JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.10900

IMPLEMENTASI SPEECH EMOTION RECOGNITION UNTUK KLASIFIKASI TINGKAT STRES DARI VOICE NOTE BERBASIS HYBRID CNN-LSTM

Muhamad Fakhri Khairil Imam[1*], Asril Adi Sunarto[2], Winda Apriandari[3]

[1,2,3]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sukabumi; Jl. R. Syamsudin, S.H. No. 50, Sukabumi; 0266-218345

Keywords:
CNN-LSTM; Speech Emotion Recognition; Klasifikasi Tingkat Stres; Hybrid Fusion; Voice

Corespondent Email:
ririfakhri888@ummi.ac.id

Abstrak. Kesehatan mental menjadi isu penting seiring meningkatnya tekanan psikologis akibat gaya hidup digital, namun deteksi stres masih mengandalkan instrumen *self-report* yang bersifat terjadwal dan subjektif. Penelitian ini mengembangkan sistem *Speech Emotion Recognition* (SER) berbasis *voice note* berbahasa Indonesia untuk mengklasifikasikan tingkat stres menggunakan pendekatan hybrid CNN-LSTM (akustik) dan *Zero-Shot Classification* IndoBERT (linguistik) dengan metodologi CRISP-DM. Model dilatih menggunakan dataset gabungan RAVDESS (1.440 file, pembagian *speaker-independent*) dan *voice note* primer berbahasa Indonesia (106 file) yang dipetakan ke tiga kategori *Non_Stress, Mild_Stress,* dan *High_Stress*, menghasilkan 1.546 sampel asli. Fitur akustik diekstraksi menggunakan 40 koefisien MFCC dengan panjang tetap 174 *frame.* Pengujian pada 256 sampel independen menghasilkan akurasi model CNN-LSTM akustik sebesar 62,11% dengan *F1-score* terbaik pada kelas *High_Stress* (0,69), sementara kelas *Mild_Stress* menunjukkan performa terendah (F1 0,41) akibat karakteristik akustik yang ambigu. Untuk mengatasi kesenjangan bahasa antara dataset RAVDESS dan *voice note* Indonesia, sistem menerapkan fusi hibrida dengan bobot linguistik 90% dan akustik 10%. Model diintegrasikan ke aplikasi *web StressVoice* berbasis Flask yang dilengkapi transkripsi Whisper AI, respons empatik, cetak surat rujukan PDF, dan integrasi konsultasi WhatsApp ke psikolog.

Abstract. *Mental health has become a critical issue amid rising psychological pressure from digital lifestyles, yet stress detection still relies on scheduled, subjective self-report instruments. This study develops a Speech Emotion Recognition (SER) system based on Indonesian voice notes to classify stress levels using a hybrid CNN-LSTM (acoustic) and IndoBERT Zero-Shot Classification (linguistic) approach under the CRISP-DM methodology. The model was trained on a combined RAVDESS dataset (1,440 files, speaker-independent split) and primary Indonesian voice notes (106 files) mapped into three categories, Non_Stress, Mild_Stress, and High_Stress, yielding 1,546 original samples. Acoustic features were extracted using 40 MFCC coefficients with a fixed length of 174 frames. Testing on 256 independent samples produced an acoustic CNN-LSTM accuracy of 62.11%, with the best F1-score on the High_Stress class (0.69), while Mild_Stress showed the lowest performance (F1 0.41) due to ambiguous acoustic characteristics. To bridge the language gap between RAVDESS and Indonesian voice notes, the system applies hybrid fusion weighting linguistic analysis at 90% and acoustic analysis at 10%. The model is integrated into the Flask-based StressVoice web application, equipped with Whisper AI transcription, empathetic responses.*

1. PENDAHULUAN

Kesehatan mental menjadi isu krusial di era modern seiring meningkatnya tekanan psikologis akibat gaya hidup digital, tuntutan pekerjaan, dan dinamika sosial. Sinyal suara manusia mengandung informasi non-verbal yang mencerminkan kondisi emosional melalui perubahan pitch, intonasi, energi, dan pola akustik tertentu [1]. Stres sebagai salah satu kondisi psikologis dapat memengaruhi konsentrasi, kualitas tidur, dan produktivitas, serta tercermin melalui perubahan karakteristik akustik suara [2].

