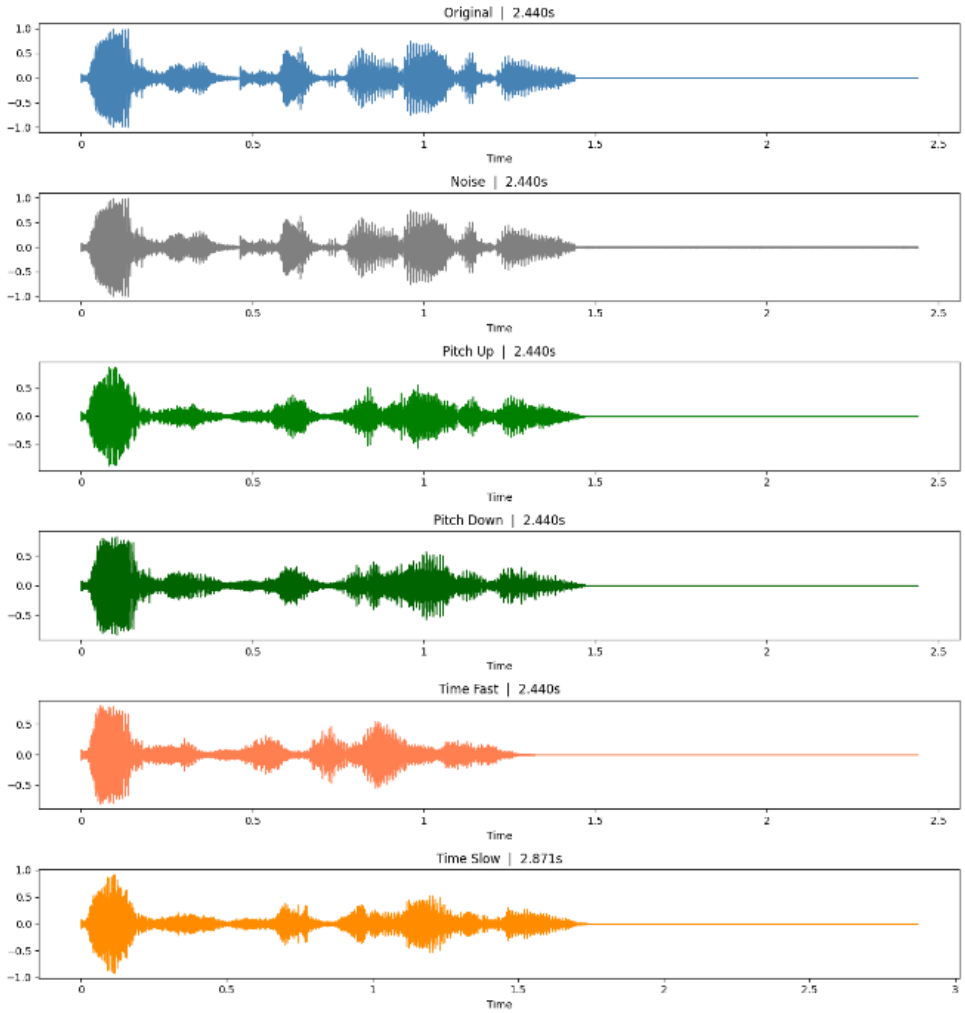
Selain proses trimming dan penyeragaman durasi, reduksi noise diterapkan menggunakan Butterworth Low-Pass Filter orde tiga dengan frekuensi cutoff 2.000 Hz melalui metode zero-phase filtering untuk menghindari distorsi fase pada sinyal hasil. Keseluruhan tahap preprocessing ini memastikan setiap sampel audio memiliki kualitas dan dimensi

yang seragam, sehingga variabilitas yang muncul pada tahap pelatihan model dapat diatribusikan pada karakteristik emosional sinyal, bukan pada inkonsistensi teknis hasil akuisisi data.

