Klasifikasi Emosi Mahasiswa Menggunakan LSTM Berbasis MFCC dengan Pendekatan Multimodal Fusion

Muhamad Fahri Muttaqin
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
muhamadfahrimuttaqin@student.telkomuniversity.ac.id

Kris Sujatmoko. S.T.,M.T.
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
Krissujatmoko@telkomuniversity.ac.id

Dr. Ir. Bambang Setia Nugroho.S.T.,M.T.
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
bambangsetianugroho@telkomuniversity.ac.id

***Abstrak* — Dosen sering mengalami kesulitan dalam mengenali kondisi emosional mahasiswa secara objektif selama proses pembelajaran berlangsung, karena isyarat emosional bersifat subjektif dan sulit diamati secara langsung. Penelitian ini mengusulkan modul klasifikasi emosi mahasiswa berbasis suara menggunakan arsitektur Long Short-Term Memory (LSTM) dengan fitur Mel-Frequency Cepstral Coefficients (MFCC), yang dipadukan dengan pendekatan Multimodal Late Fusion untuk menggabungkan hasil klasifikasi akustik dengan analisis sentimen berbasis leksikon dari teks hasil transkripsi. Emosi dipetakan ke dalam dua kelas biner, yaitu Positif (senang dan netral) dan Negatif (sedih dan marah). Model dilatih menggunakan dataset suara akademik yang dikumpulkan secara mandiri dari 4 orang aktor mahasiswa. Hasil pengujian menunjukkan bahwa meskipun arsitektur CNN memperoleh akurasi data latih lebih tinggi (85,57%) dibandingkan LSTM (81,36%), performa CNN menurun drastis menjadi 48,3% pada pengujian lapangan (live testing), sedangkan LSTM tetap tangguh dengan akurasi 78,3%. Evaluasi Leave-One-Subject-Out Cross Validation (LOSO-CV) menghasilkan rata-rata akurasi generalisasi sebesar 81,4%. Pada pengujian sistem Multimodal Late Fusion terhadap 60 sampel (4 subjek × 15 skenario), sistem berhasil mengklasifikasikan 47 sampel dengan benar (akurasi 78,3%), dengan performa terbaik pada skenario anomali linguistik seperti sarkasme dan candaan/ironi (87,5%), yang membuktikan bahwa pendekatan multimodal mampu memvalidasi silang antara nada suara dan makna teks sehingga menghasilkan klasifikasi emosi yang lebih tangguh dibandingkan pendekatan unimodal.**

Kata kunci — LSTM, MFCC, Speech Emotion Recognition, Multimodal Fusion, Late Fusion.

I. PENDAHULUAN

Keterlibatan mahasiswa dalam proses pembelajaran merupakan faktor krusial yang menentukan efektivitas pendidikan tinggi, dan dipengaruhi oleh berbagai faktor kompleks termasuk metode pengajaran, motivasi mahasiswa, serta kondisi emosional seperti kebosanan, frustrasi, dan tekanan psikologis selama pembelajaran berlangsung [1]. Namun, dosen sering kali mengalami kesulitan dalam mengenali kondisi emosional mahasiswa selama proses pembelajaran, seperti apakah mahasiswa merasa senang, sedih, atau marah, karena isyarat emosional tersebut sering kali sulit diamati secara langsung dan cenderung bersifat subjektif. Padahal, emosi memiliki peran penting dalam proses belajar karena berfungsi sebagai penghubung antara motivasi dan kinerja kognitif, di mana emosi positif dapat meningkatkan efektivitas proses belajar sedangkan emosi negatif dapat menghambat keinginan seseorang untuk belajar [2].

Beberapa pendekatan Speech Emotion Recognition (SER) telah dikembangkan untuk mengatasi permasalahan tersebut. Pendekatan pertama adalah sistem klasik yang mengandalkan ekstraksi fitur Mel-Frequency Cepstral Coefficients (MFCC) yang diklasifikasikan menggunakan classifier konvensional seperti Support Vector Machine (SVM) atau Hidden Markov Model (HMM). Pendekatan berbasis fitur akustik klasik semacam ini umumnya bergantung pada kualitas ekstraksi fitur manual sehingga kurang mampu menangkap dependensi temporal pada sinyal suara [4]. Pendekatan kedua adalah arsitektur Deep Learning modern seperti Convolutional Neural Network (CNN) yang berfokus pada analisis sinyal akustik secara unimodal, tanpa mempertimbangkan makna kata yang diucapkan, sehingga rentan tertipu oleh anomali linguistik seperti sarkasme atau ironi ketika mahasiswa menyampaikan keluhan dengan nada yang terdengar tegas atau bersemangat.

