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

Klasifikasi Emosi Suara Multi Bahasa Berbasis Deep Learning Menggunakan Mel Spectrogram dan MFCC pada Indian Speech Emotion

Gilang Maulana1*, Wahyu Herdiansyah2,Thalhah al fayadh3, Muhammad Rizky4, Supriadin5
1,2Sistem Teknologi Informasi, Universitas Muhammdiyah Mataram
email: gilanggmaulana2@gmail.com[1], wahyuherdiansyah32@gmail.com[2] ,tolhahalfayadh@gmail.com[3], dompusupriadin43@gmail.com[4], risganteng3@gmail.com[5]

Abstrak: Pengenalan emosi berbasis suara (Speech Emotion Recognition/SER) pada lingkungan multibahasa masih menjadi tantangan karena keterbatasan dataset dan variasi karakteristik akustik antar bahasa. Penelitian ini bertujuan membangun sistem klasifikasi emosi suara menggunakan arsitektur deep learning CNN-LSTM hibrida dengan fitur gabungan Mel Spectrogram dan MFCC pada Indian Speech Emotion Dataset yang mencakup tiga bahasa (Inggris, Hindi, dan Marathi) dengan 1.000 sampel audio pada lima kelas emosi (angry, happy, neutral, sad, surprise). Metode penelitian meliputi pra-pemrosesan audio, augmentasi data (time stretching, pitch shifting, penambahan noise, dan volume scaling), ekstraksi fitur Mel Spectrogram, MFCC, Delta MFCC, dan Delta-Delta MFCC, serta pelatihan model CNN-BiLSTM dengan rasio data latih dan uji 80:20. Hasil pengujian pada data uji menunjukkan akurasi sebesar 70,00% dengan F1-score rata-rata sebesar 0,70. Kelas neutral memperoleh performa terbaik (F1-score 0,78), sedangkan kelas happy paling sering tertukar dengan kelas lain (F1-score 0,58). Hasil ini menunjukkan kombinasi fitur Mel Spectrogram dan MFCC mampu mengenali emosi suara multibahasa, meskipun terdapat indikasi overfitting yang membuka peluang perbaikan melalui regularisasi dan penambahan data pada penelitian selanjutnya.

Kata Kunci: CNN-LSTM; Indian Speech Emotion Dataset; Mel Spectrogram; MFCC; multibahasa; Speech Emotion Recognition

Abstract: *Speech Emotion Recognition (SER) in multilingual settings remains challenging due to limited datasets and varying acoustic characteristics across languages. This study aims to build a speech emotion classification system using a hybrid CNN-LSTM deep learning architecture with combined Mel Spectrogram and MFCC features on the Indian Speech Emotion Dataset, which covers three languages (English, Hindi, and Marathi) with 1,000 audio samples across five emotion classes (angry, happy, neutral, sad, surprise). The method includes audio preprocessing, data augmentation (time stretching, pitch shifting, noise addition, and volume scaling), feature extraction using Mel Spectrogram, MFCC, Delta MFCC, and Delta-Delta MFCC, and training a CNN-BiLSTM model with an 80:20 train-test split. Testing results show an accuracy of 70.00% with an average F1-score of 0.70. The neutral class achieved the best performance (F1-score 0.78), while the happy class was most frequently confused with other classes (F1-score 0.58). These results indicate that the combination of Mel Spectrogram and MFCC features is able to recognize multilingual speech emotion, although signs of overfitting suggest room for improvement through stronger regularization and additional data in future work.*

Keywords: *CNN-LSTM; Indian Speech Emotion Dataset; Mel Spectrogram; MFCC; multilingual; Speech Emotion Recognition*

PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI) dan pengolahan sinyal suara telah membuka peluang besar dalam pengembangan sistem pengenalan emosi manusia berbasis suara [1][2]. Emosi merupakan aspek penting dalam komunikasi manusia karena dapat mencerminkan kondisi psikologis, niat, maupun respons seseorang terhadap suatu situasi [3]. Dalam komunikasi verbal, emosi tidak hanya disampaikan melalui kata-kata, tetapi juga melalui intonasi, tekanan suara, kecepatan berbicara, serta karakteristik akustik lainnya [4]. Oleh karena itu, analisis emosi suara

menjadi salah satu bidang penelitian yang terus berkembang, khususnya pada penerapan Human Computer Interaction (HCI), layanan pelanggan otomatis, sistem pembelajaran digital, hingga teknologi kesehatan mental [3][5].

Pengenalan emosi berbasis suara atau Speech Emotion Recognition (SER) merupakan proses identifikasi kondisi emosional seseorang dari sinyal suara dengan memanfaatkan teknik pengolahan sinyal, machine learning, maupun deep learning [3][6]. Seiring berkembangnya teknologi deep learning, berbagai metode klasifikasi emosi menunjukkan performa yang lebih baik dibandingkan pendekatan konvensional [7][8]. Model deep learning mampu mempelajari pola kompleks dari data suara secara otomatis tanpa memerlukan proses rekayasa fitur manual yang berlebihan [6][7]. Namun demikian, performa sistem SER masih sangat dipengaruhi oleh kualitas dataset, metode ekstraksi fitur, serta arsitektur model yang digunakan [9][16].