Beberapa penelitian telah mengembangkan sistem deteksi tingkat stres otomatis berbasis suara menggunakan jaringan saraf tiruan dan ekstraksi fitur MFCC, yang menunjukkan bahwa pendekatan berbasis suara dapat memberikan evaluasi kondisi psikologis yang lebih konsisten dibandingkan pengamatan manual [3]. Dibandingkan rekaman laboratorium atau emosi aktor pada dataset konvensional, *voice note* yang direkam menggunakan perangkat mobile merepresentasikan komunikasi yang lebih spontan dan lebih mampu menggambarkan kondisi emosional secara autentik [4] *Convolutional Neural Network* (CNN) memiliki kemampuan mengekstraksi fitur spektral dari representasi audio seperti *Mel-Frequency Cepstral Coefficients* (MFCC), yang merepresentasikan karakteristik spektral suara sesuai persepsi pendengaran manusia serta sensitif terhadap perubahan energi dan frekuensi terkait emosi dan stres [5]. *Long Short-Term Memory* (LSTM) di sisi lain efektif memodelkan pola temporal sinyal suara yang bersifat sekuensial, sehingga kombinasi CNN-LSTM banyak digunakan dalam *Speech Emotion Recognition* (SER) karena mampu menangkap karakteristik spasial dan temporal secara simultan [6]. Penelitian menunjukkan bahwa fitur MFCC memiliki karakteristik yang mampu membedakan suara individu yang mengalami stres dan tidak stres [7], dan pemrosesan fitur akustik menggunakan *deep learning* terbukti meningkatkan kemampuan sistem mengenali kondisi emosional manusia melalui suara [8].

Perlu ditegaskan bahwa terdapat perbedaan mendasar antara mengenali emosi dari suara, mendeteksi indikasi stres, dan mendiagnosis gangguan kesehatan mental secara klinis. Penelitian ini memetakan kategori emosi pada dataset standar ke dalam kategori indikasi tingkat stres berdasarkan kedekatan konseptualnya, sehingga sistem yang dihasilkan merupakan model klasifikasi eksperimental yang mengenali pola akustik terkait emosi/stres, dan tidak dimaksudkan sebagai alat ukur klinis maupun alat diagnosis medis/psikologis.

Sebagian besar penelitian deteksi stres berbasis suara masih menggunakan dataset emosi berbahasa asing atau rekaman terkontrol, sehingga kurang merepresentasikan variasi ekspresi emosional dalam komunikasi sehari-hari [4]. Oleh karena itu, penggunaan data voice note berbahasa Indonesia yang direkam secara alami menjadi penting untuk meningkatkan relevansi sistem deteksi stres pada konteks komunikasi digital masyarakat Indonesia [9].

Mengandalkan fitur akustik semata rentan terhadap kesalahan klasifikasi akibat variasi kualitas perangkat perekam dan gangguan suara latar, sehingga fusi antara analisis akustik berbasis deep learning dan analisis teks (linguistik) terbukti mampu meningkatkan akurasi dan mengungguli performa sistem unimodal [10]. Pendekatan *hybrid fusion* ini memerlukan tahap *Automatic Speech Recognition* menggunakan Whisper AI untuk mentranskripsikan *voice note* menjadi teks [11], yang kemudian dianalisis menggunakan *Zero-Shot Classification* berbasis Transformer untuk memetakan konteks semantik ke label tingkat stres tanpa memerlukan dataset berlabel khusus [12].

Berdasarkan permasalahan tersebut, penelitian ini mengembangkan sistem *Speech Emotion Recognition* berbasis pendekatan *hybrid* CNN-LSTM (akustik) dan *Zero-Shot Classification* IndoBERT (linguistik) untuk mengklasifikasikan tingkat stres dari *voice note* berbahasa Indonesia berdasarkan metodologi CRISP-DM, yang diimplementasikan ke dalam prototipe aplikasi *web* bernama *StressVoice.*

2. TINJAUAN PUSTAKA

2.1. Speech Emotion Recognition (SER)

SER merupakan bidang dalam *speech processing dan affective computing* yang bertujuan mengenali kondisi emosional melalui sinyal suara, menitikberatkan pada karakteristik non-verbal seperti pitch, intonasi, energi, dan durasi suara [9]. Dalam konteks kesehatan mental, kondisi stres sering tercermin melalui perubahan pola suara secara tidak sadar, sehingga fitur akustik seperti MFCC dan fitur prosodik mampu memberikan performa baik dalam mendeteksi kondisi emosional dari *voice note* [13].

2.2. Model Emosi Valence-Arousal

Emosi merupakan indikator penting dalam memahami kondisi psikologis. Teknik kecerdasan buatan memungkinkan identifikasi emosi secara otomatis melalui sinyal suara, teks, dan ekspresi wajah, membuka jalan bagi sistem monitoring kesehatan mental yang tidak invasif dan berkelanjutan [14]. Konsep deteksi emosi digunakan dalam penelitian ini untuk memahami hubungan antara keadaan emosional dan tingkat stres, dengan memetakan kategori emosi ke kategori tingkat stres melalui pendekatan SER.

2.3. Fitur Akustik MFCC

MFCC adalah metode ekstraksi fitur yang meniru cara sistem pendengaran manusia merespons frekuensi suara [15]. Proses ekstraksi MFCC terdiri dari *pre-emphasis, framing, windowing, transformasi Fourier, mel filter bank,* dan *discrete cosine transform* (DCT) [15]. Fitur MFCC banyak digunakan sebagai input model CNN dan LSTM dalam penelitian SER karena mampu merepresentasikan struktur spektral suara sekaligus mendukung pemodelan dinamika temporal [15].

