Analisis Pola Emosi dari Perspektif Speech Emotion Recognition untuk Identifikasi Tingkat Stres Menggunakan CNN dan LSTM

1st Defhita Zahra Sastika
Teknik Informatika
Telkom Univeristy Purwokerto
Purwokerto, Indonesia
defhitazahra@student.telkomuniversity.ac.id

2nd Yesy Diah Rosita
Teknik Informatika
Telkom Univeristy Purwokerto
Purwokerto, Indonesia
yesydr@telkomuniversity.ac.id

3rd M.Raafi'u Firmansyah
Teknik Informatika
Telkom Univeristy Purwokerto
Purwokerto, Indonesia
raafiu@telkomuniversity.ac.id

Abstrak* — Penelitian ini bertujuan mengidentifikasi tingkat stres melalui analisis pola emosi berbasis suara menggunakan Speech Emotion Recognition (SER). Metode yang digunakan adalah kombinasi Convolutional Neural Network (CNN) untuk ekstraksi fitur dan Long Short-Term Memory (LSTM) untuk analisis temporal. Data rekaman suara diproses melalui tahap pre-processing, augmentasi, dan ekstraksi fitur menggunakan log-Mel spectrogram. Hasil penelitian menunjukkan bahwa model CNN mencapai akurasi 96,86%, sedangkan model LSTM memperoleh akurasi 86,7% pada data uji. Evaluasi menggunakan K-Fold Cross Validation menghasilkan rata-rata akurasi 82,9% dengan F1-score 0,819. Hasil ini menunjukkan bahwa pendekatan CNN-LSTM efektif dalam mendeteksi tingkat stres secara non-invasif.

***Kata kunci*— speech emotion recognition, cnn-lstm, emosi, stres, deep learning**

I. PENDAHULUAN

Kesehatan mental adalah komponen penting dalam kehidupan manusia yang mempengaruhi produktivitas, hubungan sosial, dan juga kualitas hidup. Salah satu indikator yang mencerminkan kondisi kesehatan mental seseorang adalah keadaan emosional yang terekspresi dalam perilaku seseorang sehari-hari, termasuk dalam cara seseorang berbicara. Data empiris menunjukkan bahwa tingkat stres di masyarakat, khususnya kalangan usia produktif masih menjadi perhatian serius. Menurut *World Health Organization* (WHO), lebih dari 970 juta individu di dunia megalami gangguan kesehatan mental terutama depresi dan kecemasan [1]. Sedangkan di Indonesia, sekitar 6,1% penduduk berusia 15 tahun keatas mengalami gangguan kesehatan mental [2].

Pendekatan konvensional dalam mengidentifikasi tingkat stres masih bersifat subjektif, seperti melalui survei atau observasi manual, sehingga kurang efektif dalam memberikan hasil yang objektif dan berkelanjutan. Oleh karena itu, diperlukan pendekatan berbasis teknologi yang mampu mengidentifikasi kondisi emosional secara otomatis.

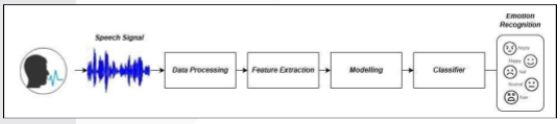
Speech Emotion Recognition (SER) menjadi salah satu solusi yang relevan karena mampu menganalisis karakteristik suara seperti intonasi, pitch, dan tempo untuk mendeteksi emosi. Penelitian sebelumnya menunjukkan bahwa kombinasi metode Convolutional Neural Network (CNN) dan Long Short-Term Memory (LSTM) mampu meningkatkan akurasi dalam klasifikasi emosi karena dapat menangkap fitur spasial dan temporal secara bersamaan.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menganalisis pola emosi dari data suara menggunakan pendekatan CNN-LSTM guna mengidentifikasi tingkat stres secara objektif.

II. KAJIAN TEORI

A. *Speech Emotion Recognition*

Speech Emotion Recognition (SER) adalah teknologi yang digunakan untuk mengidentifikasi dan mengklasifikasikan emosi seseorang berdasarkan karakteristik suara mereka. Teknologi ini bekerja dengan menganalisis sinyal suara dan mengekstraksi fitur-fitur tertentu seperti intonasi, nada (pitch), kecepatan bicara, intensitas, dan ritme yang berkaitan dengan ekspresi emosional [18].