Salah satu tantangan utama dalam penelitian Speech Emotion Recognition adalah keberagaman bahasa dan karakteristik suara manusia [3][4]. Banyak penelitian sebelumnya hanya menggunakan dataset berbahasa tunggal sehingga model yang dihasilkan kurang mampu melakukan generalisasi terhadap data multibahasa [11][19]. Untuk menjawab permasalahan tersebut, penelitian ini memanfaatkan Indian Speech Emotion Dataset yang mencakup tiga bahasa (Inggris, Hindi, dan Marathi) dengan variasi ekspresi suara yang beragam. Penggunaan dataset multibahasa diharapkan dapat meningkatkan kemampuan model dalam mengenali emosi dari berbagai karakteristik pelafalan dan intonasi.

Dataset ini terdiri atas 1.000 sampel audio format WAV yang terdistribusi merata pada lima kelas emosi, yaitu marah (angry), bahagia (happy), netral (neutral), sedih (sad), dan terkejut (surprise), masing-masing berjumlah 200 sampel. Tiga bahasa yang tercakup, yaitu bahasa Inggris (250 sampel), Hindi (250 sampel), dan Marathi (500 sampel), menjadikan dataset ini representatif untuk skenario pengenalan emosi lintas bahasa India. Berdasarkan analisis pada keseluruhan sampel, durasi rata-rata setiap file audio dalam dataset ini adalah sekitar 3–4 detik per sampel, sehingga estimasi total durasi keseluruhan dataset berkisar antara 50–67 menit audio. Rentang durasi ini cukup representatif untuk penelitian Speech Emotion Recognition skala menengah, meskipun lebih pendek dibandingkan dataset SER berukuran besar seperti RAVDESS atau CREMA-D [10].

Dalam proses ekstraksi fitur suara, Mel Spectrogram dan Mel Frequency Cepstral Coefficients (MFCC) merupakan dua metode yang paling banyak digunakan dalam penelitian pengolahan audio [9][13]. MFCC mampu merepresentasikan karakteristik frekuensi suara yang menyerupai sistem pendengaran manusia melalui pendekatan skala mel [11][13]. Sementara itu, Mel Spectrogram dapat menggambarkan distribusi energi frekuensi terhadap waktu secara visual sehingga membantu model deep learning mengenali pola emosi pada sinyal suara [12][19]. Kombinasi kedua fitur tersebut dinilai mampu membantu model menangkap informasi temporal dan frekuensi secara lebih optimal sehingga berpotensi meningkatkan akurasi klasifikasi emosi [8][16].

Beberapa penelitian terdahulu menunjukkan bahwa penggunaan deep learning seperti Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), maupun kombinasi keduanya mampu menghasilkan performa yang baik pada klasifikasi emosi suara [6][7][20]. Akan tetapi, sebagian penelitian masih memiliki keterbatasan pada jumlah data, variasi bahasa, dan jenis fitur yang digunakan [3][4]. Selain itu, masih sedikit penelitian yang mengombinasikan Mel Spectrogram dan MFCC pada dataset multibahasa India untuk klasifikasi emosi suara secara lebih mendalam [12][13].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk membangun sistem klasifikasi emosi suara berbasis deep learning menggunakan fitur Mel Spectrogram dan MFCC pada Indian Speech Emotion Dataset yang mencakup tiga bahasa. Penelitian ini diharapkan dapat menghasilkan model yang mampu mengenali emosi suara dengan tingkat akurasi yang baik serta memberikan kontribusi dalam pengembangan teknologi pengenalan emosi berbasis audio pada lingkungan multibahasa India.

TINJAUAN PUSTAKA

Penelitian mengenai Speech Emotion Recognition (SER) telah berkembang pesat dalam satu dekade terakhir seiring dengan kemajuan teknologi deep learning. Al-Dujaili dan Ebrahimi-Moghadam [3] melakukan survei komprehensif terhadap metode SER dan menyimpulkan bahwa pendekatan deep learning, khususnya CNN dan LSTM, secara konsisten menghasilkan akurasi lebih tinggi dibandingkan pendekatan berbasis fitur konvensional seperti SVM atau GMM.

Li dan Cao [20] membandingkan model CNN-LSTM standar dengan CNN-LSTM yang diperkuat mekanisme attention, dan menemukan bahwa attention layer secara signifikan meningkatkan kemampuan model dalam menangkap fitur temporal yang relevan. Tang et al. [5] mengusulkan arsitektur CNN-Transformer dengan mekanisme multidimensional attention yang berhasil mencapai akurasi tinggi pada dataset IEMOCAP. Kedua penelitian tersebut memperkuat argumen bahwa kombinasi CNN untuk ekstrasi fitur spasial dan mekanisme sekuensial untuk konteks temporal merupakan pendekatan yang efektif dalam SER.