2.4. Arsitek CNN-LSTM

Data audio memiliki karakteristik dua dimensi (waktu-frekuensi) dan bersifat sekuensial, sehingga kombinasi CNN dan LSTM dinilai efektif menangkap pola akustik dan dinamika emosi yang berubah seiring waktu [16]. CNN berperan sebagai feature extractor yang memproses representasi MFCC untuk menangkap pola lokal domain frekuensi, sedangkan keluaran CNN diteruskan ke LSTM untuk mempelajari hubungan temporal antar segmen suara. Pendekatan ini terbukti meningkatkan performa pengenalan emosi dibandingkan penggunaan model tunggal [17].

2.5. Hybrid Fusion Akustik dan Linguistik

Fusi antara analisis akustik dan fitur linguistik (teks) menghasilkan prediksi SER yang lebih stabil dan komprehensif dibandingkan sistem unimodal [10]. Whisper AI, yang dilatih menggunakan metode pengawasan lemah berskala besar, memiliki ketahanan tinggi terhadap gangguan noise dan variasi dialek sehingga cocok menjembatani domain audio ke domain teks [11]. Teks hasil transkripsi selanjutnya dianalisis menggunakan Zero-Shot Classification berbasis Transformer seperti IndoBERT untuk memetakan makna semantik ke label tingkat stres tanpa memerlukan dataset berlabel khusus untuk tugas tersebut [12].

2.6. Metode CRISP-DM

Penelitian ini menggunakan metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang terdiri dari enam tahapan sistematis, yaitu *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation*, dan *Deployment*. Metode ini banyak digunakan dalam penelitian klasifikasi berbasis data karena mampu mengakomodasi seluruh proses mulai dari pemahaman masalah hingga implementasi model [18].

3. METODE PENELITIAN

Penelitian ini dilaksanakan menggunakan kerangka kerja CRISP-DM yang sistematis dan iteratif untuk mengembangkan model klasifikasi tingkat stres berbasis *voice note* berbahasa Indonesia dengan arsitektur *hybrid* CNN-LSTM, dengan alur tahapan seperti ditunjukkan pada Gambar 1.

Gambar 1. Alur Penelitian CRISP-DM

3.1. Business Understanding

Tahap ini merumuskan permasalahan keterbatasan metode deteksi stres konvensional yang bersifat *self-report*, terjadwal, dan subjektif, serta menetapkan tujuan mengembangkan sistem klasifikasi tingkat stres otomatis berbasis *voice note* berbahasa

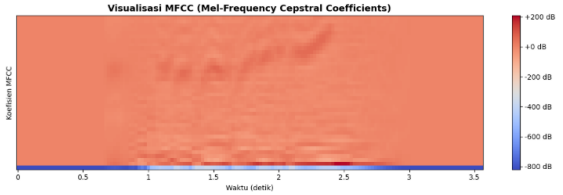
Indonesia ke dalam tiga kategori: *Non_Stress, Mild_Stress,* dan *High_Stress*, menggunakan arsitektur hybrid CNN-LSTM dan *Zero-Shot Classification* sebagai pelengkap, bukan pengganti, instrumen psikologi tervalidasi.

3.2. Data Understanding

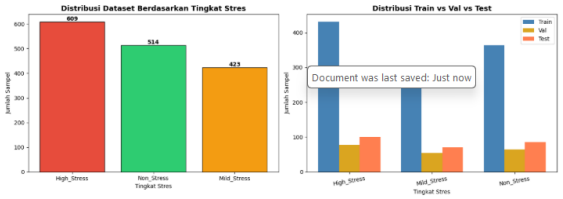
Data yang digunakan terdiri dari dataset sekunder RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) sebanyak 1.440 *file audio* berbahasa Inggris dengan 8 kategori emosi (*neutral, calm, happy, sad, angry, fearful, disgust, surprised*) yang direkam oleh 24 aktor profesional [19], serta dataset primer berupa 106 *voice note* berbahasa Indonesia berdurasi 10-30 detik yang dikumpulkan melalui Google Form dari responden dengan label self-report (34 file *Non_Stress*, 39 file *Mild_Stress*, 33 file *High_Stress*). Kedelapan kategori emosi RAVDESS dipetakan ke tiga kategori tingkat stres berdasarkan tingkat aktivasi emosi negatif menurut model *Valence-Arousal* [20], yaitu *Non_Stress* (*neutral, calm, happy*), *Mild_Stress* (*sad, surprised*), dan *High_Stress* (*angry, fearful, disgust*).

3.3. Data Preparation

Data suara diproses melalui seleksi kualitas (penghapusan file rusak atau berdurasi tidak sesuai), penataan struktur folder berdasarkan label, normalisasi amplitudo, trimming durasi, *noise reduction*, serta ekstraksi fitur akustik MFCC sebanyak 40 koefisien dengan panjang tetap 174 time frame (max_len = 174) menggunakan library librosa, sebagaimana divisualisasikan pada Gambar 2.