Penelitian ini mengusulkan modul klasifikasi emosi mahasiswa berbasis suara yang mengganti classifier konvensional dengan model Long Short-Term Memory (LSTM) berbasis fitur MFCC, sehingga model lebih mampu menangkap dependensi temporal pada sinyal suara dan lebih tangguh terhadap variasi data antar individu. Untuk mengatasi kelemahan pendekatan unimodal, diterapkan pula pendekatan multimodal yang menggabungkan hasil klasifikasi emosi dari sinyal akustik (LSTM + MFCC) dengan analisis sentimen berbasis teks hasil transkripsi menggunakan pendekatan leksikon, melalui mekanisme Multimodal Late Fusion. Dengan mekanisme fusion ini, sistem dapat memvalidasi silang antara nada suara dan makna

kata yang sebenarnya diucapkan sebelum mengambil kesimpulan akhir mengenai kondisi emosional mahasiswa, sehingga akurasi deteksi menjadi lebih tangguh dibandingkan solusi yang hanya mengandalkan satu modalitas. Seluruh pipeline pemrosesan dijalankan secara lokal (offline) pada perangkat tanpa bergantung pada layanan Cloud API pihak ketiga, sejalan dengan ketentuan Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi [3].

Penelitian ini secara khusus bertujuan untuk mengevaluasi performa model LSTM berbasis MFCC dalam mengklasifikasikan emosi mahasiswa ke dalam dua kategori polaritas biner (Positif dan Negatif), baik pada tahap pelatihan model, tahap generalisasi menggunakan Leave-One-Subject-Out Cross Validation (LOSO-CV), maupun pada tahap pengujian lapangan (live testing) dengan pendekatan Multimodal Late Fusion, sebagai bagian dari evaluasi modul klasifikasi emosi pada sistem deteksi suara mahasiswa yang lebih luas.

II. KAJIAN TEORI

A. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) merupakan salah satu metode ekstraksi fitur yang paling banyak digunakan dalam pengolahan suara, seperti pengenalan ucapan (speech recognition), identifikasi pembicara (speaker recognition), dan pengenalan emosi suara (speech emotion recognition) [5]. MFCC dirancang untuk merepresentasikan karakteristik suara dengan pendekatan yang lebih mendekati cara manusia mendengar melalui penggunaan skala Mel [7]. Secara umum, proses ekstraksi MFCC melibatkan pembagian sinyal suara menjadi beberapa frame, transformasi sinyal ke domain frekuensi menggunakan Fast Fourier Transform (FFT), pemrosesan menggunakan Mel Filter Bank, dan perhitungan menggunakan Discrete Cosine Transform (DCT) [7], sehingga menghasilkan sekumpulan koefisien yang merepresentasikan karakteristik spektral dari sistem produksi suara manusia, termasuk kontribusi pita suara, lidah, dan bibir dalam membentuk sinyal akustik.

B. *Speech Emotion Recognition* (SER)

Emosi memiliki peran penting dalam komunikasi manusia dan dapat disampaikan melalui berbagai karakteristik suara, seperti intonasi, nada, kecepatan berbicara, serta ritme ucapan. Klasifikasi emosi suara atau Speech Emotion Recognition (SER) merupakan teknologi yang digunakan untuk mengenali kondisi emosional seseorang berdasarkan karakteristik suara yang dihasilkan. Sistem SER umumnya bekerja melalui beberapa tahapan, dimulai dari preprocessing untuk meningkatkan kualitas sinyal, dilanjutkan dengan ekstraksi fitur seperti MFCC yang mampu merepresentasikan informasi emosional dari sinyal suara, kemudian diproses menggunakan metode machine learning atau deep learning untuk mengklasifikasikan emosi ke dalam kategori tertentu [10]. Pada perkembangan terbaru, pendekatan deep learning seperti Convolutional Neural Network (CNN) banyak digunakan dalam sistem SER karena kemampuannya mengekstraksi pola kompleks secara otomatis dari representasi sinyal suara seperti Mel-spectrogram, sehingga memungkinkan sistem mempelajari hubungan non-linear antara fitur suara dan kondisi emosional secara lebih efektif dibandingkan metode konvensional [10]. Penerapan SER pada lingkungan e-learning dapat membantu meningkatkan pemahaman terhadap kondisi emosional pengguna guna mendukung evaluasi keterlibatan dalam pembelajaran [6].