Pada aspek ekstraksi fitur, Mukhamediya et al. [9] menunjukkan bahwa parameter tuning pada log-mel spectrogram, termasuk jumlah filter dan ukuran frame, berdampak signifikan terhadap performa model deep learning dalam SER. Mohan et al. [11] membuktikan bahwa MFCC yang dikombinasikan dengan model ensemble menghasilkan klasifikasi emosi yang lebih robust dibandingkan penggunaan MFCC tunggal. Vidhi dan Seeja [12] mendemonstrasikan bahwa Mel Spectrogram dikombinasikan dengan CNN menghasilkan akurasi yang kompetitif pada dataset monolingual.

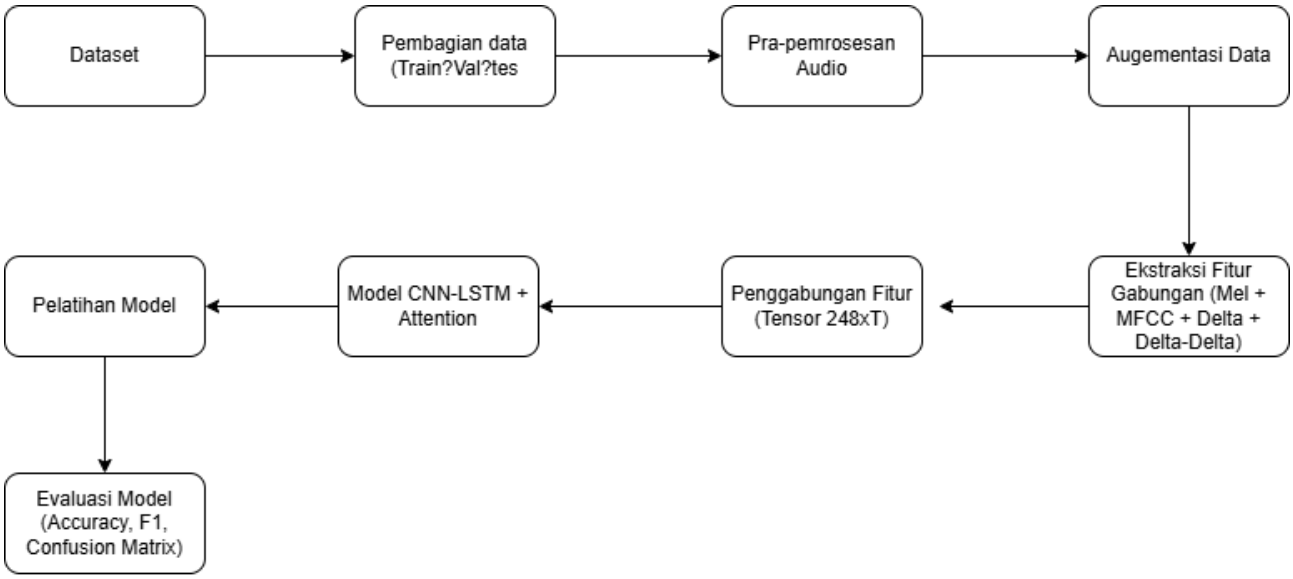
Dalam konteks augmentasi data, Mushtaq et al. [19] menunjukkan bahwa augmentasi berbasis spektral seperti time stretching dan pitch shifting secara efektif meningkatkan generalisasi model pada data audio. Alotaibi et al. [16] membuktikan dalam penelitian SER real-time bahwa augmentasi data merupakan komponen krusial untuk mengatasi keterbatasan jumlah sampel dan ketidakseimbangan kelas.

Sejauh tinjauan yang dilakukan, penelitian SER pada dataset multibahasa India yang menggabungkan Mel Spectrogram dan MFCC dengan arsitektur CNN-LSTM masih sangat terbatas. Sebagian besar penelitian menggunakan dataset berbahasa tunggal seperti RAVDESS (Inggris) atau EMO-DB (Jerman) [4][13]. Penelitian ini mengisi kesenjangan tersebut dengan menggunakan Indian Speech Emotion Dataset yang mencakup bahasa Inggris, Hindi, dan Marathi.

METODE

Penelitian ini diawali dengan pengumpulan Indian Speech Emotion Dataset yang terdiri atas tiga bahasa, yaitu English, Hindi, dan Marathi. Selanjutnya dilakukan pra-pemrosesan audio berupa resampling, normalisasi, serta penyeragaman durasi sinyal. Untuk meningkatkan variasi data, diterapkan beberapa teknik augmentasi audio, yaitu time stretching, pitch shifting, penambahan noise, dan volume scaling. Setelah itu dilakukan ekstraksi fitur menggunakan kombinasi Mel Spectrogram, MFCC, Delta MFCC, dan Delta-Delta MFCC. Fitur yang diperoleh kemudian digabungkan dan digunakan sebagai masukan model CNN-BiLSTM. Dataset dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20. Model dilatih menggunakan optimizer Adam dan dievaluasi menggunakan metrik accuracy, precision, recall, F1-score, serta confusion matrix untuk mengukur performa klasifikasi emosi suara multibahasa.
Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.

Dengan pengumpulan Indian Speech Emotion Dataset yang terdiri atas tiga bahasa, yaitu English, Hindi, dan Marathi, dengan total 1.000 sampel audio WAV dan lima kelas emosi (angry, happy, neutral, sad, surprise). Dataset dibagi menjadi data latih (680 sampel), validasi (120 sampel), dan uji (200 sampel) dengan stratifikasi berdasarkan kombinasi bahasa dan emosi untuk menjaga proporsi kelas.