GAMBAR 1
(System Overview SER)

Gambar 1 menunjukkan *system overview* dari penggunaan *Speech Emotion Recognition* dimana proses dimulai dari audio mentah dan dilakukan data

pre-processing hingga menghasilkan *output* berupa *Emotion Recognition*.

B. Emosi

Emosi dapat dipahami sebagai gambaran dari pikiran dan perasaan, yang muncul sebagai respons subjektif terhadap suatu pengalaman, disertai perubahan fisiologis dan perilaku.[15].

C. Stress

Stres adalah kondisi tekanan yang dapat memengaruhi emosi, pola pikir, serta keadaan fisik individy. Kondisi ini muncul saat individu merasa tuntutan yang dihadapi melampaui kemampuan atau sumber daya yang dimilikinya [16]. Stres memiliki hubungan antara individu dan lingkungannya yang membebani atau melebihi kapasitas individu serta mengancam kesejahteraannya.

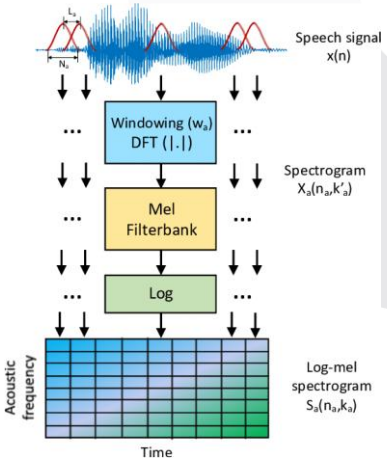
Gambar 2.2 Arsitektur CNN

Gambar 2.2 menunjukkan arsirtektur CNN bahwa CNN terdiri dari *input*, *output* dan *hidden layer*, dimana *layer-layer* ini terdiri dari tiga *tipe*, yaitu *convolution, pooling dan fully-connected layer.*

E. Log Mel Spektogram

Log- Mel spectrogram merupakan representasi visual dari spektrum frekuensi sinyal audio terhadap

waktu, di mana skala frekuensi diubah dari skala Hertz ke skala Mel.



GAMBAR 2
(Blok Diagram Log-Mel Spektrogram)

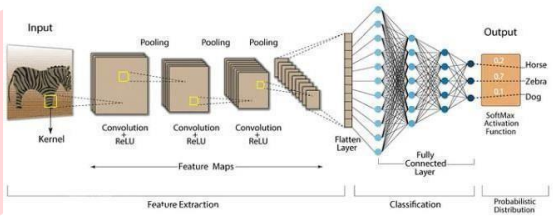
Gambar 2 diatas menunjukkan bahwa *Log-Mel spectrogram* menghasilkan matriks berbobot yang lebih informatif,

F. Time Stretching

Mengubah kecepatan suara (lebih cepat atau lebih lambat) tanpa mengubah *pitch*, untuk mensimulasikan gaya bicara yang berbeda.

D. Convolutional Neural Network

Convolutional Neural Network (CNN) adalah salah satu pendekatan dalam *deep learning* yang masuk kedalam jaringan saraf tiruan, dan umum digunakan untuk klasifikasi citra serta pengenalan objek [18]. CNN memiliki tiga komponen utama, yaitu *input layer, hidden layer*, dan *output layer*, di mana pada *hidden layer* terdapat tiga jenis lapisan, yaitu *convolution*, *pooling*, dan *fully connected layer* [19].



GAMBAR 3
(Lapisan CNN)

G. Noise Injection

Menambahkan kebisingan latar belakang untuk membuat model lebih toleran terhadap data noisy.

III. METODE

A. Alat dan Bahan Penelitian

Dalam proses penelitian ini, dibutuhkan alat dan bahan yang akan membantu baik dalam pengolahan data hasil penelitian, hingga dalam pembuatan model. Berikut adalah perlengkapan yang digunakan selama penelitian berlangsung yang terdiri dari perangkat keras dan perangkat lunak.

Bahan penelitian yang digunakan berupa rekaman suara mahasiswa yang merepresentasikan ekspresi emosi mereka. Pengambilan data dilakukan secara konsisten dengan prosedur yang sama, yaitu pada rentang waktu pukul 08.00 WIB hingga 15.00 WIB untuk meminimalkan variabel eksternal yang dapat mempengaruhi kondisi emosional subjek. Selama proses perekaman, mikrofon diletakkan pada jarak kurang lebih 10 cm dari mulut subjek guna memastikan kualitas suara yang optimal dan konsistensi dalam akuisisi data. Pengambilan dataset berdurasi pendek menggunakan naskah yang berisikan kalimat untuk masing-masing kelas emosi, responden diminta untuk mengucapkan sesuai dengan kelas emosinya. Sedangkan untuk dataset audio berdurasi panjang, responden diminta untuk menceritakan apa yang sedang dialami dan dirasakan responden pada saat pengambilan data suara.

