JURNAL SISFOTENIKA
Vol. 16 No. 2, Juli 2026
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/sisfotenika.v16i2.662

Deteksi Emosi Suara Berbahasa Indonesia Berbasis Web Menggunakan MFCC dan BiLSTM

Web-Based Indonesian Speech Emotion Recognition Using MFCC and BiLSTM

Yusfi Syawali*[1], Arnita[2]
[1,2]Universitas Negeri Medan;Jl. William Iskandar Ps. V, Kenangan Baru, Kec. Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara 20221, (061)6613365
[1,2]Jurusan Matematika, Program Studi Ilmu Komputer,FMIPA, Medan
e-mail: ***[1]yusfisyawali@gmail.com**, [2]arnita@unimed.ac.id

Abstrak

Deteksi emosi suara berbahasa Indonesia masih menghadapi tantangan karena keterbatasan dataset lokal dan belum banyaknya implementasi dalam bentuk sistem yang dapat digunakan secara langsung. Penelitian ini bertujuan mengembangkan sistem deteksi emosi suara berbahasa Indonesia berbasis web menggunakan ekstraksi fitur Mel-Frequency Cepstral Coefficients (MFCC) dan klasifikasi Bidirectional Long Short-Term Memory (BiLSTM). Dataset yang digunakan merupakan dataset hybrid yang terdiri atas 1.200 berkas audio dari 40 aktor dengan lima kelas emosi, yaitu netral, senang, terkejut, jijik, dan kecewa. Fitur suara diekstraksi menggunakan 40 koefisien MFCC yang dikombinasikan dengan Delta dan Delta-Delta, kemudian diklasifikasikan menggunakan model BiLSTM dengan attention pooling. Hasil evaluasi pada data uji menunjukkan akurasi sebesar 80,56% dan Macro F1-score sebesar 0,8060. Model juga diimplementasikan ke dalam sistem web Serambi Emosi. Hasil pengujian Black-Box menunjukkan seluruh fungsi berjalan valid, sedangkan evaluasi usability menggunakan System Usability Scale memperoleh skor 71,25 dengan kategori Good. Hasil ini menunjukkan bahwa MFCC dan BiLSTM berpotensi digunakan sebagai prototipe pendukung analisis kondisi emosional berbasis suara.

Kata kunci*—Deteksi Emosi Suara, Bahasa Indonesia, MFCC, BiLSTM, Sistem Website*

Abstract

Indonesian speech emotion detection still faces challenges due to the limited availability of local datasets and the lack of practical system implementation. This study aims to develop a web-based Indonesian speech emotion detection system using Mel-Frequency Cepstral Coefficients (MFCC) feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) classification. The dataset used in this study is a hybrid dataset consisting of 1,200 audio files from 40 actors with five emotion classes: neutral, happy, surprised, disgusted, and disappointed. Speech features were extracted using 40 MFCC coefficients combined with Delta and Delta-Delta features, then classified using a BiLSTM model with attention pooling. The evaluation results on the test data achieved an accuracy of 80.56% and a Macro F1-score of 0.8060. The trained model was also implemented into a web-based system called Serambi Emosi. Black-Box Testing showed that all main functions worked properly, while usability evaluation using the System Usability

Scale obtained a score of 71.25, categorized as Good. These results indicate that MFCC and BiLSTM has potential as a prototype to support voice-based emotional state analysis.

Keywords— *Speech Emotion Detection, Indonesian Language, MFCC, BiLSTM, Website System*

1. PENDAHULUAN

Kesehatan mental merupakan salah satu aspek penting dalam kesejahteraan manusia karena berkaitan langsung dengan kemampuan individu dalam mengelola emosi, mengambil keputusan, berinteraksi sosial, dan menjalankan aktivitas sehari-hari. Gangguan emosional seperti kecemasan dan depresi masih menjadi permasalahan yang cukup serius, terutama ketika tidak terdeteksi sejak awal [1]. Di Indonesia, tantangan tersebut semakin kompleks karena akses terhadap tenaga profesional kesehatan mental belum merata, khususnya pada wilayah yang memiliki keterbatasan layanan psikologis. Selain itu, stigma sosial juga dapat menghambat individu untuk mencari bantuan secara langsung [2], [3]. Kondisi ini menunjukkan perlunya alat bantu yang bersifat objektif, mudah diakses, dan dapat digunakan sebagai pendukung awal dalam mengenali kondisi emosional seseorang [4].

Salah satu pendekatan yang dapat digunakan untuk mendukung analisis kondisi emosional adalah Speech Emotion Recognition (SER). SER merupakan bidang penelitian yang berfokus pada pengenalan emosi manusia melalui sinyal suara. Suara tidak hanya memuat informasi linguistik, tetapi juga mengandung informasi paralinguistik seperti intonasi, energi, kecepatan bicara, tekanan suara, dan pola temporal yang dapat merepresentasikan kondisi emosional pembicara [5]. Dibandingkan pendekatan berbasis ekspresi wajah atau gestur, analisis berbasis suara memiliki potensi untuk diterapkan pada berbagai kondisi, termasuk konsultasi jarak jauh, interaksi manusia-komputer, dan sistem pemantauan emosional berbasis aplikasi [6].

Secara umum, sistem SER terdiri atas dua tahapan utama, yaitu ekstraksi fitur dan klasifikasi emosi. Pada tahap ekstraksi fitur, Mel-Frequency Cepstral Coefficients (MFCC) banyak digunakan karena mampu merepresentasikan karakteristik spektral suara berdasarkan pendekatan persepsi pendengaran manusia. MFCC dapat menangkap informasi penting dari sinyal suara, khususnya pada pola frekuensi yang berkaitan dengan karakteristik vokal. Beberapa penelitian menunjukkan bahwa MFCC masih menjadi salah satu fitur akustik yang efektif dan efisien dalam pengenalan suara maupun emosi suara [6], [7]. Kemampuan metode jaringan saraf dalam memodelkan hubungan nonlinear pada data sekuensial dengan karakteristik yang bervariasi telah dibuktikan pada berbagai domain prediksi, termasuk prediksi berbasis data temporal [8]. Pada tahap klasifikasi, model deep learning berbasis data sekuensial seperti Long Short-Term Memory (LSTM) dan Bidirectional Long Short-Term Memory (BiLSTM) banyak digunakan karena mampu mempelajari ketergantungan temporal pada urutan fitur suara. BiLSTM memiliki keunggulan karena memproses informasi secara dua arah, yaitu dari masa lalu ke masa depan dan dari masa depan ke masa lalu, sehingga dapat menangkap konteks temporal secara lebih menyeluruh [9], [10]. Kemampuan LSTM dalam mempelajari dependensi jangka panjang dan pola nonlinear pada data sekuensial juga telah ditunjukkan pada penelitian sebelumnya [11].