A. Dataset

Dataset yang digunakan adalah Indian Speech Emotion Dataset yang terdiri atas 1.000 file audio berformat WAV. Dataset mencakup tiga bahasa, yaitu English, Hindi, dan Marathi, dengan lima kategori emosi yang terdiri atas Angry, Happy, Neutral, Sad, dan Surprise. Struktur dataset disusun berdasarkan bahasa dan emosi dengan pola folder:

Tabel 1. Distribusi Sampel Indian Speech Emotion Dataset

Bahasa	Angry	Happy	Neutral/Sad/Surprise	Total
English	50	50	50 each	250
Hindi	50	50	50 each	250
Marathi	100	100	100 each	500
Total	**200**	**200**	**200 each**	**1.000**

Distribusi data terdiri atas 250 sampel bahasa Inggris, 250 sampel bahasa Hindi, dan 500 sampel bahasa Marathi. Masing-masing kategori emosi memiliki jumlah data yang seimbang, yaitu sebanyak 200 sampel.

B. Pra-Pemrosesan dan Augmentasi

Setiap audio diresampling ke 22.050 Hz, dipotong/dipad ke durasi 3 detik, dan dinormalisasi amplitudo. Untuk meningkatkan variasi data pelatihan dan mengurangi overfitting, dilakukan augmentasi pada setiap sampel latih dengan menghasilkan **7 varian** (1 data asli + 6 data augmentasi), meliputi:

- Time stretching (faktor 0,8 dan 1,2)
- Pitch shifting (±2 semitone)
- Penambahan Gaussian noise (SNR ~15 dB)
- Volume scaling (faktor 0,7 dan 1,3)

Augmentasi **hanya** diterapkan pada data latih, sedangkan data validasi dan uji tetap dalam bentuk asli untuk menjamin evaluasi yang tidak bias.

C. Ekstraksi Fitur Gabungan

Dari setiap sinyal audio diekstrak **empat representasi akustik**:

1. Mel Spectrogram (128 Mel bands)
2. MFCC (40 koefisien)
3. Delta MFCC (turunan pertama)
4. Delta-Delta MFCC (turunan kedua)

Keempat fitur tersebut dinormalisasi per-frekuensi menggunakan StandardScaler dan kemudian **dikonkatenasi** menjadi satu tensor 3D dengan dimensi **(248, T, 1)**, di mana 248 = 128 + 40 + 40 + 40, dan T adalah jumlah frame waktu (≈130 frame untuk durasi 3 detik). Representasi ini digunakan sebagai masukan model deep learning.

D. Arsitektur Model CNN-LSTM dengan Attention

Model yang digunakan adalah **CNN-LSTM hibrida** yang terdiri dari:

- **4 blok konvolusi** (Conv2D + BatchNormalization + ReLU + MaxPooling + Dropout) dengan jumlah filter 32, 64, 128, dan 256.
- **Reshape** untuk mengubah dimensi spasial menjadi sekuens temporal.
- **LSTM 256** (return sequences) + **LayerNormalization**.
- **Mekanisme Attention** sederhana: setiap frame temporal diberi bobot melalui Dense(1, activation='tanh') dan Softmax, lalu dikalikan dengan output LSTM untuk fokus pada frame paling informatif.
- **LSTM 128** (tanpa return sequences) untuk menangkap konteks temporal akhir.
- **Dua lapisan fully-connected** (256 dan 128 neuron) dengan Dropout (0,4 dan 0,3) dan L2 regularisasi (1e-4).
- **Output layer** dengan 5 neuron dan aktivasi softmax.

Total parameter model sekitar **4,85 juta**, dengan regularisasi yang kuat untuk mengatasi overfitting.

F. Pelatihan Model

Model dikompilasi dengan:

- Optimizer **Adam** (learning rate awal 0,001)
- Loss **categorical crossentropy**
- Metrik **accuracy**

Pelatihan menggunakan:

- Batch size 32
- Maksimum 100 epoch
- Callback:
    - **EarlyStopping** (monitor val_accuracy, patience 15, restore best weights)
    - **ReduceLROnPlateau** (monitor val_loss, factor 0,5, patience 7)

Bobot kelas dihitung secara otomatis (class_weight='balanced') untuk mengatasi ketidakseimbangan data.