Gambar 2. Visualisasi MFCC



Gambar 3. Distribusi Dataset Berdasarkan Tingkat Stres

Distribusi dataset gabungan berdasarkan tingkat stres ditunjukkan pada Gambar 3. Pembagian dataset dilakukan berbeda untuk tiap sumber data guna mencegah kebocoran data (*data leakage*). Dataset RAVDESS dibagi secara *speaker-independent* berdasarkan nomor aktor: aktor 1-17 (1.020 sampel) sebagai data latih, aktor 18-20 (180 sampel) sebagai data validasi, dan aktor 21-24 (240 sampel) sebagai data uji. Dataset primer dibagi secara *stratified random split* (74 data latih, 16 data validasi, 16 data uji) karena setiap file berasal dari responden yang berbeda. Kedua subset digabungkan menghasilkan 1.094 sampel data latih, 196 sampel data validasi, dan 256 sampel data uji (total 1.546 sampel asli).

Untuk memperkaya variasi data latih tanpa menambah risiko kebocoran data, dilakukan *augmentasi audio* (*pitch shifting, time stretching, dan noise injection*) yang hanya diterapkan pada 1.020 file data latih RAVDESS, menghasilkan tambahan sampel sehingga total data latih menjadi 2.188 sampel, sementara data validasi dan data uji tidak diaugmentasi agar tetap merepresentasikan kondisi data asli saat evaluasi.

3.4. Modelling

Sistem dikembangkan menggunakan pendekatan fusi multimodal (*hybrid)*. Modalitas pertama menggunakan CNN-LSTM untuk mengekstraksi pola spasial-temporal dari fitur MFCC dengan bobot probabilitas 10%. Modalitas kedua menggunakan Whisper AI (*model medium*) untuk transkripsi *voice note* ke teks Bahasa Indonesia, yang kemudian dianalisis menggunakan *Zero-Shot Classification* berbasis Transformer (facebook/bart-large-mnli) dengan bobot probabilitas 90%. Bobot linguistik yang lebih dominan dipilih karena model CNN-LSTM dilatih pada dataset RAVDESS berbahasa Inggris, sehingga terdapat kesenjangan karakteristik akustik dengan tuturan Bahasa Indonesia.

Struktur model CNN-LSTM terdiri dari tiga blok konvolusi (Conv1D 64/128/256 filter, masing-masing disertai *BatchNormalization, MaxPooling*, dan *Dropout* 0,3), dua lapisan LSTM (128 dan 64 unit), serta dua dense layer (128 dan 64 unit dengan Dropout 0,4 dan 0,3) sebelum lapisan *output dense* 3 unit beraktivasi *softmax*. Model dikompilasi menggunakan optimizer Adam (*learning rate* 0,001) dengan

fungsi *loss categorical crossentropy*, dilatih maksimal 100 *epoch* dengan *batch size* 32, serta tiga *callback: EarlyStopping (patience* 15, memantau *val_loss* dan mengembalikan bobot terbaik), *ModelCheckpoint* (menyimpan bobot dengan *val_accuracy* terbaik), dan ReduceLROnPlateau (menurunkan *learning rate* 50% apabila *val_loss* tidak membaik selama 5 epoch berturut-turut).

3.5. Evaluation

Kinerja model CNN-LSTM diukur pada data uji yang independen dari data latih menggunakan metrik Akurasi, Presisi, Recall, dan F1-Score, serta *Confusion Matrix* untuk menganalisis distribusi prediksi benar dan salah pada tiap kelas tingkat stres.

$$Akurasi = (TP + TN) / (TP + TN + FP + FN)$$

$$Presisi = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1\text{-}Score = 2 \times (Presisi \times Recall) / (Presisi + Recall)$$

3.6. Evaluation

Model terbaik disimpan dalam format .h5 dan diintegrasikan ke dalam prototipe aplikasi *web StressVoice* berbasis *framework Flask* (*Python*) sebagai *backend* dan HTML/CSS/JavaScript sebagai *frontend,* terkoneksi dengan basis data MySQL (MariaDB via Laragon). Sistem menyediakan dua antarmuka: (1) antarmuka pengguna untuk merekam atau mengunggah *voice note* dan menampilkan hasil klasifikasi, transkripsi, respons empatik, cetak surat rujukan psikologis (PDF), serta tautan konsultasi WhatsApp; dan (2) antarmuka administrator yang dilindungi autentikasi login untuk mengelola riwayat data (CRUD) dan memutar ulang audio pengguna. Pendekatan deployment berbasis *web* lokal ini sejalan dengan riset serupa di lingkungan Universitas Muhammadiyah Sukabumi yang telah mendemonstrasikan integrasi model *deep learning* ke dalam arsitektur *web* untuk otomatisasi komoditas lokal [21], serta konsisten dengan pola *deployment* berbasis CRISP-DM pada penelitian klasifikasi citra berbasis *web* lainnya [22]