C. *Long Short-Term Memory* (LSTM)

LSTM (*Long Short-Term Memory*) merupakan salah satu jenis jaringan saraf tiruan yang dirancang untuk menangkap pola sekuensial dalam data, sehingga ideal digunakan untuk menganalisis dinamika temporal yang terdapat pada sinyal suara [11]. Kemampuan LSTM menangkap pola sekuensial jangka panjang diperoleh melalui struktur sel (cell state) yang diatur oleh tiga gerbang, yaitu forget gate yang menentukan informasi lama mana yang perlu dibuang, input gate yang menentukan informasi baru mana yang perlu disimpan, dan output gate yang menentukan bagian mana dari cell state yang dikeluarkan sebagai hidden state pada tiap langkah waktu [11]. Mekanisme ini memungkinkan LSTM mempelajari pola perubahan intonasi, energi, dan pitch suara secara bertahap sepanjang durasi ucapan, tanpa mengalami masalah vanishing gradient yang umum terjadi pada jaringan saraf berulang (RNN) konvensional. Sebagai perbandingan, Convolutional Neural Network (CNN) menangkap pola spektral yang bersifat lebih statis dibandingkan pola temporal, sehingga performanya cenderung lebih rentan menurun pada kondisi suara spontan dibandingkan pendekatan berbasis LSTM.

D. *Multimodal Fusion (Late Fusion)*

Late fusion atau disebut juga *decision-level fusion* merupakan salah satu pendekatan *multimodal fusion* yang banyak digunakan dalam pengenalan emosi multimodal. Pendekatan ini bekerja dengan cara mengekstraksi fitur dan melatih model secara independen untuk setiap modalitas, kemudian menggabungkan hasil prediksi dari masing-masing modalitas tersebut melalui proses seperti rata-rata, penjumlahan berbobot, voting mayoritas, atau jaringan saraf tiruan untuk memperoleh kategori emosi akhir [13]. Dibandingkan dengan pendekatan early fusion yang menggabungkan fitur mentah dari berbagai modalitas sejak awal sebelum masuk ke model, late fusion memberikan keleluasaan bagi setiap modalitas untuk diproses menggunakan arsitektur model yang paling sesuai dengan karakteristik datanya masing-masing, sekaligus mempermudah proses evaluasi karena kontribusi tiap modalitas terhadap keputusan akhir dapat ditelusuri secara terpisah. Karakteristik ini menjadikan late fusion sesuai digunakan pada kasus di mana sinyal dari satu modalitas dapat saling melengkapi atau bahkan bertentangan dengan modalitas lainnya, seperti pada kondisi sarkasme di mana nada suara dan pilihan kata menunjukkan arah emosi yang berbeda.

III. METODE

Penelitian ini merupakan bagian dari evaluasi modul klasifikasi emosi pada sistem Student Voice Detection, yaitu sistem deteksi suara mahasiswa berbasis pengolahan audio dan machine learning yang dijalankan secara lokal untuk mendukung analisis keterlibatan dan evaluasi pembelajaran di kelas.

A. Dataset Emosi

Pada tahap awal, penelitian ini mempertimbangkan penggunaan dataset publik yang tersedia. Namun, setelah dilakukan eksplorasi terhadap beberapa dataset emosi wicara yang umum digunakan, ditemukan bahwa dataset tersebut memiliki keterbatasan seperti perbedaan bahasa, aksen, serta konteks budaya yang kurang merepresentasikan karakteristik tutur mahasiswa Indonesia. Dataset emosi wicara berbahasa Indonesia yang tersedia pun dinilai kurang sesuai karena durasi rekaman yang relatif singkat serta kurang mampu merepresentasikan variasi ekspresi emosi dalam konteks kalimat akademik, yang berpotensi mengurangi kemampuan model dalam mempelajari pola temporal emosi secara optimal, khususnya untuk arsitektur berbasis sekuens seperti LSTM.

Berdasarkan pertimbangan tersebut, penelitian ini membangun dataset emosi wicara secara mandiri (self-collected dataset) dengan melibatkan 4 orang aktor mahasiswa. Setiap aktor diminta mengucapkan 10 kalimat akademik yang sama secara berulang dengan variasi ekspresi emosi yang berbeda, guna menjaga konsistensi konten sehingga perbedaan yang ditangkap model lebih berfokus pada aspek emosional dari suara. Distribusi dataset yang dihasilkan disajikan pada Tabel 1.