G. Evaluasi

Model dievaluasi pada **data uji** (200 sampel asli, tanpa augmentasi) menggunakan metrik:

- Accuracy
- Precision, Recall, F1-score per kelas
- Confusion Matrix

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

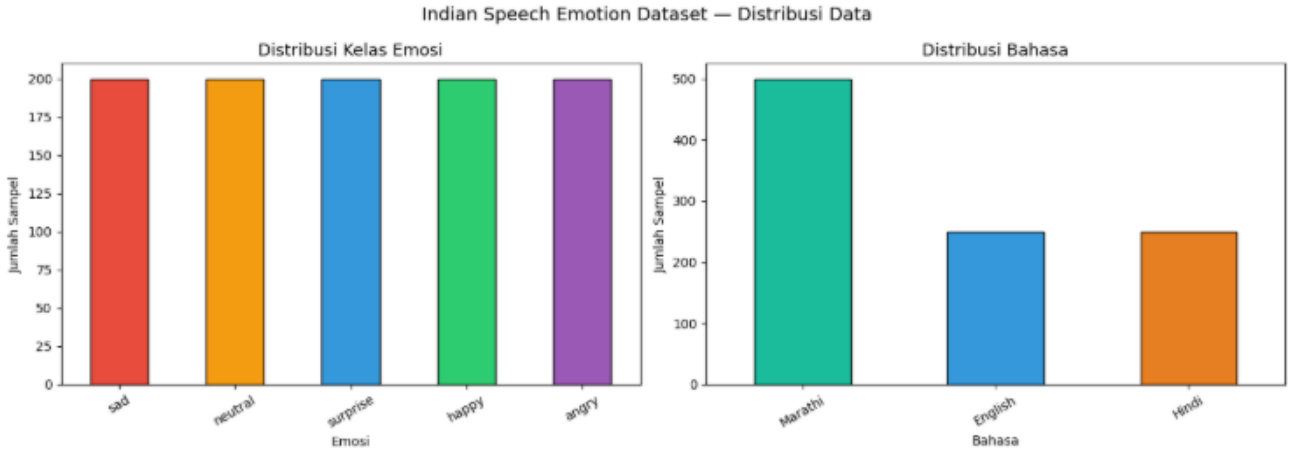
HASIL DAN PEMBAHASAN

Hasil pengujian pada data uji menunjukkan akurasi sebesar **70,00%** dan F1-score rata-rata **0,70**. Kelas *neutral* mencapai performa terbaik (F1=0,78), sedangkan *happy* paling sulit dikenali (F1=0,58) dan sering tertukar dengan *angry* dan *surprise*.

Kurva pelatihan menunjukkan akurasi data latih mencapai **96,40%** sedangkan akurasi validasi tertinggi hanya **68,33%**, menandakan adanya **overfitting** dengan gap 28,06 poin persentase. Hal ini mengindikasikan bahwa meskipun augmentasi dan regularisasi telah diterapkan, kompleksitas model masih terlalu besar untuk ukuran dataset yang relatif kecil (1.000 sampel asli). Penelitian selanjutnya disarankan menambah data asli, menggunakan teknik regularisasi lebih agresif, atau mengadopsi arsitektur berbasis transformer (misal Wav2Vec 2.0) untuk meningkatkan generalisasi.

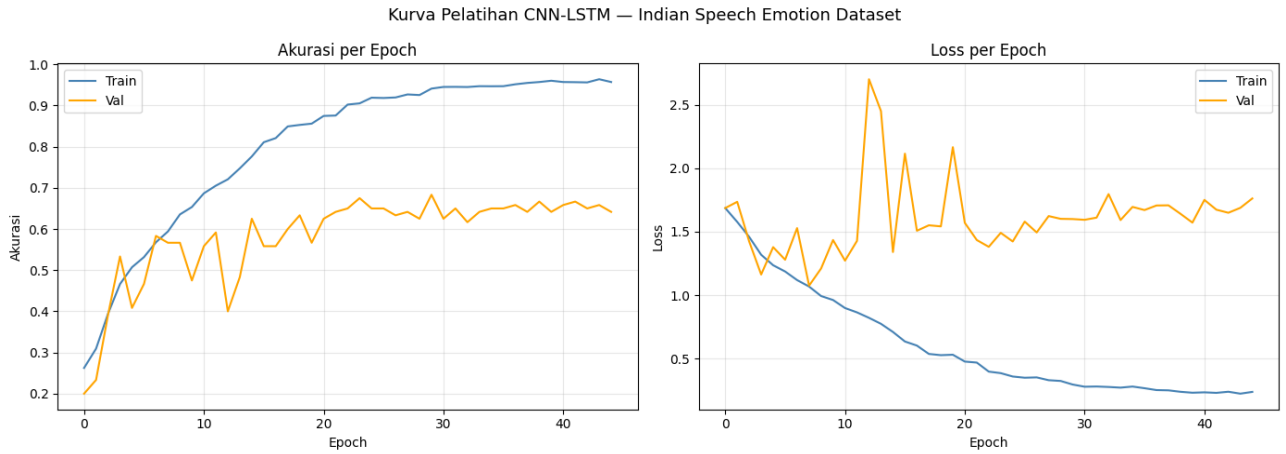
A. Distribusi Dataset

Indian Speech Emotion Dataset memiliki distribusi yang seimbang pada lima kelas emosi, masing-masing terdiri atas 200 sampel (20% dari total). Keseimbangan distribusi kelas ini menguntungkan proses pelatihan karena mengurangi risiko bias model terhadap kelas mayoritas. Distribusi bahasa menunjukkan proporsi yang tidak merata, di mana bahasa Marathi mendominasi dengan 500 sampel (50%), diikuti bahasa Inggris dan Hindi masing-masing 250 sampel (25%). Ketidakseimbangan distribusi bahasa ini menjadi tantangan tersendiri bagi kemampuan generalisasi model.