B. Subjek dan Objek Penelitian

Subjek yang dianalisis dalam penelitian ini adalah penggunaan metode *Speech Emotion Recognition (SER)* yang diimplementasikan ke dalam model berbasis CNN dan LSTM untuk

mengidentifikasi emosi dalam suara. Objek dari penelitian ini adalah rekaman suara mahasiswa yang mengandung berbagai ekspresi emosi.

C. Diagram Alir Penelitian

GAMBAR 4
(Diagram Alir Penelitian)

Gambar 4 menjelaskan bahwa penelitian dilakukan dengan melakukan pengumpulan data suara sebagai bahan utama analisis.

D. *Labelling* Data

Label yang diberikan berupa klasifikasi emosi, yaitu marah, senang, sedih, netral, dan takut. Selain itu, untuk kebutuhan analisis tingkat stres, setiap potongan suara juga dilabeli berdasarkan kategori tingkat stres yaitu tinggi, sedang, dan rendah oleh Psikolog Klinis & Praktisi Anak – Remaja, Qurrata A'yun, M.Psi., Psikolog (STR: PE00001669119339) di Insight Psikologi & Enderun Islamic Preschool.

E. Data Preparation

1. Augmentasi

Menambahkan noise buatan ke dalam rekaman suara asli dengan tujuan meningkatkan ketahanan model terhadap gangguan suara dari lingkungan nyata. Pada penelitian ini, noise ditambahkan dengan rasio signal-to-noise (SNR) sebesar 20 dB. Teknik augmentasi berupa *time stretching* diterapkan dengan rentang ±10%, yaitu mempercepat atau memperlambat tempo suara tanpa mengubah *pitch*.

2. Ekstraksi Fitur

Feature extraction adalah proses pengambilan informasi penting dari data suara untuk dijadikan input ke dalam model. Fitur yang digunakan dalam pengenalan emosi suara adalah log mel spektogram.

3. Modelling

Setelah fitur diekstraksi, tahap selanjutnya adalah membangun dan melatih model menggunakan arsitektur CNN-LSTM. CNN menggunakan Conv2D 3 *layer* dengan besaran *layer* 32, 64, dan 128 untuk menangkap fitur secara kompleks. LSTM digunakan untuk menganalisis pola temporal atau urutan dari data tersebut dengan menggunakan 1 *layer* LSTM dan 2 *fully connected layer* untuk mempelajari pola emosi. Kombinasi kedua arsitektur ini efektif dalam memahami konteks suara dalam dimensi waktu dan ruang, sehingga mampu mengenali pola emosi atau stres dengan lebih baik.

4. Model Evaluation

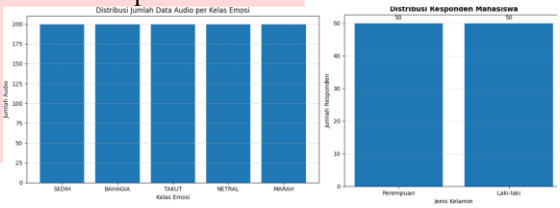
Setelah model selesai dilatih, langkah berikutnya adalah melakukan evaluasi terhadap performa model tersebut. Evaluasi dilakukan dengan menggunakan data uji (testing set) dan mengukur berbagai metrik seperti akurasi, precision, recall, F1- score, dan confusion matrix dengan Cross Validation sebagai metrik utama.

IV. HASIL DAN PEMBAHASAN

A. Hasil

1. Pengumpulan dan Karakteristik Data

Penelitian ini menggunakan dua jenis dataset audio dengan karakteristik dan tujuan yang berbeda, yaitu dataset audio emosi mahasiswa berdurasi pendek (3 detik) dan dataset audio mahasiswa berdurasi panjang (10 menit). Tahap ini juga mencakup Exploratory Data Analysis (EDA) yang bertujuan untuk memahami distribusi, karakteristik, serta kualitas dataset sebelum dilakukan pemodelan.