4.3 Hasil Data Augmentasi

Untuk mengatasi keterbatasan jumlah data pelatihan, diterapkan lima teknik augmentasi audio secara kombinatif, yaitu noise injection dengan faktor 0,005, pitch shift naik sebesar +2 semitone, pitch shift turun sebesar −2 semitone, time stretch dipercepat sebesar 1,15×, dan time stretch diperlambat sebesar 0,85×. Setiap file audio original menghasilkan lima versi augmentasi tambahan, sehingga setiap sampel memiliki enam representasi yang berbeda secara akustik namun tetap mempertahankan label emosi yang sama.

Dengan total 1.075 file pada data latih hasil pembagian 80% dari keseluruhan dataset, proses augmentasi menghasilkan 6.450 sampel pelatihan. Oleh karena pembagian data dilakukan secara stratified berdasarkan kelas emosi sebelum augmentasi, distribusi antarkelas pada data latih tetap proporsional setelah augmentasi diterapkan, sehingga tidak terjadi bias terhadap kelas tertentu selama proses pelatihan model. Seluruh versi augmentasi diseragamkan pada durasi 2,44 detik yang telah ditetapkan pada tahap preprocessing, melalui mekanisme zero-padding atau pemotongan sebagaimana diterapkan pada sampel original. Visualisasi sinyal audio hasil augmentasi untuk setiap teknik yang diterapkan disajikan pada Gambar 3.

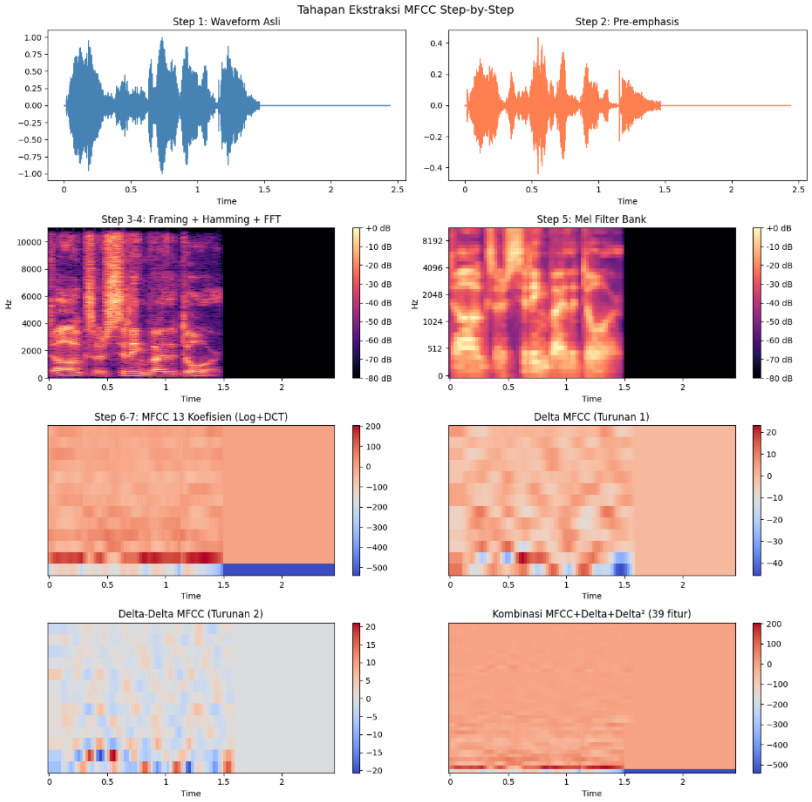


Gambar 3. Visualisasi Hasil Augmentasi

Penerapan augmentasi pada level sinyal audio mentah, sebelum proses ekstraksi fitur dilakukan, memungkinkan variasi akustik yang dihasilkan tetap konsisten dengan karakteristik alami suara manusia, berbeda dengan augmentasi pada level fitur yang berpotensi menghasilkan representasi tidak realistis. Peningkatan volume data pelatihan sebesar enam kali lipat ini diharapkan dapat meningkatkan kemampuan generalisasi model, khususnya pada model CNN yang secara inheren memerlukan jumlah data lebih besar dibandingkan model klasik seperti SVM untuk mencapai performa optimal.