TABEL 1
DISTRIBUSI DATASET EMOSI

Aktor	Netral	Senang	Sedih	Marah
Aktor 01	50	25	25	25
Aktor 02	50	25	25	25
Aktor 03	50	25	25	25
Aktor 04	50	25	25	25
Total	200	100	100	100

Distribusi data menunjukkan bahwa jumlah sampel emosi netral dibuat lebih banyak dibandingkan emosi lainnya untuk merepresentasikan kondisi komunikasi sehari-hari yang cenderung didominasi oleh ekspresi netral, sekaligus berfungsi sebagai baseline bagi model dalam membedakan emosi positif dan negatif. Skenario emosi kemudian dipetakan ke dalam dua kelas biner untuk mempermudah proses klasifikasi dan membentuk batas keputusan (decision boundary) yang lebih jelas, yaitu Kelas Positif yang mencakup emosi senang dan netral, serta Kelas Negatif yang mencakup emosi sedih dan marah. Proses perekaman dilakukan menggunakan mikrofon dengan kondisi lingkungan yang dijaga relatif konsisten untuk meminimalkan noise eksternal, di mana setiap aktor menyesuaikan intonasi, tekanan suara, dan ekspresi sesuai emosi yang direpresentasikan.

B. Ekstraksi Fitur MFCC

Sinyal audio yang telah melalui tahap Voice Activity Detection (VAD) untuk memangkas keheningan (silence) dengan ambang batas 20 desibel, kemudian ditransformasikan menjadi representasi numerik menggunakan metode ekstraksi fitur MFCC. Proses ini menghasilkan 40 koefisien MFCC per frame yang merepresentasikan karakteristik spektral dari sistem produksi suara manusia. Matriks fitur kemudian dipotong (chunking) menjadi jendela waktu beresolusi 130 frame, sehingga tersusun matriks temporal berukuran 130 frame × 40 koefisien sebagai input ke model LSTM. Sebelum diumpankan ke dalam arsitektur jaringan saraf tiruan, matriks fitur tersebut dinormalisasi menggunakan Z-Score Normalization untuk memusatkan distribusi data agar berpusat pada nilai rata-rata nol dengan standar deviasi satu.

C. Arsitektur dan Pelatihan Model LSTM

Modul klasifikasi emosi diimplementasikan menggunakan arsitektur Long Short-Term Memory (LSTM), yang merupakan pengembangan dari Recurrent Neural Network (RNN) dan dirancang untuk mengatasi permasalahan vanishing gradient pada pemrosesan data sekuensial. Arsitektur model terdiri dari lapisan Masking untuk mengabaikan bagian data hasil zero-padding, lapisan LSTM dengan 128 unit untuk mempelajari pola urutan waktu dari sinyal suara, dua lapisan Dropout sebesar 0,4 untuk mematikan 40% neuron secara acak pada setiap epoch guna mencegah overfitting, serta lapisan Dense berdimensi 64 dengan fungsi aktivasi ReLU sebelum menghasilkan keputusan klasifikasi biner akhir. Sebelum tahap pelatihan model final ditetapkan, dilakukan evaluasi komparatif antara dua kandidat arsitektur akustik, yaitu Convolutional Neural Network (CNN) dan LSTM, untuk menentukan model yang paling representatif dalam mengenali pola emosi dari fitur MFCC.

D. Multimodal Late Fusion

Penelitian ini menyadari bahwa suara dengan intonasi tinggi atau menggebu-gebu tidak selamanya bermakna positif; dalam beberapa kondisi, mahasiswa dapat menyampaikan keluhan, kebingungan, atau bahkan sarkasme dengan nada yang terdengar tegas atau bersemangat. Oleh karena itu, diimplementasikan mekanisme fusi akhir (late fusion) berbasis multimodal dengan menyertakan analisis leksikon berbasis teks untuk menganalisis makna semantik dari kata yang diucapkan [12]. Hasil klasifikasi emosi akustik dari model LSTM digabungkan dengan hasil analisis sentimen teks transkripsi melalui aturan keputusan fusi yang disajikan pada Tabel 2.

TABEL 2
ATURAN KEPUTUSAN FUSI AKHIR (LATE FUSION) MULTIMODAL

Teks	Audio	Output
Positif	Positif	Positif
Negatif	Negatif	Negatif
Positif	Negatif	Negatif
Negatif	Positif	Positif
Netral	Negatif	Sesuai Nada

Melalui matriks keputusan pada Tabel 2, sistem tidak hanya mempertimbangkan bagaimana suatu kalimat diucapkan, tetapi juga apa makna semantik dari kalimat tersebut secara berimbang, sehingga dapat mereduksi tingkat kesalahan klasifikasi (false positive) yang kerap terjadi jika hanya mengandalkan karakteristik akustik seperti tinggi-rendahnya nada suara. Pada sisi akustik, sistem mengekstraksi nilai probabilitas dari setiap potongan (chunk) audio, kemudian menghitung rata-rata keseluruhan probabilitas tersebut menggunakan metode Soft Voting. Jika rata-rata probabilitas tersebut mencapai atau melebihi ambang batas (threshold) sebesar 60%, emosi akustik dari modul LSTM ditetapkan sebagai label akustik akhir; label ini kemudian digabungkan dengan hasil analisis emosi teks

melalui aturan Late Fusion pada Tabel 2 untuk menghasilkan emosi final.