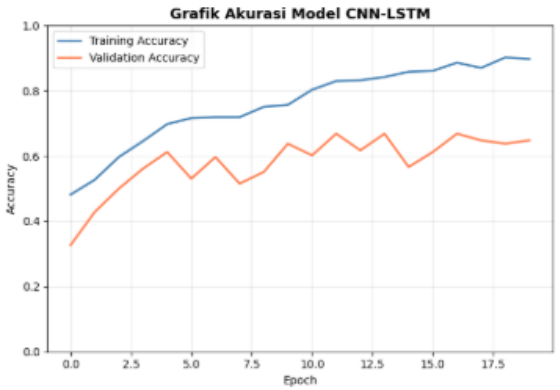
4. HASIL DAN PEMBAHASAN

4.1. Hasil Pelatihan Model

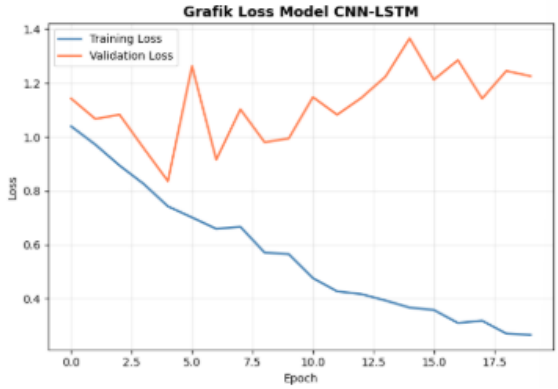
Proses pelatihan model dieksekusi dengan batasan maksimal 100 epoch dan dihentikan otomatis oleh *EarlyStopping* pada *epoch* ke-20. Model mencapai akurasi training terbaik 82,97% dan akurasi validasi terbaik 66,84% pada *epoch* ke-12 (Tabel 1). Karena kriteria *EarlyStopping* (berbasis val_loss, optimal pada *epoch* 5) dan ModelCheckpoint (berbasis *val_accuracy*, optimal pada epoch 12) mengarah ke *epoch* yang berbeda, bobot model terbaik dimuat ulang secara eksplisit dari *checkpoint val_accuracy* tertinggi menggunakan *load_model*(). Pola divergensi antara *training loss* yang turun konsisten dari 1,05 ke 0,27 dan *validation loss* yang hanya menurun sampai *epoch* 5 (0,84) lalu berfluktuasi naik ke rentang 1,0-1,4 (Gambar 4 dan Gambar 5), mengindikasikan model masih menunjukkan gejala *overfitting* terhadap data latih.

Tabel 1. Ringkasan Hasil Pelatihan Model CNN-LSTM

Parameter	Nilai	Keterangan
Akurasi Training Terbaik	82,97%	*Epoch* 12
Akurasi Validasi Terbaik	66,84%	*Epoch* 12
Akurasi Data Testing	62,11%	256 sampel uji
Best Epoch (checkpoint dipakai)	*Epoch* 12	Dimuat ulang via *load_model*()
Epoch Training (Total)	20 *Epoch*	Dihentikan *EarlyStopping*



Gambar 4. Grafik Akurasi Model CNN-LSTM



Gambar 5. Grafik Loss Model CNN-LSTM

4.2. Evaluasi Kinerja Klasifikasi (Testing)

Pengujian pada 256 sampel data uji yang sepenuhnya independen dari data latih menghasilkan akurasi model CNN-LSTM sebesar 62,11% (159 dari 256 sampel diprediksi benar) sesuai Persamaan (1). Tabel 2 merangkum *classification report* untuk setiap kelas tingkat stres.

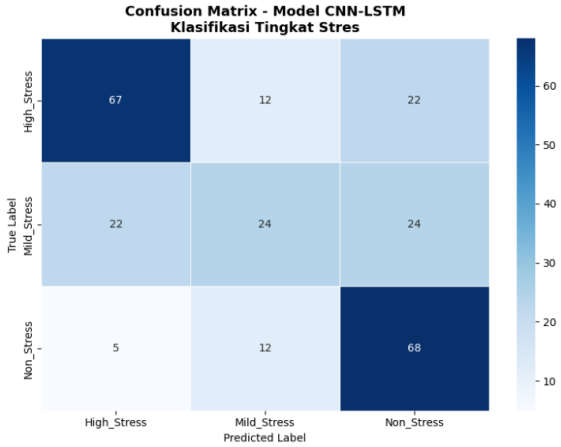
Tabel 2. *Classification Report Model* CNN-LSTM

Kelas	Precision	Recall	F1-Score	Support
High_Stress	0,71	0,66	0,69	101
Mild_Stress	0,50	0,34	0,41	71
Non_Stress	0,60	0,80	0,68	85
Macro Avrage	0,60	0,60	0,59	256
Weighted Average	0,62	0,62	0,61	256

Berdasarkan Tabel 2, kelas *High_Stress* memperoleh F1-score tertinggi (0,69), sedangkan kelas *Mild_Stress* memperoleh F1-score terendah (0,41) karena secara konseptual berada pada area transisi antara dua kutub emosi ekstrem, sehingga karakteristik akustiknya paling ambigu.