4.4 Hasil Ekstraksi Fitur

Hasil ekstraksi fitur menghasilkan representasi yang sesuai dengan rancangan pada subbab metodologi. Fitur prosodi yang diekstraksi menghasilkan 11 dimensi yang identik untuk kedua model, sedangkan fitur spektral menghasilkan representasi yang berbeda sesuai kebutuhan masing-masing algoritma: vektor berdimensi 37 (26 fitur MFCC tersarikan + 11 fitur prosodi) untuk model SVM, dan matriks (13, 107, 1) yang mempertahankan struktur temporal untuk model CNN, dengan fitur prosodi diproses melalui cabang *dense* terpisah sebelum digabungkan pada lapisan klasifikasi akhir. Tahapan pembentukan fitur MFCC beserta representasi fitur yang dihasilkan divisualisasikan pada Gambar 4.

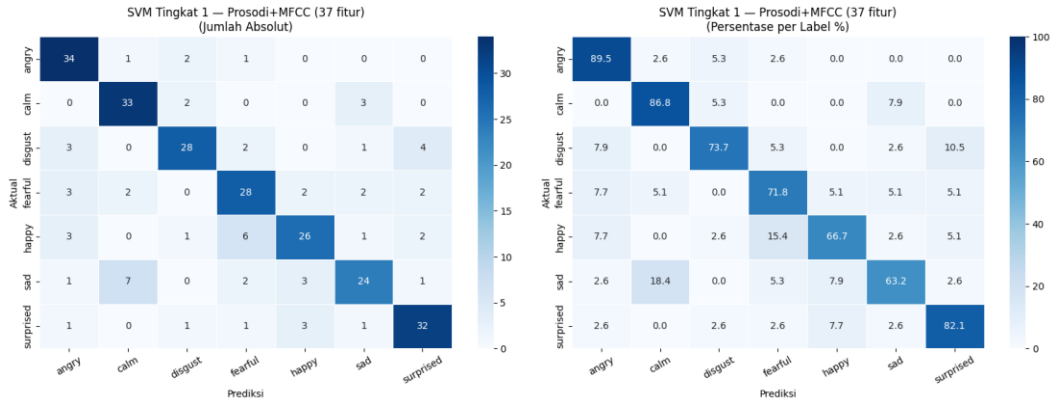


Gambar 4. Visualisasi Tahapan Ekstraksi Fitur MFCC Dari Sinyal Audio Hasil Preprocessing Hingga Representasi Fitur Akhir.

Konsistensi dimensi fitur ini memastikan bahwa kedua model menerima informasi akustik yang setara dari sumber data yang sama, sehingga perbedaan performa klasifikasi pada tahap pengujian dapat diatribusikan pada kemampuan masing-masing algoritma dalam memanfaatkan struktur fitur yang tersedia, bukan pada perbedaan kandungan informasi akustik antarmodel.

4.5 Performa Model SVM

Evaluasi model SVM dilakukan menggunakan kernel RBF dengan optimasi *hyperparameter* melalui GridSearchCV lima-fold cross-validation pada kombinasi fitur prosodi dan spektral MFCC. Hasil pengujian pada parameter optimal C=10 menunjukkan akurasi sebesar 76,21% dengan precision 0,7629, recall 0,7621, dan weighted F1-score sebesar 0,7596. Visualisasi hasil klasifikasi model SVM dalam bentuk confusion matrix disajikan pada Gambar 5.



Gambar 5. Confusion matrix hasil klasifikasi model SVM

Berdasarkan confusion matrix, model menunjukkan kemampuan klasifikasi yang cukup baik pada seluruh kelas emosi. Hal ini terlihat dari tingginya tingkat prediksi benar pada beberapa kelas, seperti angry sebesar 89,5%, calm sebesar 86,8%, dan surprised sebesar 82,1%. Namun, masih terdapat beberapa kesalahan prediksi, terutama pada emosi sad yang hanya mencapai 63,2% dan happy sebesar 66,7%, yang mengindikasikan adanya kemiripan karakteristik dengan kelas emosi lainnya sehingga lebih sulit dibedakan oleh model.