Meskipun penelitian SER telah berkembang cukup pesat, sebagian besar penelitian masih menggunakan dataset berbahasa asing seperti RAVDESS dan EmoDB. Kondisi ini menjadi kendala ketika sistem SER ingin diterapkan pada konteks Bahasa Indonesia, karena karakteristik prosodi, pelafalan, intonasi, dan ekspresi emosional dalam bahasa lokal dapat berbeda dari bahasa lain. Beberapa penelitian mulai mengembangkan dataset dan model SER berbahasa Indonesia, salah satunya melalui IndoWaveSentiment yang memuat sampel audio dalam lima kategori emosi [12]. Namun, ketersediaan dataset SER berbahasa Indonesia masih relatif terbatas dan implementasi model ke dalam sistem berbasis web yang dapat digunakan oleh pengguna akhir

juga belum banyak dikembangkan [13].

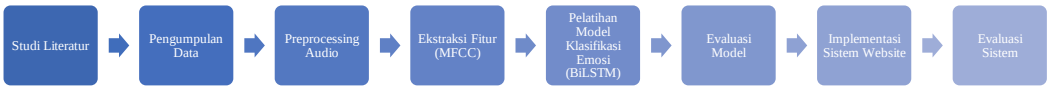
Berdasarkan permasalahan tersebut, penelitian ini mengembangkan sistem deteksi emosi suara berbahasa Indonesia berbasis web menggunakan kombinasi MFCC dan BiLSTM. Penelitian ini menggunakan dataset hybrid berbahasa Indonesia yang terdiri atas data sekunder dan data primer, dengan lima kelas emosi, yaitu netral, senang, terkejut, jijik, dan kecewa. Ekstraksi fitur dilakukan menggunakan MFCC yang diperkaya dengan fitur Delta dan Delta-Delta, sedangkan klasifikasi emosi dilakukan menggunakan model BiLSTM. Selain pengembangan model, penelitian ini juga mengimplementasikan sistem dalam bentuk aplikasi web agar proses deteksi emosi dapat digunakan secara lebih praktis oleh pengguna.

Kontribusi utama penelitian ini terletak pada dua aspek. Pertama, penelitian ini memanfaatkan dataset hybrid berbahasa Indonesia untuk mendukung pengembangan model SER pada konteks lokal. Kedua, penelitian ini tidak hanya berfokus pada evaluasi model, tetapi juga mengimplementasikan model ke dalam sistem berbasis web sebagai prototipe yang dapat digunakan untuk mendukung analisis kondisi emosional secara mandiri. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem SER berbahasa Indonesia serta menjadi dasar bagi penelitian lanjutan dalam bidang analisis emosi berbasis suara.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan kategori penelitian terapan. Pendekatan kuantitatif digunakan karena proses deteksi emosi dilakukan melalui pengolahan sinyal suara menjadi representasi numerik, kemudian diklasifikasikan menggunakan model deep learning. Penelitian ini juga bersifat terapan karena hasil model tidak hanya dievaluasi secara komputasional, tetapi juga diimplementasikan ke dalam sistem berbasis web. Secara umum, penelitian terdiri atas dua komponen utama, yaitu eksperimen machine learning untuk membangun model deteksi emosi suara dan rekayasa perangkat lunak untuk mengembangkan sistem web Serambi Emosi.

Tahapan penelitian meliputi pengumpulan dataset, pra-pemrosesan audio, augmentasi data latih, ekstraksi fitur MFCC, pelatihan model BiLSTM, evaluasi model klasifikasi, implementasi sistem web, serta evaluasi fungsionalitas dan usability sistem. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur penelitian deteksi emosi suara berbasis web menggunakan MFCC-BiLSTM

2.1 Dataset Penelitian

Dataset yang digunakan dalam penelitian ini merupakan dataset hybrid berbahasa Indonesia yang menggabungkan data sekunder dan data primer. Data sekunder berasal dari dataset IndoWaveSentiment yang terdiri atas 300 berkas audio dari 10 aktor [12]. Untuk memperluas variasi pembicara, dilakukan pengumpulan data primer dari 30 aktor tambahan dengan mengikuti struktur emosi dan format perekaman yang sama. Dengan demikian, total dataset yang digunakan berjumlah 1.200 berkas audio dari 40 aktor.

Dataset mencakup lima kategori emosi, yaitu netral, senang, terkejut, jijik, dan kecewa. Kelima kelas tersebut dipilih karena sesuai dengan struktur kelas pada IndoWaveSentiment dan memiliki karakteristik vokal yang dapat dianalisis melalui sinyal suara [12]. Setiap berkas audio diberi label berdasarkan kelas emosi, identitas aktor, intensitas, dan pengulangan rekaman.

Pembagian dataset dilakukan menggunakan pendekatan speaker-independent split, yaitu seluruh rekaman dari satu aktor ditempatkan hanya pada satu subset data. Strategi ini digunakan untuk mencegah kebocoran data antar subset dan memastikan evaluasi model lebih objektif terhadap pembicara yang tidak muncul pada data latih. Pembagian dataset ditunjukkan pada Tabel 1.