Tabel 3. *Confusion Matrix* Model CNN-LSTM (256 Sampel Uji)

Aktual \ Prediksi	*High Stress*	*Mild Stress*	*Non_ Stress*	Total
High_Stress	67	12	22	101
Mild_Stress	22	24	24	70
Non_Stress	5	12	68	85



Gambar 6. Visualisasi *Heatmap Confusion Matrix Model* CNN-LSTM

Visualisasi *heatmap* pada Gambar 6 memperkuat pembacaan Tabel 3. Berdasarkan Tabel 3, dari 101 sampel *High_Stress*, 67 sampel diprediksi benar, 12 sampel salah diprediksi sebagai *Mild_Stress*, dan 22 sampel salah diprediksi sebagai *Non_Stress.* Dari 70 sampel *Mild_Stress*, 22 sampel salah diprediksi sebagai *High_Stress*, 24 sampel diprediksi benar, dan 24 sampel salah diprediksi sebagai Non_Stress — lebih dari separuh sampel kelas ini (46 dari 70, atau 66%) salah diklasifikasikan, konsisten dengan F1-score terendahnya. Dari 85 sampel *Non_Stress*, 5 sampel salah diprediksi sebagai *High_Stress*, 12 sampel salah diprediksi sebagai *Mild_Stress*, dan 68 sampel diprediksi benar. Dengan demikian, total prediksi benar (diagonal utama) berjumlah 67 + 24 + 68 = 159 sampel dari 256 sampel uji, menghasilkan akurasi keseluruhan 62,11% sesuai Persamaan (1). Kesalahan klasifikasi terkonsentrasi pada kelas *Mild_Stress*, sejalan dengan F1-score kelas tersebut yang jauh lebih rendah dibandingkan dua kelas lainnya.

Sebagai perbandingan, penelitian SER lain yang memanfaatkan arsitektur pretrained berbasis *self-supervised learning*, yaitu Wav2Vec 2.0 yang di-fine-tuning pada dataset CREMA-D untuk empat kelas emosi (*angry, happy, neutral, sad*), berhasil mencapai akurasi validasi 78–79% dengan *weighted* F1-score 0,79 [22]. Perbandingan ini mengindikasikan bahwa model akustik berbasis representasi pretrained seperti Wav2Vec 2.0 berpotensi memberikan performa yang lebih tinggi dibandingkan arsitektur CNN-LSTM yang dilatih dari awal (*trained from scratch*) pada penelitian ini, sehingga eksplorasi arsitektur

serupa menjadi salah satu arah pengembangan yang relevan untuk meningkatkan akurasi modalitas akustik pada penelitian selanjutnya.

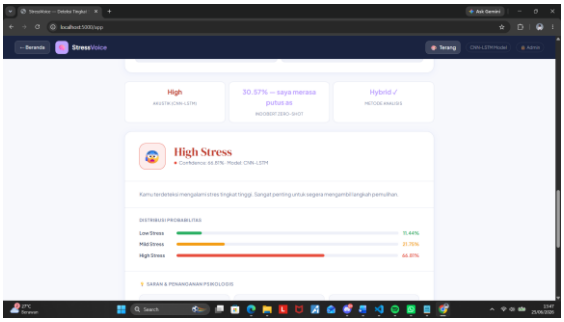
4.3. Analisis Efektivitas Hybrid Fusion

Pendekatan *hybrid fusion* yang menggabungkan analisis akustik CNN-LSTM (10%) dan analisis linguistik IndoBERT Zero-Shot (90%) dirancang untuk mengatasi keterbatasan performa akustik murni akibat kesenjangan bahasa antara dataset pelatihan (RAVDESS, berbahasa Inggris) dan *voice note* pengguna (berbahasa Indonesia). Dengan bobot linguistik yang dominan, sistem tetap dapat mengklasifikasikan tingkat stres berdasarkan makna semantik dan konteks isi percakapan meskipun kontribusi akustik terbatas, sehingga sistem lebih relevan diterapkan pada konteks pengguna berbahasa Indonesia dibandingkan mengandalkan model akustik saja. Perlu dicatat bahwa evaluasi menyeluruh terhadap skema fusi gabungan (accuracy, F1-score, dan *confusion matrix* hasil fusi) belum dilakukan pada penelitian ini; akurasi 62,11% yang dilaporkan merupakan hasil evaluasi murni modalitas CNN-LSTM akustik, sehingga evaluasi menyeluruh terhadap skema *hybrid fusion* diusulkan sebagai penelitian lanjutan.