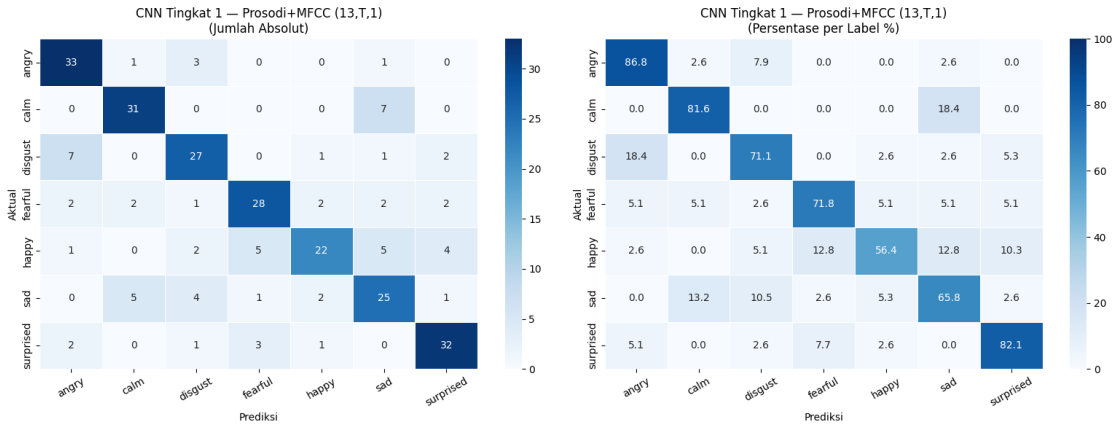
Tabel 2. Performa Model SVM

Metrik	Nilai
Accuracy	76,21%
Precision	0,7629
Recall	0,7621
F1-Score	0,7596

Berdasarkan hasil evaluasi, model memperoleh akurasi sebesar 76,21%, yang menunjukkan bahwa sebagian besar sampel berhasil diklasifikasikan dengan benar. Nilai precision sebesar 0,7629, recall sebesar 0,7621, dan F1-score sebesar 0,7596 menunjukkan keseimbangan yang baik antara kemampuan model dalam mengidentifikasi kelas emosi dan meminimalkan kesalahan klasifikasi. Selain itu, model memerlukan waktu pelatihan (training time) selama 458,5 detik dan mampu melakukan prediksi dengan cepat, yaitu hanya 0,083 detik pada tahap inferensi. Secara keseluruhan, hasil ini menunjukkan bahwa model memiliki performa klasifikasi yang baik dengan waktu prediksi yang efisien.

4.6 Performa Model CNN

Evaluasi model CNN dilakukan menggunakan arsitektur dual-input yang menerima gambar spektral MFCC dan vektor prosodi secara terpisah, dilengkapi mekanisme early stopping berdasarkan validation accuracy dan ReduceLROnPlateau untuk stabilisasi konvergensi. Hasil pengujian menunjukkan akurasi sebesar 73,61% yang tercapai pada epoch ke-36, dengan precision 0,7392, recall 0,7361, dan weighted F1-score sebesar 0,7339. Visualisasi hasil klasifikasi model CNN dalam bentuk confusion matrix disajikan pada Gambar 6.



Gambar 6. Confusion matrix hasil klasifikasi model SVM

Analisis classification report menunjukkan kelas calm memperoleh F1-score tertinggi sebesar 0,81, diikuti angry dan surprised yang masing-masing bernilai 0,80. Sebaliknya, kelas sad memperoleh F1-score terendah sebesar 0,63, diikuti happy sebesar 0,66. Rendahnya performa pada kelas sad konsisten dengan karakteristik akustiknya yang memiliki energi rendah dan variasi pitch yang minim, sehingga sulit dibedakan secara tegas dari kelas calm yang memiliki karakteristik energi serupa. Hasil komparasi selengkapnya disajikan pada Tabel 4.

Tabel 3 Performa Model CNN.