Tabel 1. Pembagian Dataset Penelitian

Subset	Jumlah Aktor	Jumlah Berkas Audio	Persentase	Keterangan
Train	28 aktor	840 berkas	70%	Digunakan untuk pelatihan model
Validasi	6 aktor	180 berkas	15%	Digunakan untuk pemantauan performa model
Test	6 aktor	180 berkas	15%	Digunakan untuk evaluasi akhir model
Total	40 aktor	1.200 berkas	100%	Dataset keseluruhan

2.2 Pra-pemrosesan Audio

Tahap pra-pemrosesan dilakukan untuk menyeragamkan format audio dan mengurangi gangguan yang dapat memengaruhi proses ekstraksi fitur. Seluruh berkas audio dikonversi ke format WAV, sampling rate 16 kHz, dan satu kanal mono. Normalisasi amplitudo dilakukan untuk menyamakan skala intensitas antar sampel sehingga model tidak terlalu dipengaruhi oleh perbedaan keras atau lemahnya suara.

Selain itu, dilakukan proses reduksi noise untuk menekan gangguan latar yang tidak relevan dengan karakteristik emosi. Reduksi noise dilakukan menggunakan pendekatan bandpass filter atau spectral gating dengan mempertahankan rentang frekuensi yang relevan terhadap suara manusia. Setelah audio dinormalisasi dan dibersihkan, data disiapkan untuk proses augmentasi dan ekstraksi fitur.

2.3 Augmentasi Data

Augmentasi data diterapkan hanya pada subset pelatihan untuk meningkatkan variasi data dan mengurangi risiko overfitting. Subset validasi dan test tidak diberi augmentasi agar evaluasi model tetap dilakukan pada data asli. Teknik augmentasi yang digunakan meliputi pitch shifting, time stretching, additive noise, dan gain perturbation.

Pitch shifting digunakan untuk menghasilkan variasi tinggi-rendah nada suara, time stretching digunakan untuk memodelkan variasi kecepatan bicara, additive noise digunakan untuk mensimulasikan kondisi rekaman yang tidak sepenuhnya bersih, sedangkan gain perturbation digunakan untuk memodifikasi intensitas suara secara moderat. Augmentasi dilakukan secara probabilistik dengan peluang augmentasi sampel sebesar 0,7. Setiap jenis transformasi memiliki peluang pemilihan sebesar 0,3, dengan maksimal dua transformasi pada satu sampel. Kombinasi time stretching dan gain perturbation dihindari dalam satu sampel yang sama agar perubahan tempo dan intensitas tidak terlalu agresif.

2.4 Ekstraksi Fitur MFCC

Ekstraksi fitur dilakukan menggunakan Mel-Frequency Cepstral Coefficients (MFCC). MFCC dipilih karena mampu merepresentasikan karakteristik spektral suara berdasarkan persepsi pendengaran manusia. Dalam penelitian ini, setiap audio diekstraksi menjadi 40 koefisien MFCC. Untuk memperkaya informasi temporal, fitur MFCC dikombinasikan dengan fitur Delta dan

Delta-Delta. Fitur Delta merepresentasikan perubahan koefisien antar frame, sedangkan Delta-Delta merepresentasikan perubahan tingkat kedua dari fitur tersebut.

Gabungan MFCC, Delta, dan Delta-Delta menghasilkan 120 fitur pada setiap frame audio. Karena jumlah frame pada setiap audio dapat berbeda akibat variasi durasi sinyal, dilakukan penyeragaman panjang urutan menggunakan nilai maksimum 119 frame. Dengan demikian, setiap sampel audio direpresentasikan dalam bentuk tensor berukuran (119, 120), yaitu 119 langkah waktu dan 120 fitur pada setiap frame. Bentuk ini sesuai dengan kebutuhan input model BiLSTM yang memproses data dalam format sekuensial.

2.5 Arsitektur dan Pelatihan Model BiLSTM

Model klasifikasi yang digunakan adalah Bidirectional Long Short-Term Memory (BiLSTM). BiLSTM digunakan karena mampu mempelajari pola temporal dari dua arah, yaitu arah maju dan arah mundur, sehingga model dapat menangkap konteks sinyal suara secara lebih menyeluruh. Input model berupa tensor fitur hasil ekstraksi MFCC dengan ukuran (119, 120). Ringkasan arsitektur model BiLSTM yang digunakan ditunjukkan pada Tabel 2. Arsitektur terdiri atas 11 lapisan dengan total parameter yang disesuaikan untuk klasifikasi lima kelas emosi.

Tabel 2. Ringkasan Arsitektur Model BiLSTM

No	Lapisan	Konfigurasi	Fungsi
1	Input Layer	Shape (119, 120)	Menerima tensor fitur MFCC+Delta+Delta-Delta
2	Masking Layer	mask_value=0.0	Mengabaikan nilai padding
3	SpatialDropout1D	rate=0,3	Mengurangi ketergantungan pada fitur tertentu
4	Bidirectional LSTM 1	128 unit, return_seq=True	Menangkap pola temporal dua arah (tingkat awal)
5	Bidirectional LSTM 2	96 unit, return_seq=True	Menangkap pola temporal lanjutan
6	Masked Attention Pooling	attention_units=64	Merangkum frame paling informatif terhadap emosi
7	Dense 1	128 unit, ReLU	Pemetaan fitur ke ruang klasifikasi
8	Dropout 1	rate=0,4	Mengurangi overfitting
9	Dense 2	96 unit, ReLU	Pemrosesan fitur akhir sebelum output
10	Dropout 2	rate=0,4	Mengurangi overfitting
11	Output Layer	5 unit, Softmax	Menghasilkan probabilitas lima kelas emosi

Arsitektur model terdiri atas input layer, masking layer, SpatialDropout1D, dua lapis Bidirectional LSTM, attention pooling, dense layer, dropout, dan output layer. Masking layer digunakan untuk mengabaikan nilai padding pada urutan audio. SpatialDropout1D dan dropout digunakan untuk mengurangi risiko overfitting [14]. Attention pooling digunakan untuk membantu model memberikan bobot lebih besar pada bagian frame yang lebih informatif terhadap emosi. Output layer menggunakan fungsi aktivasi softmax untuk menghasilkan probabilitas pada lima kelas emosi. Kemampuan jaringan saraf dalam mempelajari pola nonlinear melalui proses backpropagation pada berbagai konfigurasi arsitektur telah ditunjukkan pada penelitian sebelumnya di berbagai domain prediksi berbasis data [15].