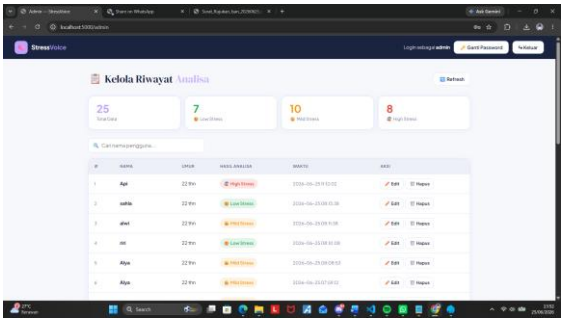
4.4. Implementasi Sistem StressVoice

Model CNN-LSTM diintegrasikan ke dalam prototipe *StressVoice* berbasis *Flask* yang menggabungkan tiga komponen: model CNN-LSTM (analisis akustik), Whisper AI model medium (transkripsi suara ke teks Bahasa Indonesia), dan facebook/bart-large-mnli (*Zero-Shot Classification linguistik*). Sistem terdiri dari empat halaman utama *landing page* (edukasi kesehatan mental berdasarkan literatur DSM-5, WHO, dan APA), halaman analisis (unggah audio atau rekam langsung, mendukung format .wav, .mp3, .ogg, .m4a, .webm, dan .flac dengan durasi maksimal 60 detik), halaman riwayat analisis (statistik agregat distribusi *Low/Mild/High Stress*), dan halaman admin (CRUD data dengan autentikasi login). Sistem juga dilengkapi fitur cetak surat rujukan psikologis dalam format PDF serta tombol konsultasi WhatsApp langsung ke psikolog dengan pesan otomatis berisi hasil analisis. Rata-rata waktu pemrosesan total sistem (ekstraksi MFCC, CNN-LSTM, Whisper, dan *Zero-Shot Classification*) adalah 30-60 detik per *voice note* menggunakan CPU tanpa akselerasi GPU. Tampilan antarmuka *landing page*, hasil analisis, dan *dashboard* admin *StressVoice* berturut-turut ditunjukkan pada Gambar 7, Gambar 8, dan Gambar 9.

Gambar 7. Tampilan Halaman *Landing Page StressVoice*



Gambar 8. Contoh Tampilan Analisis Tingkat *High Stress*



Gambar 9. Tampilan *Dashboard Admin StressVoice*

4.5. Keterbatasan Sistem

Model CNN-LSTM masih menunjukkan gejala *overfitting* akibat kesenjangan antara data latih (berbahasa Inggris, RAVDESS) dan konteks penggunaan nyata (berbahasa Indonesia), meskipun keterbatasan ini sebagian diatasi melalui bobot fusi yang mengurangi kontribusi akustik. Transkripsi Whisper AI kurang akurat pada tuturan dengan bahasa gaul yang sangat informal, nama orang, atau kualitas suara yang buruk. Selain itu, waktu pemrosesan 30-60 detik per analisis pada CPU tanpa GPU

dapat terasa lama bagi pengguna yang mengharapkan respons instan.

5. KESIMPULAN

1. Hasil yang Diperoleh Sistem SER berbasis *hybrid* CNN-LSTM untuk klasifikasi tingkat stres dari *voice note* Indonesia berhasil dikembangkan dan diuji pada 256 sampel independen, mencapai akurasi 62,11% dengan performa terbaik pada kelas *High_Stress* (F1-score 0,69).
2. Kelebihan Sistem *Hybrid fusion* (linguistik 90%, akustik 10%) berhasil mengatasi kesenjangan bahasa antara dataset RAVDESS dan pengguna Indonesia. *StressVoice* juga dilengkapi transkripsi otomatis Whisper AI, respons empatik, surat rujukan PDF, dan konsultasi WhatsApp ke psikolog.
3. Kekurangan Sistem Model CNN-LSTM masih *overfitting* dengan performa rendah pada kelas *Mild_Stress* (F1-score 0,41) akibat karakteristik akustik yang ambigu. Transkripsi Whisper AI kurang akurat pada bahasa gaul informal, dan waktu pemrosesan 30-60 detik per analisis pada CPU tanpa GPU masih tergolong lama.
4. Pengembangan Selanjutnya Disarankan memperluas dataset *voice note* berbahasa Indonesia yang lebih besar dan representatif, mengeksplorasi arsitektur berbasis Transformer (Wav2Vec 2.0 atau HuBERT) serta model bahasa Indonesia yang lebih spesifik (IndoBERT/IndoNLU), melakukan evaluasi menyeluruh terhadap skema *hybrid fusion* gabungan, dan mengembangkan dukungan analisis *real-time* serta aplikasi mobile untuk memperluas jangkauan sistem.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sukabumi, serta seluruh pihak yang telah memberikan bimbingan dan dukungan terhadap penyelesaian penelitian ini.

DAFTAR PUSTAKA

[1] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism," *Electron.*, vol. 12, no. 20, pp. 1–31, 2023, doi: 10.3390/electronics12204376.

[2] Vamsinath J, Varshini Bonagiri, Sandeep T, Meghana V, and Latha B, "Stress Detection Through Speech Analysis Using Machine Learning," *Int. J. Sci. Res. Sci. Technol.*, pp. 334–342, 2022, doi: 10.32628/ijsrst229437.

[3] M. N. Aljufri and B. H. Prasetio, "Sistem Deteksi Tingkat Stress Menggunakan Suara dengan Metode Jaringan Saraf Tiruan dan Ekstraksi Fitur MFCC berbasis Raspberry Pi," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 11, pp. 5278–5285, 2022, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/11842

[4] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1634–1654, 2023, doi: 10.1109/TAFFC.2021.3114365.