Metrik	Nilai
Accuracy	73,61%
Precision	0,7392
Recall	0,7361
F1-Score	0,7339

Waktu pelatihan model tercatat sebesar 75,7 detik, jauh lebih singkat dibandingkan waktu yang dibutuhkan SVM untuk proses pencarian hyperparameter. Efisiensi ini disebabkan oleh konvergensi yang relatif cepat pada epoch ke-36 serta pemanfaatan akselerasi GPU pada proses backpropagation. Sebaliknya, waktu inferensi yang mencapai 1,428 detik menunjukkan bahwa proses prediksi CNN memerlukan komputasi lebih intensif dibandingkan SVM, sebagai konsekuensi dari kompleksitas arsitektur dual-input dengan empat lapisan konvolusi.

4.7 Komparasi Performa CNN dan SVM

Komparasi performa dilakukan antara model SVM dan CNN yang menggunakan kombinasi fitur prosodi dan spektral MFCC secara setara, sehingga perbandingan yang diperoleh dapat diatribusikan secara langsung pada perbedaan karakteristik algoritma, bukan pada perbedaan kompleksitas representasi fitur. Hasil komparasi selengkapnya disajikan pada Tabel 5.

Tabel 4. Perbandingan Performa SVM dan CNN

Metrik	SVM	CNN	Selisih
Accuracy	76,21%	73,61%	+2,60% (SVM)
Precision	0,7629	0,7392	+0,0237 (SVM)
Recall	0,7621	0,7361	+0,0260 (SVM)
F1-Score (weighted)	0,7596	0,7339	+0,0257 (SVM)
Waktu Pelatihan	458,5 detik	75,7 detik	CNN 6× lebih cepat
Waktu Inferensi	0,083 detik	1,428 detik	SVM 17× lebih cepat

Berdasarkan keempat metrik evaluasi klasifikasi, SVM secara konsisten mengungguli CNN, dengan selisih akurasi sebesar 2,60 poin persentase dan selisih weighted F1-score sebesar 0,0257. Keunggulan SVM ini dapat dijelaskan melalui beberapa faktor yang saling berkaitan. Pertama, volume data pelatihan sebesar 6.450 sampel hasil augmentasi tergolong relatif terbatas untuk melatih model CNN secara optimal, mengingat arsitektur dual-input dengan empat lapisan

konvolusi memiliki jumlah parameter yang jauh lebih besar dibandingkan SVM, sehingga lebih rentan terhadap kondisi data tidak memadai meskipun mekanisme regularisasi melalui dropout dan early stopping telah diterapkan. Kedua, SVM dengan kernel RBF memiliki keunggulan yang telah dikonfirmasi pada berbagai penelitian SER sebelumnya dalam menangani data berdimensi rendah dengan jumlah sampel terbatas, sebagaimana representasi fitur yang digunakan pada penelitian ini yang hanya berjumlah 37 dimensi.

Meskipun demikian, CNN menunjukkan keunggulan signifikan dari sisi efisiensi komputasi pada tahap pelatihan, dengan waktu yang hanya 75,7 detik dibandingkan 458,5 detik pada SVM, suatu selisih yang mencapai sekitar enam kali lipat. Efisiensi ini terutama disebabkan oleh proses optimasi hyperparameter pada SVM yang melibatkan pencarian grid melalui dua belas kombinasi parameter dengan five-fold cross-validation, sementara CNN langsung dilatih dengan konfigurasi arsitektur yang telah ditentukan. Sebaliknya, pada tahap inferensi, SVM unggul signifikan dengan waktu prediksi 0,083 detik dibandingkan 1,428 detik pada CNN, menjadikan SVM lebih sesuai untuk skenario penerapan yang membutuhkan respons waktu nyata dengan keterbatasan sumber daya komputasi.