Tabel 3. Konfigurasi Pelatihan Model

Komponen	Nilai

Model	BiLSTM dengan attention pooling
Input	(119, 120)
Jumlah kelas	5 kelas emosi
Optimizer	Adam
Learning rate	0,0004
Batch size	32
Epoch maksimum	150
Loss function	Categorical cross-entropy
Label smoothing	0,08
Metrik utama	Accuracy, Precision, Recall, Macro-F1
Dasar pemilihan model	Validation Macro-F1

Pelatihan model dilakukan menggunakan optimizer Adam dengan learning rate 0,0004, batch size 32, dan maksimum 150 epoch. Fungsi loss yang digunakan adalah categorical cross-entropy karena tugas yang dilakukan merupakan klasifikasi multikelas [16]. Label smoothing sebesar 0,08 diterapkan untuk mengurangi kecenderungan model terlalu yakin terhadap satu kelas tertentu. Pemilihan model terbaik dilakukan berdasarkan nilai validation Macro-F1 agar evaluasi tidak hanya bergantung pada akurasi, tetapi juga memperhatikan keseimbangan performa antar kelas.

2.6 Implementasi Sistem Web

Model hasil pelatihan diimplementasikan ke dalam sistem berbasis web bernama Serambi Emosi. Sistem ini dirancang agar pengguna dapat mengunggah atau merekam suara, kemudian memperoleh hasil prediksi emosi secara langsung melalui antarmuka web. Secara umum, sistem terdiri atas antarmuka pengguna, backend pemrosesan audio, modul ekstraksi fitur, model klasifikasi BiLSTM, dan halaman hasil prediksi.

Proses kerja sistem dimulai ketika pengguna memasukkan audio melalui halaman input. Audio yang diterima kemudian divalidasi berdasarkan format dan kelayakan input. Setelah validasi berhasil, audio diproses melalui tahap pra-pemrosesan dan ekstraksi fitur MFCC dengan konfigurasi yang sama seperti pada tahap pelatihan model. Fitur yang dihasilkan dimasukkan ke model BiLSTM untuk memperoleh probabilitas masing-masing kelas emosi. Hasil prediksi kemudian ditampilkan kepada pengguna dalam bentuk label emosi dan informasi hasil analisis. Implementasi model klasifikasi ke dalam sistem berbasis web juga telah diterapkan pada penelitian sebelumnya untuk memungkinkan pengguna memasukkan data dan memperoleh hasil klasifikasi secara langsung melalui antarmuka sistem [17].

2.7 Evaluasi Model dan Sistem

Evaluasi dilakukan pada dua aspek, yaitu evaluasi model klasifikasi dan evaluasi sistem web. Evaluasi model dilakukan menggunakan data test yang tidak digunakan pada proses pelatihan maupun validasi. Metrik evaluasi yang digunakan meliputi accuracy, precision, recall, F1-score, Macro F1-score, dan confusion matrix. Penggunaan Macro F1-score penting karena dataset memiliki lima kelas emosi, sehingga performa setiap kelas perlu diperhatikan secara seimbang.

Evaluasi sistem web dilakukan menggunakan Black-Box Testing dan System Usability Scale (SUS) . Black-Box Testing digunakan untuk memastikan setiap fitur sistem berjalan sesuai kebutuhan, seperti input audio, validasi format, proses prediksi, penanganan kesalahan, dan tampilan hasil analisis. Sementara itu, SUS digunakan untuk menilai tingkat kemudahan penggunaan sistem dari sudut pandang pengguna [18]. Evaluasi SUS dilakukan kepada 30 responden setelah menggunakan sistem, kemudian skor dihitung dalam rentang 0–100 untuk menentukan tingkat usability sistem.

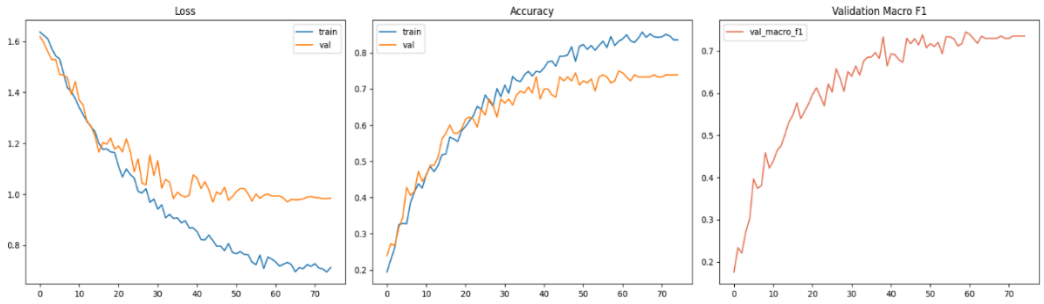
3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil pelatihan dan evaluasi model BiLSTM, analisis klasifikasi pada setiap kelas emosi, implementasi sistem web Serambi Emosi, serta evaluasi fungsionalitas dan usability sistem. Hasil disajikan dalam bentuk tabel, gambar, dan uraian untuk menjelaskan capaian model serta keterbatasan yang muncul pada proses deteksi emosi suara berbahasa Indonesia.

3.1 Hasil Pelatihan dan Evaluasi Model

Model BiLSTM dilatih menggunakan fitur MFCC yang diperkaya dengan Delta dan Delta-Delta. Setiap sampel audio direpresentasikan dalam bentuk tensor berukuran (119, 120), yaitu 119 langkah waktu dan 120 fitur pada setiap frame. Pemilihan model terbaik dilakukan berdasarkan nilai validation Macro-F1 agar performa model tidak hanya bergantung pada akurasi, tetapi juga mempertimbangkan keseimbangan kinerja pada seluruh kelas emosi.

Proses pelatihan model berlangsung selama 150 epoch dengan pemilihan model terbaik berdasarkan nilai validation Macro-F1 tertinggi. Gambar 2 menyajikan grafik riwayat pelatihan yang mencakup kurva loss, akurasi, dan validation Macro-F1 per epoch.