[5] M. E. P. Rismanto and I. Handayani, "Klasifikasi Emosi Berdasarkan Suara dengan Metode Convolutional Neural Network," *J. Inform. Univ. Pamulang*, vol. 9, no. 4, pp. 163–171, 2025, doi: 10.32493/informatika.v9i4.45236.

[6] Y. Hifny and A. Ali, "Efficient Arabic Emotion Recognition Using Deep Neural Networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 6710–6714, 2024, doi: 10.1109/ICASSP.2019.8683632.

[7] H. Rheza Paleva and B. Henryranu Prasetio, "Penerapan Short Time Fourier Transform pada MFCC untuk Sistem Pengenalan Ucapan Tingkat Stres," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 1, pp. 1–9, 2024, [Online]. Available: http://j-ptiik.ub.ac.id

[8] A. A. Kasim, M. Bakri, I. Mahmudi, R. Rahmawati, and Z. Zulnabil, "Artificial Intelligent for Human Emotion Detection with the Mel-Frequency Cepstral Coefficient (MFCC)," *JUITA J. Inform.*, vol. 11, no. 1, p. 47, 2023, doi: 10.30595/juita.v11i1.15435.

[9] F. KASYIDI, R. ILYAS, and N. M. ANNISA, "Peningkatan Kemampuan Pengenalan Emosi Melalui Suara dalam Bahasa Indonesia," *MIND J.*, vol. 6, no. 2, pp. 194–204, 2021, doi: 10.26760/mindjournal.v6i2.194-204.

[10] N. N. Y. Truong *et al.*, "DFAT: Dual-stage Fusion of Acoustic and Text feature for Speech Emotion Recognition," *Proc. 11th Int. Work. Vietnamese Lang. Speech Process.*, pp. 36–44, 2025, [Online]. Available: https://aclanthology.org/2025.vlsp-1.6/

[11] T. Srivastaval, J. C. Chou, P. Shroffl, K. Livescu, and C. Graziul, "Speech Recognition For Analysis of Police Radio

Communication," *Proc. 2024 IEEE Spok. Lang. Technol. Work. SLT 2024*, pp. 906–912, 2024, doi: 10.1109/SLT61566.2024.10832157.

[12] S. G. Tesfagergish, J. Kapočiūtė-Dzikienė, and R. Damaševičius, "Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning," *Appl. Sci.*, vol. 12, no. 17, 2022, doi: 10.3390/app12178662.

[13] A. D. Fathulramdhan and B. H. Prasetyo, "Pengembangan Sistem Deteksi Stres Berbasis Suara Menggunakan Fitur MFCC, ZCR, dan SC Dengan Metode Artificial Neural Network (ANN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 9, pp. 2548–964, 2025, [Online]. Available: http://j-ptiik.ub.ac.id

[14] M. N. Hadi and R. W. Sari, "Multi-Label Emotion Detection for Mental Health Monitoring Using Deep CNN and Visual Attention," *J. Artif. Intell. Softw. Eng.*, vol. 5, no. 2, pp. 597–605, 2025, doi: 10.30811/jaise.v5i2.6961.

[15] A. Roihan, R. Zein, M. R. Aprianti, F. N. Izzah, and A. Luthfunnisa, "Analisis Teknologi Speech Emotion Recognition (SER): Pendekatan Fitur Akustik, Klasifikasi, Keamanan Dan Implementasi Pada Sistem Portabel," *J. Sist. Inf. dan Teknol.*, vol. 5, no. 2, pp. 144–149, 2025, doi: 10.56995/sintek.v5i2.167.

[16] A. S. Sams and A. Zahra, "Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers," *Bull. Electr. Eng. Informatics*, vol. 12, no. 1, pp. 355–364, 2023, doi: 10.11591/eei.v12i1.4231.

[17] F. Makhmudov, A. Kutlimuratov, and Y. I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," *Appl. Sci.*, vol. 14, no. 23, 2024, doi: 10.3390/app142311342.

[18] A. Rianti, N. W. A. Majid, and A. Fauzi, "CRISP-DM: Metodologi Proyek Data Science," *Pros. Semin. Nas. Teknol. Inf. dan Bisnis 2023*, pp. 107–104, 2023.

[19] G. T. Waleed and S. H. Shaker, "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN," *Inf.*, vol. 16, no. 7, 2025, doi: 10.3390/info16070518.

[20] R. Ullah *et al.*, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer," *Sensors*, vol. 23, no. 13, pp. 1–20, 2023, doi: 10.3390/s23136212.

[21] R. Akyas Hifdzi Rahman, A. Adi Sunarto, and A. Asriyanik, "Penerapan You Only Look Once (Yolo) V8 Untuk Deteksi Tingkat Kematangan Buah Manggis," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 5, pp. 10566–10571, 2024, doi: 10.36040/jati.v8i5.10979.

[22] T. Putra, Dzaky, Putratama , Al Fikri, "Speech Emotion Recognition Deep Learning Karakter Kucing Interaktif Di Game Unity," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 14, no. 1, pp. 1583–1591, 2026.