E. Skema Pengujian

Pengujian dilakukan dalam dua tahap utama. Tahap pertama adalah analisis hasil pelatihan model yang mencakup perbandingan performa CNN dan LSTM pada data latih, serta pengujian kemampuan generalisasi model terpilih menggunakan metode Leave-One-Subject-Out Cross Validation (LOSO-CV) untuk mensimulasikan skenario dunia nyata di mana sistem diuji kemampuannya beradaptasi terhadap variasi karakteristik vokal individu yang berbeda. Tahap kedua adalah pengujian kinerja sistem secara langsung (live testing) untuk memvalidasi performa sistem Multimodal Late Fusion pada kondisi nyata. Pengujian lapangan dilakukan oleh 4 orang subjek mahasiswa yang berbeda, di mana setiap subjek diinstruksikan membacakan 15 skenario kalimat uji yang mencakup validasi konsisten (positif/negatif), teks netral dengan nada marah, sarkasme (teks positif nada negatif), dan candaan/ironi (teks negatif nada positif), sehingga diperoleh total 60 sampel pengujian (15 skenario × 4 subjek).

Metrik evaluasi yang digunakan meliputi Accuracy, Precision, Recall, dan F1-Score, dengan persamaan sebagai berikut:

$$Accuracy = \frac{TP + TN}{N} \times 100\% \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

dengan TP (True Positive), TN (True Negative), FP (False Positive), dan FN (False Negative) diperoleh dari distribusi hasil prediksi sistem terhadap label aktual pada confusion matrix.

IV. HASIL DAN PEMBAHASAN

A. Perbandingan Arsitektur CNN dan LSTM

Pada tahap perancangan awal modul klasifikasi emosi, dilakukan evaluasi komparatif antara dua arsitektur jaringan saraf, yaitu CNN dan LSTM, untuk menentukan model akustik yang paling representatif dalam mengenali pola emosi dari fitur MFCC. Kedua arsitektur dilatih menggunakan dataset suara yang sama agar perbandingan performa dapat dilakukan secara setara. Hasil perbandingan disajikan pada Tabel 3.

TABEL 3
PERBANDINGAN AKURASI PELATIHAN DAN PENGUJIAN LAPANGAN

Arsitektur	Akurasi Data Latih	Akurasi Live Testing
CNN	85,57%	48,3%
LSTM	81,36%	78,3%

Pada hasil pelatihan, CNN memperoleh akurasi validasi sebesar 85,57%, lebih tinggi dibandingkan LSTM yang memperoleh akurasi 81,36%. Berdasarkan metrik ini saja, CNN tampak sebagai kandidat yang lebih unggul. Namun, ketika kedua model diuji pada kondisi pengujian lapangan (live testing) dengan suara spontan dalam lingkungan tidak terkontrol, performa keduanya berbalik secara signifikan: akurasi CNN turun drastis dari 85,57% menjadi 48,3%, sedangkan LSTM hanya mengalami penurunan dari 81,36% menjadi 78,3%. Hal ini menunjukkan bahwa CNN, yang menangkap pola spektral bersifat lebih statis, mengalami overfitting terhadap karakteristik akustik spesifik pembicara pada dataset pelatihan, sedangkan LSTM yang menangkap dinamika temporal terbukti jauh lebih tangguh terhadap pergeseran domain (domain shift) dan gangguan lingkungan (noise) pada kondisi nyata. Berdasarkan hasil ini, arsitektur LSTM dipilih sebagai model akustik final pada sistem.

B. Confusion Matrix dan Classification Report Pelatihan LSTM

Distribusi hasil prediksi model LSTM pada data pelatihan dievaluasi melalui confusion matrix, dengan parameter True Positive (TP) = 266, True Negative (TN) = 140, False Positive (FP) = 59, dan False Negative (FN) = 34, dari total sampel (N) sebesar 499 data wicara. Berdasarkan parameter tersebut, diperoleh nilai akurasi keseluruhan sebesar 81,36%, yang membuktikan bahwa model LSTM biner telah berhasil mempelajari pola umum ekstraksi sinyal wicara manusia dengan sangat baik. Melalui capaian persentase di atas 80% ini, model dikategorikan sangat layak dan memiliki fondasi klasifikasi akustik yang kokoh sebelum dipadukan dengan leksikon teks pada mekanisme Multimodal Late Fusion. Laporan klasifikasi (classification report) per kelas disajikan pada Tabel 4.