- Angry 200
- Happy 200
- Neutral 200
- Sad 200
- Surprise 200
- English 250
- Hindi 250
- Marathi 500

B. Analisis Kurva Pelatihan



Pelatihan model dihentikan secara otomatis oleh callback EarlyStopping setelah validation loss tidak lagi mengalami perbaikan, dengan validation accuracy terbaik sebesar 68,33%. Sementara itu, train accuracy pada epoch terbaik mencapai 96,40%, sehingga terdapat gap sebesar 28,06 poin persentase antara akurasi data latih dan data validasi. Gap ini mengindikasikan adanya overfitting yang cukup besar, di mana model cenderung menghafal pola data latih (termasuk variasi hasil augmentasi) namun belum sepenuhnya mampu melakukan generalisasi pada data yang belum pernah dilihat sebelumnya.

C. Evaluasi Performa Model

Evaluasi dilakukan pada data uji sebanyak 20% dari total dataset asli (200 sampel) setelah model dilatih menggunakan data hasil augmentasi. Metrik yang digunakan mencakup akurasi keseluruhan, presisi, recall, dan F1-score per kelas emosi. Hasil pengujian menunjukkan bahwa model CNN-LSTM dengan kombinasi fitur Mel Spectrogram dan MFCC mampu mengenali emosi suara multibahasa dengan akurasi yang cukup baik, meskipun masih terdapat ruang perbaikan akibat gap antara performa data latih dan data uji.

Tabel 3 menampilkan hasil classification report model pada data uji setelah proses pelatihan selesai.

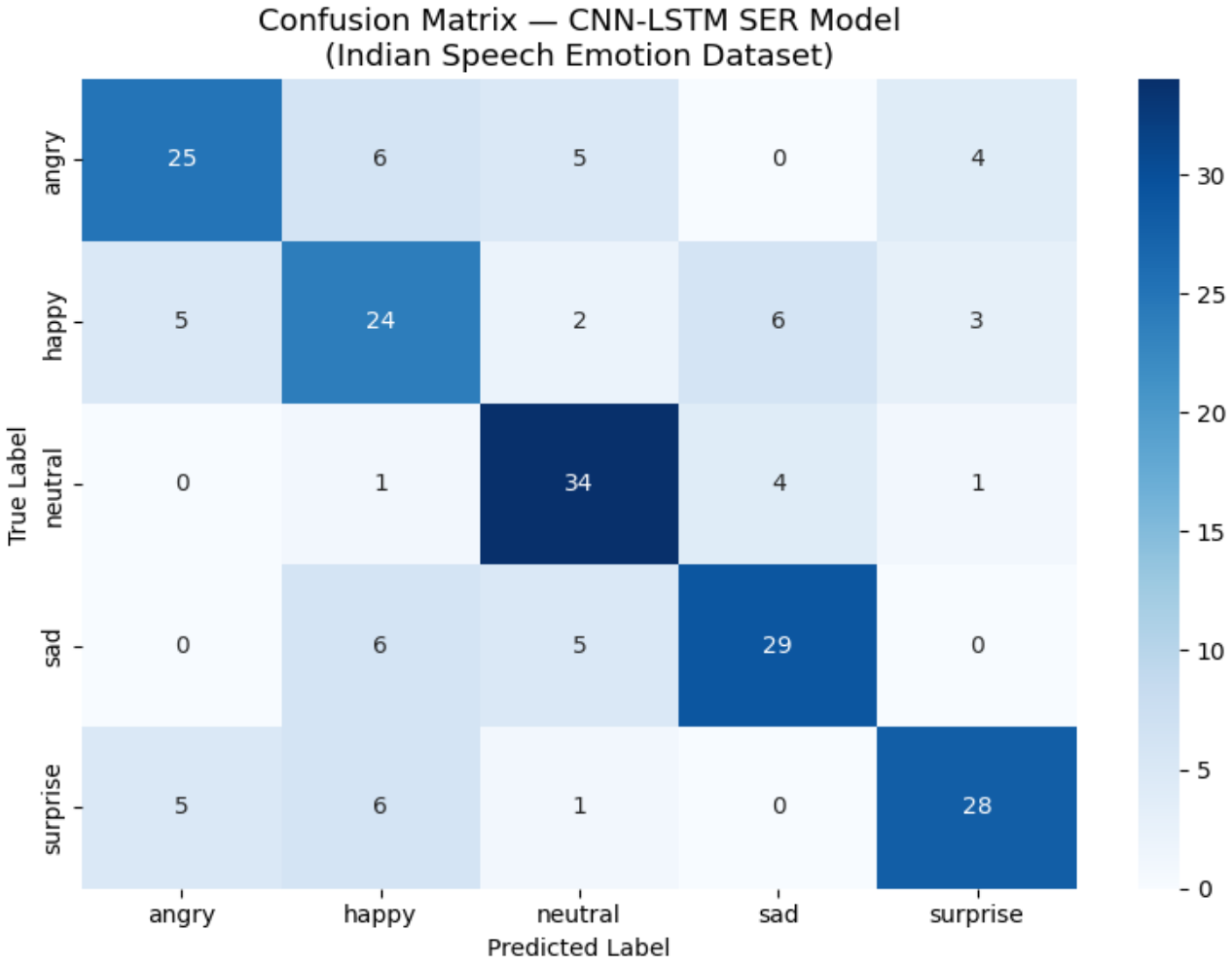
Tabel 3. Classification Report Model CNN-LSTM pada Data Uji