Gambar 2. Kurva Training vs Validation: Loss, Accuracy, dan Macro F1-score per Epoch

Berdasarkan grafik pada Gambar 2, training loss menurun secara konsisten dari sekitar 1,63 pada epoch awal menjadi sekitar 0,71 pada epoch akhir, sedangkan validation loss turun hingga mendekati 0,98. Penurunan yang terjadi pada kedua kurva loss menunjukkan bahwa model berhasil mempelajari pola data secara bertahap. Jarak antara kurva training dan validation mengindikasikan adanya kecenderungan overfitting ringan pada fase akhir pelatihan, namun hal ini telah diantisipasi melalui penerapan SpatialDropout1D, Dropout, dan label smoothing pada arsitektur model.

Pada grafik akurasi, training accuracy meningkat dari sekitar 0,19 menjadi 0,84–0,85, sedangkan validation accuracy meningkat dari 0,24 menjadi 0,75 pada epoch terbaik. Grafik validation Macro-F1 juga menunjukkan tren konsisten meningkat dari sekitar 0,18 hingga 0,74 pada epoch ke-60, yang kemudian dipilih sebagai model final. Pemilihan berdasarkan Macro-F1 dilakukan agar model terbaik tidak hanya memiliki akurasi total yang tinggi, tetapi juga menunjukkan performa yang seimbang pada seluruh lima kelas emosi.

Berdasarkan hasil pelatihan, model terbaik diperoleh pada epoch ke-60 dengan validation accuracy sebesar 0,7500 dan validation Macro-F1 sebesar 0,7449. Setelah model terbaik diperoleh, evaluasi akhir dilakukan menggunakan data uji yang terdiri atas 180 sampel audio dari 6 aktor yang tidak digunakan pada proses pelatihan maupun validasi. Hasil evaluasi umum pada data uji ditunjukkan pada Tabel 4.

Tabel 4. Hasil Pengujian Umum Model pada Data Uji

Metrik	Nilai
Test Accuracy	0,8056
Macro Precision	0,8083

Macro Recall	0,8056
Macro F1-score	0,8060

Berdasarkan Tabel 4, model memperoleh test accuracy sebesar 0,8056 atau 80,56% dan Macro F1-score sebesar 0,8060. Hasil ini menunjukkan bahwa model mampu mengenali emosi suara pada data uji dengan performa yang cukup baik. Kedekatan nilai accuracy dan Macro F1-score juga menunjukkan bahwa performa model relatif seimbang pada keseluruhan kelas emosi. Namun, nilai akurasi data uji yang lebih tinggi dibandingkan validasi tetap perlu ditafsirkan secara hati-hati karena subset validasi dan uji masing-masing hanya terdiri atas 6 aktor, sehingga variasi karakteristik suara antaraktor dapat memengaruhi tingkat kesulitan setiap subset.

3.2 Analisis Klasifikasi Setiap Kelas Emosi

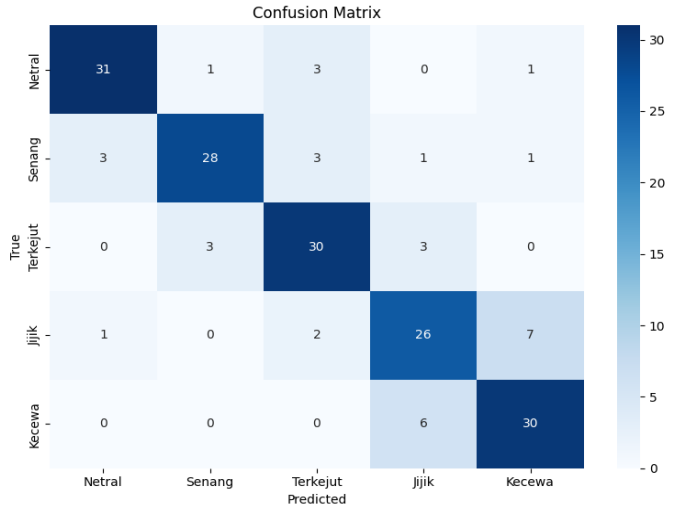
Evaluasi lebih rinci dilakukan untuk mengetahui performa model pada setiap kelas emosi. Hasil classification report model pada data uji ditunjukkan pada Tabel 5.

Tabel 5. Classification Report Model pada Data Uji

Kelas Emosi	Precision	Recall	F1-score	Support
Netral	0,8857	0,8611	0,8732	36
Senang	0,8750	0,7778	0,8235	36
Terkejut	0,7895	0,8333	0,8108	36
Jijik	0,7222	0,7222	0,7222	36
Kecewa	0,7692	0,8333	0,8000	36
Macro Average	0,8083	0,8056	0,8060	180
Weighted Average	0,8083	0,8056	0,8060	180

Berdasarkan Tabel 5, kelas Netral memperoleh performa terbaik dengan F1-score sebesar 0,8732. Hal ini menunjukkan bahwa model cukup baik dalam mengenali karakteristik suara netral yang cenderung memiliki pola vokal lebih stabil. Kelas Senang, Terkejut, dan Kecewa juga menunjukkan performa cukup baik dengan F1-score masing-masing sebesar 0,8235, 0,8108, dan 0,8000. Sementara itu, kelas Jijik memperoleh F1-score terendah, yaitu 0,7222, sehingga menjadi kelas yang paling sulit dikenali oleh model.

Evaluasi lebih rinci dilakukan menggunakan confusion matrix untuk menganalisis distribusi prediksi benar dan salah pada masing-masing kelas emosi. Gambar 3 menyajikan heatmap confusion matrix model pada data uji.



Gambar 3. Confusion matrix model klasifikasi emosi suara

Berdasarkan confusion matrix pada Gambar 3, dari total 180 sampel uji, model berhasil mengklasifikasikan 145 sampel dengan benar, sehingga diperoleh error rate keseluruhan sebesar 19,44% (35 sampel salah klasifikasi). Prediksi benar tertinggi terdapat pada kelas Netral (31/36 sampel), diikuti kelas Terkejut dan Kecewa (masing-masing 30/36), kelas Senang (28/36), dan kelas Jijik sebagai yang terendah (26/36 sampel).