TABEL 4
LAPORAN KLASIFIKASI MODEL LSTM

Kelas Emosi	Precision	Recall	F1 Score	Support
Negatif	80,46%	70,35%	75,07%	199
Positif	81,85%	88,67%	85,10%	300
Total			81,36%	499

Hasil evaluasi kelas Positif menunjukkan tingkat presisi sebesar 81,85% dan recall sebesar 88,67%, membuktikan bahwa arsitektur LSTM biner memiliki sensitivitas tinggi dalam mengeksplorasi fitur MFCC bermakna antusias, ditandai dengan keberhasilan identifikasi sebanyak 266 sampel True Positive dan minimnya angka False Negative (34 sampel), sehingga menghasilkan nilai F1-score yang kuat di angka 85,10%. Meskipun dataset sebaran awal bersifat tidak seimbang (imbalanced support), nilai F1-score kelas Negatif tetap bertahan kokoh pada angka 75,07% dengan presisi 80,46% dan recall 70,35%, yang memvalidasi keandalan model dalam mengklasifikasikan 140 sampel True Negative secara presisi tanpa mengalami pergeseran garis keputusan (decision boundary bias) ke arah kelas mayoritas (Positif).

C. Generalisasi Model (LOSO-CV)

Untuk memvalidasi kemampuan model dalam mengenali emosi pembicara baru, diterapkan metode Leave-One-Subject-Out Cross-Validation (LOSO-CV), yang mensimulasikan skenario dunia nyata di mana sistem diuji kemampuannya beradaptasi terhadap variasi karakteristik vokal individu yang berbeda. Rekapitulasi performa generalisasi disajikan pada Tabel 5.

TABEL 5
REKAPITULASI PERFORMA GENERALISASI (LOSO-CV)

Subjek Uji	Tingkat Pengenalan (%)
Fahri	94,4%
Nayla	78,2%
Dika	79,2%
Ara	73,6%
Rata-Rata Akurasi	81,4%

Hasil rata-rata akurasi sebesar 81,4% mengonfirmasi bahwa model memiliki daya generalisasi yang baik dan stabil terhadap subjek pembicara yang tidak dilibatkan dalam proses pelatihan, sehingga model layak diterapkan pada kondisi operasional dengan variasi mahasiswa yang berbeda-beda.

D. Pengujian Multimodal Late Fusion (Live Testing)

Pengujian klasifikasi emosi Multimodal Late Fusion dilakukan untuk memastikan bahwa gabungan model LSTM (Audio) dan analisis leksikon (Teks) dapat mendeteksi emosi mahasiswa secara akurat, termasuk dalam mengatasi situasi komunikasi kompleks seperti sarkasme dan candaan/ironi. Pengujian melibatkan 4 subjek mahasiswa yang masing-masing membacakan 15 skenario kalimat uji, sehingga diperoleh total 60 sampel pengujian. Hasil analisis performa per penguji disajikan pada Tabel 6, sedangkan hasil per skenario pengujian disajikan pada Tabel 7.

TABEL 6
ANALISA PERFORMA PER PENGUJI

Subjek Penguji	Total Uji	Sesuai	Tidak Sesuai	Akurasi
Subjek 1	15	12	3	80%
Subjek 2	15	11	4	73%
Subjek 3	15	12	3	80%
Subjek 4	15	12	3	80%

Data pada Tabel 6 menunjukkan bahwa arsitektur Multimodal Late Fusion terbukti memiliki ketangguhan yang cukup stabil terhadap perbedaan karakteristik suara vokal pengguna. Tiga dari empat subjek (Subjek 1, 3, dan 4) berhasil mendapatkan tingkat akurasi sebesar 80%, sementara Subjek 2 berada di angka 73%. Konsistensi nilai akurasi di atas 70% ini mengindikasikan bahwa sistem mampu melakukan generalisasi dengan baik, dengan penurunan pada Subjek 2 dievaluasi sebagai akibat dari kebiasaan berbicara (speaking style) subjek yang cenderung bernada datar atau kurang mengekspresikan energi emosi pada pita suaranya, sehingga membingungkan penarikan kesimpulan pada model LSTM.