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, metode Support Vector Machine (SVM) dan Convolutional Neural Network (CNN) mampu melakukan klasifikasi emosi suara pada dataset RAVDESS menggunakan kombinasi fitur prosodi dan MFCC dengan tingkat akurasi di atas 70%. Model SVM menunjukkan performa terbaik dengan accuracy sebesar 76,21%, precision sebesar 0,7629, recall sebesar 0,7621, dan F1-score sebesar 0,7596. Sementara itu, model CNN memperoleh accuracy sebesar 73,61%, precision sebesar 0,7392, recall sebesar 0,7361, dan F1-score sebesar 0,7339. Hasil tersebut menunjukkan bahwa SVM lebih unggul dibandingkan CNN pada konfigurasi fitur dan dataset yang digunakan dalam penelitian ini. Selain itu, kedua model menunjukkan performa terbaik pada emosi angry, calm, dan surprised, sedangkan emosi sad dan happy masih menjadi kelas yang paling sulit diklasifikasikan karena memiliki karakteristik akustik yang relatif saling tumpang tindih. Dengan demikian, kombinasi fitur prosodi dan MFCC yang direpresentasikan dalam bentuk vektor statistik terbukti efektif untuk mendukung klasifikasi emosi suara menggunakan SVM.

Penelitian selanjutnya dapat mengeksplorasi penggunaan fitur akustik tambahan dan teknik augmentasi data untuk meningkatkan performa klasifikasi emosi suara. Selain itu, penerapan arsitektur deep learning yang lebih kompleks serta pengujian pada dataset yang lebih beragam perlu dilakukan untuk meningkatkan kemampuan generalisasi model.

DAFTAR PUSTAKA

[1] L. Ode, A. Hazmar, and B. H. Prasetio, “Sistem Pengenalan Tingkatan Emosi Ketakutan Melalui Ucapan menggunakan Ekstraksi Gammatone-Frequency Cepstral Coefficients dan Klasifikasi Random Forest Classifier berbasis Raspberry Pi 4,” vol. 7, no. 2, pp. 1003–1011, 2023.

[2] A. P. Pinheiro, “Behind a Voice There is a Speaker : Why Vocal Emotion Research Needs to Become ‘ Personal ,’” *Affect. Sci.*, vol. 6, no. 3, pp. 562–574, 2025, doi: 10.1007/s42761-025-00317-w.

[3] W. Swastika, A. A. Oepojo, and P. L. T. Irawan, “Perbandingan Akurasi Deteksi Emosi Pada Suara Menggunakan Multilayer Perceptron , Random Forest , Decision Tree dan K-NN,” vol. 05, no. 01, pp. 1–6, 2023, doi: 10.52985/insyst.v5i1.264.

[4] A. P. Wibowo, “Pengembangan Algoritma Deteksi Emosi Melalui Analisis Suara untuk Aplikasi Konseling Digital berbagai metode ekstraksi fitur seperti Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), dan Short Time Energy (STE) telah digunakan untuk,” vol. 3, 2025.

[5] L. Schewski, “Measuring negative emotions and stress through acoustic correlates in speech : A systematic review,” pp. 1–28, 2025, doi: 10.1371/journal.pone.0328833.

[6] A. Altheneyan and A. Alhadlaq, “Feature selection for emotion recognition in speech : a comparative study of fi lter and wrapper methods,” 2025, doi: 10.7717/peerj-cs.3180.

[7] E. L. Febrianti and T. Christi, “Peneraan Forward Chaining Untuk Mendianogsa Penyakit Malaria Dan Pencegahanya Berbasis Web,” *Jurteksi*, vol. 4, no. 1, pp. 93–100, 2017, doi: https://doi.org/10.33330/jurteksi.v4i1.32.

[8] E. Jordan, V. Lucarini, and M. Alrahabi, “Speech Emotion Recognition in Mental Health : Systematic Review of Voice-Based Applications Corresponding Author :,” vol. 12, pp. 1–17, 2025, doi: 10.2196/74260.

[9] S. Y. Chowdhury, B. Banik, T. Hoque, and S. Banerjee, “A Novel Hybrid Deep Learning Technique for Speech Emotion Detection using Feature Engineering,” 2026.

[10] C. Barhoumi and Y. Benayed, “Real-time speech emotion recognition using deep learning and data augmentation,” 2025.

[11] K. V. K. ; N. S. ; A. M. Posonia, “Speech Emotion Recognition using Machine Learning,” 2022, doi: 10.1109/ICCMC53470.2022.9753976.