Analisis mendalam terhadap pola kesalahan klasifikasi dilakukan menggunakan kerangka valence-arousal model dua dimensi yang merepresentasikan emosi berdasarkan valensi (positif–negatif) dan tingkat arousal (tinggi–rendah). Ringkasan kesalahan klasifikasi dominan beserta analisis penyebabnya disajikan pada Tabel 6.

Tabel 6. Analisis Kesalahan Klasifikasi Dominan Berdasarkan Kerangka Valence-Arousal

Kelas Aktual	**Kelas Prediksi**	**Jumlah**	**Error Rate**	**Analisis Penyebab**
Jijik	Kecewa	7	19,4%	Kedua emosi bervalensi negatif, arousal rendah–sedang; pola pitch dan energi tumpang tindih di ruang fitur MFCC
Kecewa	Jijik	6	16,7%	Sama dengan pasangan di atas; konfusi bersifat simetris antar dua kelas negatif
Senang	Terkejut	3	8,3%	Keduanya bervalensi positif dengan arousal tinggi; peningkatan energi dan pitch yang serupa
Senang	Netral	3	8,3%	Ekspresi Senang yang kurang menonjol menyerupai karakteristik vokal Netral
Terkejut	Senang	3	8,3%	Kemiripan dinamika temporal dan intensitas suara pada kedua emosi
Terkejut	Jijik	3	8,3%	Kedekatan pola spektral tertentu pada frekuensi menengah
Netral	Terkejut	3	8,3%	Kemiripan sebagian pola intensitas pada segmen tertentu

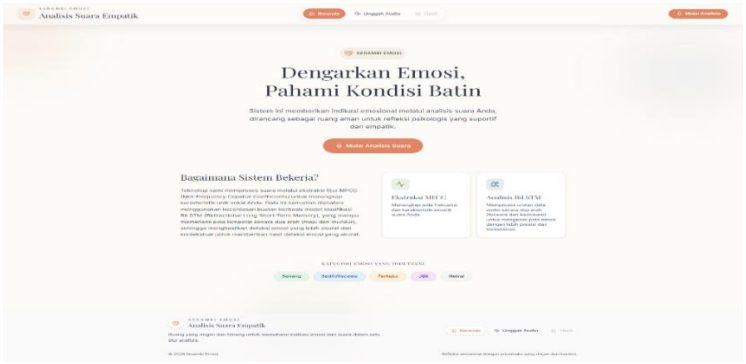
Berdasarkan Tabel 6, kesalahan terbesar terjadi pada pasangan kelas Jijik dan Kecewa (masing-masing 7 dan 6 kesalahan, berkontribusi 36,1% dari total error). Hal ini dapat dijelaskan secara teoritis: kedua emosi tersebut sama-sama menempati kuadran valensi negatif dengan arousal rendah hingga sedang dalam ruang valence-arousal. Kelas Jijik cenderung disertai ketegangan vokal yang teredam, sementara Kecewa ditandai dengan penurunan energi dan nada yang lesu. Kedekatan posisi keduanya menyebabkan representasi akustiknya khususnya pola pitch, energi, dan dinamika temporal menjadi tumpang tindih di ruang fitur MFCC.

Kesalahan juga muncul pada pasangan Senang dan Terkejut (masing-masing 3 kesalahan), yang dapat dijelaskan oleh kesamaan posisi di kuadran valensi positif dengan arousal tinggi keduanya ditandai oleh peningkatan energi dan pitch yang tajam. Temuan ini sejalan dengan Chowdhury dkk. (2025) yang menyatakan bahwa emosi dengan arousal dan valensi yang serupa cenderung sulit dibedakan oleh model berbasis akustik, karena perbedaannya lebih bersifat nuansif dan sulit ditangkap hanya melalui fitur spektral [19]. Secara keseluruhan, error rate 19,44% yang tercatat pada penelitian ini bukan semata-mata mencerminkan kelemahan model, melainkan juga merupakan tantangan yang melekat pada tugas Speech Emotion Recognition, khususnya pada emosi-emosi yang berdekatan secara dimensional.

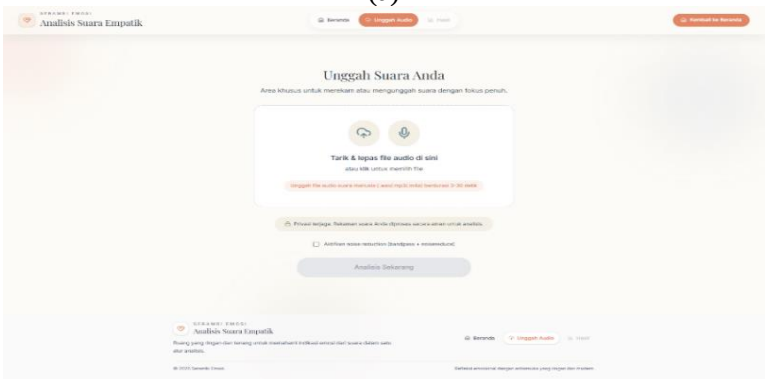
3.3 Implementasi Sistem Web Serambi Emosi

Model BiLSTM yang telah dilatih diimplementasikan ke dalam sistem berbasis web bernama Serambi Emosi. Sistem ini dirancang agar pengguna dapat mengunggah audio, menjalankan proses analisis, dan memperoleh hasil prediksi emosi melalui antarmuka web. Secara umum, sistem terdiri atas antarmuka pengguna, backend pemrosesan audio, modul pra-pemrosesan, ekstraksi fitur MFCC, model klasifikasi BiLSTM, dan halaman hasil prediksi.