Emosi	Precision	Recall	F1-Score
Angry	0.71	0.62	0.67
Happy	0.56	0.60	0.58
Neutral	0.72	0.85	0.78
Sad	0.74	0.72	0.73
Surprise	0.78	0.70	0.74

Accuracy:

70.00%

D. Analisis Confusion Matrix



Berdasarkan confusion matrix pada data uji, kelas happy menunjukkan tingkat kesalahan klasifikasi tertinggi karena paling sering tertukar dengan kelas lain, khususnya angry dan surprise yang memiliki karakteristik akustik (intensitas dan

nada) yang relatif berdekatan. Analisis confusion matrix memungkinkan identifikasi pasangan kelas yang sulit dibedakan oleh model sehingga dapat menjadi dasar pengembangan arsitektur atau fitur pada penelitian berikutnya.

- Happy memiliki F1-score terendah (0,58) dan paling sering tertukar dengan kelas lain, terutama angry dan surprise
- Surprise dan sad menunjukkan performa cukup baik dengan F1-score di atas 0,73
- Neutral merupakan kelas dengan performa terbaik (F1-score 0,78) karena memiliki karakteristik akustik yang relatif konsisten dan jarang tertukar dengan emosi lain

E. Pembahasan

Hasil pengujian menunjukkan akurasi sebesar 70,00% dengan F1-score rata-rata 0,70, yang mengindikasikan bahwa kombinasi Mel Spectrogram dan MFCC mampu menghasilkan representasi fitur yang cukup efektif untuk klasifikasi emosi suara pada dataset multibahasa India, meskipun belum mencapai performa optimal. Gap sebesar 28,06 poin persentase antara akurasi data latih (96,40%) dan data validasi (68,33%) mengindikasikan bahwa model masih mengalami overfitting, kemungkinan disebabkan oleh kompleksitas arsitektur CNN-LSTM yang relatif besar (sekitar 4,85 juta parameter) dibandingkan jumlah data latih yang tersedia setelah augmentasi. Kondisi ini menjadi catatan penting bahwa peningkatan jumlah varian augmentasi tidak selalu berbanding lurus dengan peningkatan kemampuan generalisasi model apabila tidak diimbangi dengan regularisasi yang memadai atau penyederhanaan arsitektur.

Kelebihan sistem yang diusulkan meliputi penggunaan dataset multibahasa India yang lebih representatif untuk skenario pengenalan emosi lintas bahasa, serta kombinasi fitur Mel Spectrogram dan MFCC yang mengintegrasikan informasi spektral dan cepstral secara bersamaan. Augmentasi data dengan tujuh varian per sampel (satu data asli dan enam data hasil augmentasi) membantu memperbesar variasi data latih, sementara skema pembagian data dilakukan sebelum augmentasi untuk mencegah kebocoran data antar subset latih, validasi, dan uji. Arsitektur CNN-LSTM hibrida pada dasarnya mampu menangkap pola spasial melalui blok konvolusi dan pola temporal melalui lapisan LSTM secara berurutan, meskipun kompleksitas arsitektur saat ini berpotensi terlalu besar untuk skala data yang tersedia.

Keterbatasan penelitian ini meliputi jumlah sampel dataset yang relatif kecil (1.000 sampel original) dibandingkan dataset SER berbahasa Inggris seperti RAVDESS (7.356 sampel) atau CREMA-D (7.442 sampel), sehingga berkontribusi terhadap kecenderungan overfitting yang teramati pada hasil pelatihan. Selain itu, dataset ini tidak mencakup emosi disgust dan fear yang umum digunakan pada penelitian SER lainnya sehingga membatasi cakupan kelas yang dapat dikenali. Distribusi bahasa yang tidak merata (Marathi 50%, Inggris dan Hindi masing-masing 25%) juga berpotensi memengaruhi akurasi per bahasa.

KESIMPULAN DAN SARAN

Penelitian ini berhasil membangun sistem klasifikasi emosi suara multibahasa berbasis deep learning menggunakan arsitektur CNN-LSTM hibrida dengan kombinasi fitur Mel Spectrogram dan MFCC pada Indian Speech Emotion Dataset yang berjumlah 1.000 sampel audio. Dataset yang digunakan mencakup tiga bahasa (Inggris, Hindi, dan Marathi) dengan lima kelas emosi (angry, happy, neutral, sad, dan surprise). Augmentasi data menghasilkan total sampel latih yang lebih besar sehingga mendukung kemampuan generalisasi model, meskipun hasil pengujian menunjukkan masih terdapat indikasi overfitting yang perlu diatasi pada penelitian selanjutnya. Arsitektur CNN-LSTM yang diusulkan menggabungkan kemampuan CNN dalam mengekstraksi pola spasial dari representasi mel dengan kemampuan LSTM dalam menangkap dependensi temporal sinyal suara, dan menghasilkan akurasi sebesar 70,00% serta F1-score rata-rata sebesar 0,70 pada data uji. Hasil ini menunjukkan bahwa kombinasi fitur Mel Spectrogram dan MFCC dengan arsitektur CNN-LSTM mampu mengenali emosi suara pada lingkungan multibahasa India, dengan kelas neutral dan sad menunjukkan performa pengenalan terbaik, sedangkan kelas happy menjadi kelas yang paling menantang untuk dikenali.