[12] J. C. &Vasan A. Tulika Jha , Ramisetty Kavya, “Machine learning techniques for speech emotion recognition using paralinguistic acoustic features,” 2022, doi: https://doi.org/10.1007/s10772-022-09985-6.

[13] P. Ghosh, “Comparative Study of Classification Models for Emotion Detection from Speech,” vol. 8, no. 3, pp. 837–840, 2023.

[14] S. H. S. Gheed T. Waleed, “Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN”, doi: https://doi.org/10.3390/info16070518.

[15] D. R. I. M. S. Kristiawan Nugroho, Imam Husni Al Amin, Nina Anggraeni Noviasari, "Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition", doi: https://doi.org/10.3390/computers14090361.
[16] C. Lombardo, G. Esposito, S. Carbone, S. Serrano, and C. Mento, "Speech analysis and speech emotion recognition in mental disease : a scoping review," no. November, 2025, doi: 10.3389/fpsyg.2025.1645860.
[17] G. T. Waleed and S. H. Shaker, "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN," 2025.
[18] N. Kumar, S. Kobir, and R. Ahmed, "Enhanced Speech Emotion Recognition with Efficient Channel Attention Guided Deep CNN-BiLSTM Framework".
[19] Q. Ouyang, "Speech Emotion Detection Based on MFCC and CNN-LSTM Architecture," 2006.
[20] L. Gueguen and R. Hamti, "xView2: A Dataset for Interpretable Disaster Damage Assessment," *arXiv Prepr.*, vol. arXiv:2004, 2020, doi: https://doi.org/10.48550/arXiv.2004.01912 Focus to learn more.
[21] S. Madanian *et al.*, "Intelligent Systems with Applications Speech emotion recognition using machine learning — A systematic review," *Intell. Syst. with Appl.*, vol. 20, no. July, p. 200266, 2023, doi: 10.1016/j.iswa.2023.200266.
[22] F. S. Alamri, A. R. Mirdad, T. Saba, M. A. Raza, and U. Siddique, "Investigating the Impact of Combined Spectral and Prosodic Features on Speech Emotion," *IEEE Access*, vol. PP, p. 1, 2025, doi: 10.1109/ACCESS.2025.3650618.
[23] I. N. Afifah, T. B. Santoso, and T. Dutono, "Indonesian speech emotion recognition : feature extraction and neural network approaches," vol. 15, no. 4, pp. 3769–3778, 2025, doi: 10.11591/ijece.v15i4.pp3769-3778.
[24] D. B. Javaheri, "Speech & Song Emotion Recognition Using Multilayer Perceptron and Standard Vector Machine," pp. 1–7, 2021.
[25] R. Al Zoubi, B. Mahfood, and S. Abbas, *Proceedings of International Conference on Information Technology and Applications (ICITA 2021)*. 2020. doi: 10.1007/978-981-16-7618-5_52.
[26] D. Krasnoproshin and M. Vashkevich, "Speech emotion recognition using SVM classifier with suprasegmental MFCC features," pp. 118–121.
[27] L. Wuttisittikulkij, S. Shah, S. M. Ali, and M. Alibakhshikenari, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer," pp. 1–20, 2023.
[28] T. M. L. Flower and T. Jaya, "A Novel Concatenated 1D-CNN Model for Speech Emotion Recognition," *Biomed. Signal Process. Control*, vol. 93, no. 106201, 2024, doi: 10.1016/j.bspc.2024.106201.
[29] N. Barsainyan and Dileep Kumar Singh, *Speech Emotion Recognition Using Deep Learning Algorithm on RAVDESS Dataset*. 2024. doi: 10.1007/978-981-99-9554-7_33.
[30] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," pp. 1–14, 2025.
[31] "hyper.pdf."
[32] L. Azhari and E. Widarti, "Evaluasi Kinerja Kernel Linear , RBF , dan Polynomial pada Model Support Vector Machine untuk Prediksi Risiko Hipertensi," vol. 17, no. 2, pp. 192–201, 2025, doi: 10.22441/fifo.2025.v17i2.008.
[33] E. Lieskovsk, M. Jakubec, R. Jarina, and M. Chmul, "A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism," 2021.