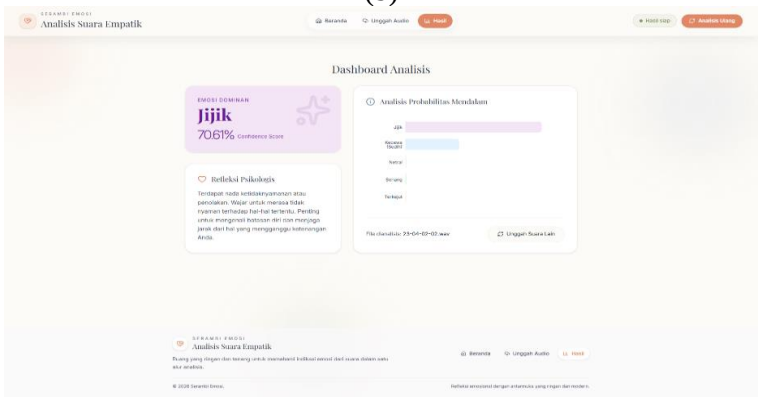
Sistem mendukung input audio dalam format .wav, .mp3, dan .m4a. Audio yang dimasukkan pengguna divalidasi terlebih dahulu berdasarkan format, ukuran, dan kelayakan file. Setelah validasi berhasil, audio diproses melalui tahapan pra-pemrosesan dan ekstraksi fitur MFCC dengan konfigurasi yang sama seperti pada tahap pelatihan. Fitur yang dihasilkan kemudian dimasukkan ke dalam model BiLSTM untuk memperoleh hasil prediksi. Output sistem ditampilkan dalam bentuk emosi dominan, confidence score, dan distribusi probabilitas untuk seluruh kelas emosi.



(a)



(b)



(c)

Gambar 4. Tampilan sistem web Serambi Emosi: (a) landing page, (b) halaman input audio, dan (c) halaman hasil analisis

Gambar 4 menunjukkan tampilan utama sistem web Serambi Emosi. Halaman awal berfungsi memperkenalkan sistem dan kelas emosi yang dapat dideteksi. Halaman input audio digunakan untuk mengunggah suara dan menjalankan proses analisis. Halaman hasil analisis menampilkan emosi dominan, tingkat keyakinan model, serta distribusi probabilitas dari setiap kelas emosi. Implementasi ini menunjukkan bahwa model deteksi emosi suara tidak hanya dapat dievaluasi pada lingkungan eksperimen, tetapi juga dapat diintegrasikan ke dalam sistem web yang dapat digunakan oleh pengguna akhir.

3.4 Evaluasi Sistem Web

Evaluasi sistem web dilakukan menggunakan Black-Box Testing dan System Usability Scale (SUS). Black-Box Testing digunakan untuk memastikan fungsi utama sistem berjalan sesuai kebutuhan, sedangkan SUS digunakan untuk menilai tingkat kemudahan penggunaan sistem dari sudut pandang pengguna. Ringkasan hasil evaluasi sistem ditunjukkan pada Tabel 7.

Tabel 7. Ringkasan Evaluasi Sistem Web

Aspek Evaluasi	Hasil
Jumlah skenario Black-Box Testing	13
Skenario valid	13
Skenario tidak valid	0
Persentase keberhasilan Black-Box Testing	100%
Jumlah responden SUS	30
Skor minimum SUS	37,5
Skor maksimum SUS	92,5
Rata-rata skor SUS	71,25
Kategori usability	Good / Acceptable

Berdasarkan Tabel 7, seluruh skenario Black-Box Testing dinyatakan valid dengan tingkat keberhasilan 100%. Hasil ini menunjukkan bahwa fungsi utama sistem, mulai dari input audio, validasi file, proses analisis, hingga tampilan hasil prediksi, telah berjalan sesuai dengan rancangan. Sementara itu, evaluasi usability menggunakan SUS memperoleh skor rata-rata sebesar 71,25 dari 30 responden. Nilai tersebut berada pada kategori Good dan termasuk dalam tingkat acceptable, sehingga sistem dapat dinilai cukup mudah digunakan oleh pengguna.

Secara keseluruhan, hasil penelitian menunjukkan bahwa kombinasi MFCC, Delta, Delta-Delta, dan BiLSTM mampu digunakan untuk deteksi emosi suara berbahasa Indonesia dengan performa yang cukup baik. Selain itu, implementasi model ke dalam sistem web memperkuat kontribusi penelitian karena sistem yang dikembangkan tidak hanya berhenti pada evaluasi model, tetapi juga diwujudkan dalam bentuk prototipe yang dapat digunakan secara praktis. Meskipun demikian, sistem ini masih memiliki keterbatasan, terutama pada jumlah aktor uji yang relatif terbatas dan performa kelas Jijik yang masih lebih rendah dibandingkan kelas lainnya. Oleh karena itu, pengembangan lanjutan diperlukan melalui perluasan dataset, penambahan variasi kondisi rekaman, dan pengujian pada konteks penggunaan nyata.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa:

1. Kombinasi MFCC, Delta, Delta-Delta, dan BiLSTM dengan attention pooling mampu digunakan untuk deteksi emosi suara berbahasa Indonesia. Model memperoleh accuracy sebesar 80,56% dan Macro F1-score sebesar 0,8060 pada data uji, sehingga menunjukkan performa yang cukup baik dalam mengenali lima kelas emosi, yaitu netral, senang, terkejut, jijik, dan kecewa.
2. Hasil evaluasi per kelas menunjukkan bahwa emosi Netral memperoleh performa terbaik dengan F1-score sebesar 0,8732, sedangkan emosi Jijik memperoleh performa terendah dengan F1-score sebesar 0,7222. Kesalahan klasifikasi paling dominan terjadi antara

kelas Jijik dan Kecewa, yang mengindikasikan adanya kemiripan karakteristik akustik pada emosi bernuansa negatif.

3. Model yang dikembangkan berhasil diimplementasikan ke dalam sistem web Serambi Emosi. Hasil Black-Box Testing menunjukkan seluruh fungsi utama sistem berjalan valid, sedangkan evaluasi usability memperoleh skor SUS sebesar 71,25 dengan kategori Good dan tingkat acceptable. Dengan demikian, sistem ini dapat digunakan sebagai prototipe pendukung analisis kondisi emosional berbasis suara, meskipun belum ditujukan sebagai alat diagnosis klinis.