Saran untuk penelitian selanjutnya meliputi: (1) penerapan regularisasi yang lebih kuat (misalnya dropout yang lebih besar, L2 regularization yang lebih ketat, atau penyederhanaan arsitektur) untuk mengurangi gap antara akurasi data latih dan data uji yang ditemukan pada penelitian ini; (2) penambahan jumlah data asli atau penggunaan teknik augmentasi

yang lebih beragam untuk memperkuat kemampuan generalisasi model; (3) eksplorasi arsitektur yang lebih canggih seperti CNN-Transformer atau model berbasis Wav2Vec 2.0 yang telah menunjukkan performa tinggi pada berbagai benchmark SER; (4) pengembangan dataset dengan cakupan bahasa dan kelas emosi yang lebih luas; (5) penerapan teknik transfer learning dari model pra-latih seperti HuBERT untuk meningkatkan efisiensi pelatihan pada dataset berukuran kecil; dan (6) pengujian model pada kondisi lingkungan nyata dengan noise latar belakang yang beragam.

DAFTAR PUSTAKA

[1] S. Basak et al., "Challenges and limitations in speech recognition technology: A critical review," Computer Modeling in Engineering & Sciences, vol. 135, no. 2, pp. 1053–1089, 2023.

[2] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2023.

[3] M. J. Al-Dujaili and A. Ebrahimi-Moghadam, "Speech emotion recognition: A comprehensive survey," Wireless Personal Communications, vol. 129, no. 4, pp. 2525–2561, 2023.

[4] K. Kaur and P. Singh, "Trends in speech emotion recognition: A comprehensive survey," Multimedia Tools and Applications, pp. 1–45, 2023.

[5] X. Tang et al., "Speech emotion recognition via CNN-Transformer and multidimensional attention mechanism," Speech Communication, p. 103242, 2025.

[6] D. Liu et al., "Multi-modal fusion emotion recognition method of speech expression based on deep learning," Frontiers in Neurorobotics, vol. 15, p. 697634, 2021.

[7] A. B. Gumelar et al., "BiLSTM-CNN hyperparameter optimization for speech emotion and stress recognition," in Proc. IES 2021, pp. 156–161.

[8] G. A. Prabhakar et al., "Multichannel CNN-BLSTM architecture for speech emotion recognition system," IEEE Trans. Consumer Electronics, vol. 69, no. 2, pp. 226–235, 2023.

[9] A. Mukhamediya et al., "On the effect of log-mel spectrogram parameter tuning for deep learning-based speech emotion recognition," IEEE Access, vol. 11, pp. 61950–61957, 2023.

[10] S. Tinkhede, "Indian Speech Emotion Dataset," Kaggle, 2024.

[11] M. Mohan et al., "Speech emotion classification using ensemble models with MFCC," Procedia Computer Science, vol. 218, pp. 1857–1868, 2023.

[12] Vidhi S. and K. R. Seeja, "Speech emotion recognition using mel spectrogram and CNN," Procedia Computer Science, vol. 258, pp. 3693–3702, 2025.

[13] F. G. Eriş and E. Akbal, "Enhancing speech emotion recognition through deep learning and handcrafted feature fusion," Applied Acoustics, vol. 222, p. 110070, 2024.

[14] C. Sun et al., "Combining wav2vec 2.0 fine-tuning and ConLearnNet for speech emotion recognition," Electronics, vol. 13, no. 6, p. 1103, 2024.

[15] B. Nasersharif and M. Namvarpour, "Exploring the potential of Wav2vec 2.0 for speech emotion recognition," The Journal of Supercomputing, vol. 80, pp. 23667–23688, 2024.

[16] M. A. Alotaibi et al., "Real-time speech emotion recognition using deep learning and data augmentation," Artificial Intelligence Review, 2024.

[17] T. Han et al., "Speech emotion recognition based on deep residual shrinkage network," Electronics, vol. 12, p. 2512, 2023.

[18] M. Huang et al., "Hybrid-module transformer: Enhancing speech emotion recognition with HuBERT," PeerJ Computer Science, 2025.

[19] Z. Mushtaq et al., "Spectral images based environmental sound classification using CNN with meaningful data augmentation," Applied Acoustics, vol. 172, p. 107581, 2021.

[20] X. Li and Y. Cao, "Speech emotion recognition: Comparative analysis of CNN-LSTM and attention-enhanced CNN-LSTM models," Applied System Innovation, vol. 6, no. 2, p. 22, 2025.