TABEL 7
ANALISA PERFORMA PER SKENARIO

Skenario Pengujian	Total Data Uji	Sesuai	Tidak Sesuai	Akurasi
Konsisten Positif	16	13	3	81,3%
Konsisten Negatif	16	11	5	68,8%
Teks Netral, Nada Marah	12	9	3	75%
Teks Positif, Nada Negatif	8	7	1	87,5%
Teks Negatif, Nada Positif	8	7	1	87,5%
Total	60	47	13	78,3%

Melalui Tabel 7, dapat dilihat bahwa secara keseluruhan sistem berhasil menebak 47 dari 60 sampel dengan benar, menghasilkan akurasi total sebesar 78,3%. Menariknya, keunggulan utama sistem ini justru terlihat pada skenario anomali linguistik yang kompleks, yaitu skenario Sarkasme (Teks Positif, Nada Negatif) dan Candaan/Ironi (Teks Negatif, Nada Positif), dengan tingkat akurasi masing-masing menyentuh 87,5%. Hal ini membuktikan bahwa metode Late Fusion berhasil menjalankan fungsinya: ketika leksikon teks secara keliru menangkap makna karena adanya sarkasme, fitur akustik (audio) berhasil memberikan bobot kompensasi yang lebih dominan untuk mengarahkan prediksi sistem pada emosi final yang tepat. Di sisi lain, akurasi terendah (68,8%) terjadi pada skenario Konsisten Negatif, yang umumnya disebabkan oleh artikulasi yang terburu-buru sehingga fitur MFCC kehilangan batas frekuensi emosi.

E. Confusion Matrix Evaluasi Multimodal

Untuk menganalisis distribusi penyebaran prediksi model secara lebih menyeluruh, disusun confusion matrix total dari seluruh pengujian lapangan (60 sampel), sebagaimana disajikan pada Tabel 8.

TABEL 8
CONFUSION MATRIX TOTAL EVALUASI MULTIMODAL (60 PENGUJIAN)

	Prediksi: Positif	Prediksi: Negatif
Aktual: Positif	20 (TP)	4 (FN)
Aktual: Negatif	9 (FP)	27 (TN)

Dari total 24 sampel audio yang secara aktual bernilai emosi Positif (gabungan skenario konsisten positif dan candaan/ironi), sistem berhasil memprediksi dengan benar sebagai Positif (True Positive) sebanyak 20 sampel, dengan 4 sampel mengalami salah klasifikasi menjadi emosi Negatif (False Negative). Sementara itu, dari total 36 sampel audio yang secara aktual bernilai emosi Negatif (gabungan skenario konsisten negatif, teks netral nada marah, dan sarkasme), sistem sukses mengidentifikasi sebanyak 27 sampel sebagai kelas Negatif (True Negative), dengan 9 sampel salah terdeteksi sebagai emosi Positif (False Positive). Tingginya angka kemunculan True Positive (20) dan True Negative (27)

secara bersamaan mencerminkan nilai akurasi proporsional (balanced accuracy) yang solid, membuktikan bahwa garis keputusan (decision boundary) yang dibentuk oleh fusi bobot antara model LSTM dan leksikon teks telah optimal, tanpa mengalami gejala overfitting atau kecondongan hanya pada satu kelas emosi saja.

F. Pembahasan

Secara keseluruhan, hasil pengujian menunjukkan bahwa model LSTM berbasis fitur MFCC jauh lebih tangguh terhadap gangguan lingkungan (noise) dan pergeseran domain (domain shift) dibandingkan arsitektur CNN, sehingga lebih layak digunakan pada kondisi operasional nyata meskipun akurasi data latihnya sedikit lebih rendah. Penerapan pendekatan Multimodal Late Fusion, yang menggabungkan model LSTM (akustik) dengan analisis leksikon berbasis teks, terbukti mampu meningkatkan ketangguhan sistem dalam menghadapi situasi komunikasi yang kompleks, khususnya pada skenario anomali linguistik seperti sarkasme dan candaan/ironi yang menjadi kelemahan utama pendekatan unimodal. Dengan demikian, kombinasi arsitektur LSTM berbasis MFCC dan mekanisme Multimodal Late Fusion memberikan solusi klasifikasi emosi mahasiswa yang lebih akurat, tangguh, dan kontekstual dibandingkan pendekatan konvensional maupun unimodal deep learning.

V. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan, model LSTM berbasis fitur MFCC terbukti lebih tangguh dibandingkan CNN dalam mengklasifikasikan emosi mahasiswa pada kondisi lapangan, dengan akurasi data latih sebesar 81,36% yang hanya menurun menjadi 78,3% pada pengujian lapangan, dibandingkan CNN yang menurun drastis dari 85,57% menjadi 48,3%. Evaluasi Leave-One-Subject-Out Cross Validation (LOSO-CV) menghasilkan rata-rata akurasi generalisasi sebesar 81,4%, yang mengonfirmasi kemampuan model dalam beradaptasi terhadap variasi karakteristik vokal individu yang berbeda. Pada pengujian sistem Multimodal Late Fusion terhadap 60 sampel lapangan, sistem berhasil mengklasifikasikan emosi mahasiswa dengan akurasi total 78,3%, dengan performa terbaik pada skenario anomali linguistik seperti sarkasme dan candaan/ironi (87,5%), yang membuktikan bahwa pendekatan multimodal mampu memvalidasi silang antara nada suara dan makna teks sehingga menghasilkan klasifikasi emosi yang lebih tangguh dibandingkan pendekatan unimodal.

Dengan demikian, arsitektur LSTM berbasis MFCC yang dipadukan dengan Multimodal Late Fusion terbukti menjadi solusi klasifikasi emosi mahasiswa yang layak digunakan untuk mendukung sistem evaluasi keterlibatan mahasiswa berbasis suara pada lingkungan kelas nyata. Penelitian selanjutnya disarankan untuk mengeksplorasi penerapan Bandpass Filter (BPF) pada tahap prapemrosesan audio guna meningkatkan kualitas sinyal, serta memperluas variasi dataset emosi wicara agar model dapat lebih mengoptimalkan pembedaan batas frekuensi emosi khususnya pada skenario dengan artikulasi yang cepat.

REFERENSI

[1] R. Das and S. Dev, “Optimizing student engagement detection using facial and behavioral features,” Neural Comput. Appl., vol. 37, no. 23, pp. 19063–19085, 2025, doi: 10.1007/s00521-025-11317-z.

[2] U. Negeri Gorontalo, J. Jend Sudirman No, D. Tim, K. Kota Tengah, and K. Gorontalo, “Peran Emosi Dalam Proses Belajar Isniarty Gintulangi,” J. Educ. Emosi Dalam Proses Belajar Isniarty Gintulangi, vol. 07, no. 01, pp. 7871–7878, 2024.

[3] Pemerintah Pusat, “Undang-undang (UU) Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi,” Pemerintah Pus., no. 016999, p. 1, 2022, [Online]. Available: https://peraturan.bpk.go.id/Download/224884/UU Nomor 27 Tahun 2022.pdf

[4] A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, and T. K. Whangbo, “Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm,” Sensors, vol. 22, no. 21, 2022, doi: 10.3390/s22218122.

[5] S. Ali, S. Tanweer, S. Khalid, and N. Rao, “Mel Frequency Cepstral Coefficient: A Review,” 2021, doi: 10.4108/eai.27-2-2020.2303173.

[6] A. Boughida, M. N. Kouahla, and I. Chebata, “Speech emotion recognition using MFCC features: application in e-learning environements,” Int. J. Speech Process., pp. 0–3, 2023,[Online].Available:https://www.researchgate.net/publication/338987485

[7] R. Sharma, D. Govind, J. Mishra, A. K. Dubey, K. T. Deepak, and S. R. M. Prasanna, Milestones in speaker recognition, vol. 57, no. 3. Springer Netherlands, 2024. doi: 10.1007/s10462-023-10688-w.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Proc. Mach. Learn. Res., vol. 202, pp. 28492–28518, 2023.

[9] M. M. Or-Rashid, A. K. Nondi, A. Al Sadnun, M. A. H. Wadud, T. M. A. U. H. Bhuiyan, and M. S. Hossain, “Speech Emotion Recognition from Audio Data Using LSTM Model,” Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 7, pp. 247–254, 2025, doi: 10.14569/IJACSA.2025.0160726.

[10] N. Penumajji, “Deep Learning for Speech Emotion Recognition: A CNN Approach Utilizing Mel Spectrograms,” ICEIEC 2025 - Proc. 2025 IEEE 15th Int. Conf. Electron. Inf. Emerg. Commun., pp. 74–78, 2025, doi: 10.1109/ICEIEC65904.2025.11273118.

[11] R. Begazo, A. Aguilera, I. Dongo, and Y. Cardinale, “A Combined CNN Architecture for Speech Emotion Recognition,” Sensors, vol. 24, no. 17, pp. 1–39, 2024, doi: 10.3390/s24175797.

[12] R. Firdaus, I. A. #2, and A. Herdiani, “Lexicon-Based Sentiment Analysis of Indonesian Language Student Feedback Evaluation,” J. Comput., vol. 6, no. 1, pp. 1–12, 2021, doi: 10.34818/indojc.2021.6.1.408.

[13] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, “Recognition: Speech, Text, and Face,” MDPI, J. Basel, Switzerland., pp. 1–33, 2023.