5. SARAN

Berdasarkan keterbatasan penelitian, beberapa saran untuk pengembangan selanjutnya adalah sebagai berikut:

1. Penelitian selanjutnya dapat memperluas dataset dengan menambah jumlah aktor, variasi usia, gender, aksen, serta kondisi perekaman yang lebih beragam agar model memiliki kemampuan generalisasi yang lebih baik terhadap penutur baru.
2. Pengembangan model dapat dilakukan dengan menambahkan kelas emosi lain, seperti marah, sedih, takut, atau tenang, sehingga sistem mampu mengenali spektrum emosi yang lebih luas dan lebih mendekati kondisi penggunaan nyata.
3. Penelitian berikutnya dapat membandingkan metode MFCC-BiLSTM dengan arsitektur lain seperti CNN-BiLSTM, Transformer, atau model berbasis self-attention untuk mengetahui pendekatan yang paling optimal dalam deteksi emosi suara berbahasa Indonesia.
4. Sistem web Serambi Emosi dapat dikembangkan lebih lanjut melalui pengujian pada lingkungan nyata dengan melibatkan lebih banyak pengguna, termasuk praktisi psikologi, agar evaluasi usability dan manfaat sistem dapat diperoleh secara lebih komprehensif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, atas dukungan dalam pelaksanaan penelitian ini. Penulis juga mengucapkan terima kasih kepada dosen pembimbing, responden, serta seluruh pihak yang telah membantu dalam proses pengumpulan data, pengembangan sistem, dan evaluasi penelitian.

REFERENSI

[1] O. K. Sari, N. Ramdhani, and S. Subandi, “Kesehatan Mental di Era Digital: Peluang Pengembangan Layanan Profesional Psikolog,” *Media Penelitian dan Pengembangan Kesehatan*, vol. 30, no. 4, Dec. 2020, doi: 10.22435/mpk.v30i4.3311.

[2] World Health Organization, “Mental disorders,” World Health Organization. Accessed: Jan. 15, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[3] Badan Kebijakan Pembangunan Kesehatan, “Laporan Tematik SKI,” Kementrian Kesehatan Republik Indonesia. Accessed: Jun. 17, 2026. [Online]. Available: https://www.badankebijakan.kemkes.go.id/laporan-tematik-ski/

[4] S. Bambang, M. Irawaty, S. N. Aizzun, and S. Sugiono, “Sistem pengenal emosi berbasis suara menggunakan ekstraksi ciri Fast Fourier Transform,” *Vortex*, vol. 5, no. 1, Mar. 2024, doi: 10.28989/vortex.v5i1.2092.

[5] Anandappa and K. Mudnal, “Analysis Of Emotions Through Speech Recognition,” *Journal of Scientific Research and Technology*, pp. 30–34, Apr. 2024, doi: 10.61808/jsrt95.

[6] S. Madanian *et al.*, "Speech emotion recognition using machine learning — A systematic review," *Intelligent Systems with Applications*, vol. 20, p. 200266, Nov. 2023, doi: 10.1016/j.iswa.2023.200266.

[7] G. Ajinurseto, L. O. Bakrim, and N. Islamuddin, "Penerapan Metode Mel Frequency Cepstral Coefficients pada Sistem Pengenalan Suara Berbasis Desktop," *Infomatek*, vol. 25, no. 1, pp. 11–20, Jun. 2023, doi: 10.23969/infomatek.v25i1.6109.

[8] M. As, T. Mine, and T. Yamaguchi, "Prediction of Bus Travel Time Over Unstable Intervals between Two Adjacent Bus Stops," *International Journal of Intelligent Transportation Systems Research*, vol. 18, no. 1, pp. 53–64, Jan. 2020, doi: 10.1007/s13177-018-0169-3.

[9] Z. Kh. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[10] D. R. Alghifari, M. Edi, and L. Firmansyah, "Implementasi Bidirectional LSTM untuk Analisis Sentimen Terhadap Layanan Grab Indonesia," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 12, no. 2, pp. 89–99, Sep. 2022, doi: 10.34010/jamika.v12i2.7764.

[11] M. S. Sinaga, S. Iskandar, S. Manullang, A. Arnita, F. Marpaung, and F. Buulolo, "ENHANCING LQ45 STOCK PRICE FORECASTING USING LSTM MODEL," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 20, no. 1, pp. 0427–0438, Nov. 2025, doi: 10.30598/barekengvol20iss1pp0427-0438.

[12] A. Bustamin, A. M. Rizky, E. Warni, I. S. Areni, and Indrabayu, "IndoWaveSentiment: Indonesian audio dataset for emotion classification," *Data Brief*, vol. 57, p. 111138, Dec. 2024, doi: 10.1016/j.dib.2024.111138.

[13] 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, Dec. 2021, doi: 10.26760/mindjournal.v6i2.194-204.

[14] S. Rahmadani, C. S. Rahayu, A. Salim, and K. N. Cahyo, "Deteksi Emosi Berdasarkan Wicara Menggunakan Deep Learning Model," *Jurnal Informatika, Teknologi dan Sains*, vol. 4, no. 3, pp. 220–224, Aug. 2022, doi: 10.51401/jinteks.v4i3.1952.

[15] P. D. Sari and F. Ahyaningsih, "Jaringan Saraf Tiruan Backpropagation untuk Prediksi Harga Bahan Pangan di Wilayah Kabupaten Deli Serdang," *Algoritma : Jurnal Matematika, Ilmu pengetahuan Alam, Kebumian dan Angkasa*, vol. 2, no. 6, pp. 105–117, Nov. 2024, doi: 10.62383/algoritma.v2i6.287.

[16] N. Y. Hussain, "Deep Learning Architectures Enabling Sophisticated Feature Extraction and Representation for Complex Data Analysis," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 2290–2300, Nov. 2024, doi: 10.38124/ijisrt/IJISRT24OCT1521.

[17] W. P. Ananta, K. Saputra, S. Iskandar, F. Ahyaningsih, and S. Manullang, "Implementasi Sistem Klasifikasi Jenis Kayu Jati dan Akasia menggunakan Convolutional Neural Network Berbasis Web," *Jurnal Sistem Informasi Kaputama (JSIK)*, vol. 10, no. 1, pp. 35–43, Jan. 2026, doi: 10.59697/jsik.v10i1.1230.

[18] J. Brooke, "SUS -- a quick and dirty usability scale," 1996, pp. 189–194.

[19] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," *Sci. Rep.*, vol. 15, no. 1, p. 11824, Apr. 2025, doi: 10.1038/s41598-025-95734-z.