GAMBAR 5
(Distribusi Dataset)

2. Pra-pemrosesan Data Audio

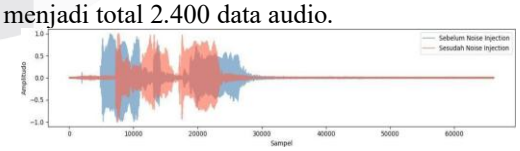
Berdasarkan proses pra-pemrosesan yang telah dilakukan, seluruh data audio pada penelitian ini berhasil diseragamkan dengan sampling rate sebesar 22.050 Hz serta dinormalisasi ke dalam rentang amplitudo −1 hingga 1.

3. Pembagian Data

Pembagian data dilakukan setelah tahap pra-pemrosesan, dengan tujuan memisahkan data menjadi data latih dan data uji menggunakan rasio 80:20. Proses pembagian data dilakukan sebelum augmentasi untuk memastikan bahwa data uji tetap mempresentasikan kondisi asli sehingga evaluasi model dapat dilakukan secara objektif tanpa adanya kebocoran data *(data leakage).*

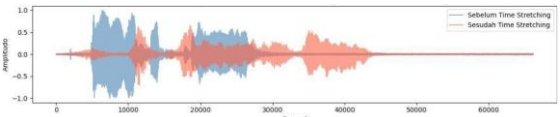
4. Augmentasi Data

Berdasarkan proses augmentasi data yang telah dilakukan pada data latih, jumlah dataset audio emosi meningkat dari 800 data asli menjadi total 2.400 data audio.



GAMBAR 6
(*Waveform Comparison* Sebelum *vs* Sesudah *Noise Injection)*

Gambar 6 menampilkan perbandingan amplitudo sinyal sebelum dan sesudah ditambahkan *noise*.

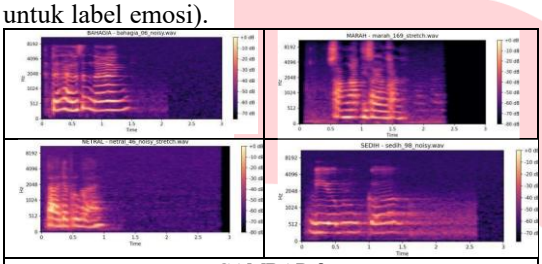


GAMBAR 7
(Waveform *Comparion Time Stretching)*

Gambar 7 menampilkan perbandingan gelombang audio sebelum dan sesudah dilakukan *time stretching*.

5. Ekstraksi Fitur

Setelah augmentasi, dataset audio sebanyak 2.400 file berhasil diekstraksi fiturnya menggunakan *Log-Mel Spectrogram* dengan 128 filter Mel dan *sampling rate* 22.050 Hz, setiap sinyal audio terlebih dahulu diseragamkan durasinya menjadi 3 detik untuk menjaga konsistensi panjang input kemudian disimpan dalam format *.npy* (*x.npy* untuk fitur dan *y.npy* untuk label emosi).

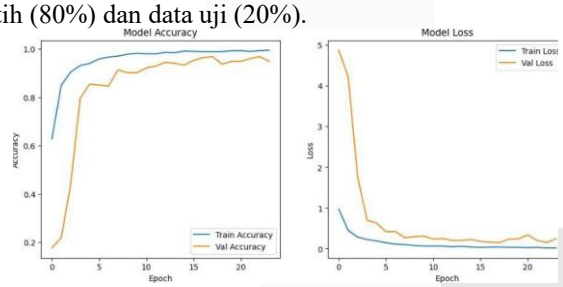


GAMBAR 8
(Ekstraksi Fitur)

Gambar 8 menunjukkan visualisasi dari hasil ekstraksi fitur untuk masing-masing kelas emosi, dengan frekuensi dan juga temporal untuk masing-masing kelas.

6. Pelatihan Model

Berdasarkan program pelatihan model CNN 2D yang dijalankan, data fitur log-Mel *Spectrogram* telah berhasil dibagi menjadi data latih (80%) dan data uji (20%).



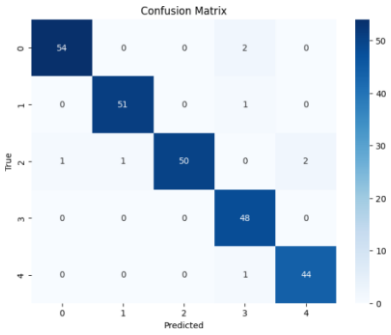
GAMBAR 9
(Kurva *Train & Val Accuracy* dan *Loss)*

Berdasarkan **Gambar 9** diatas, Permodelan CNN yang dibangun mampu mencapai performa klasifikasi yang tinggi terhadap data emosi. Hal ini menunjukkan bahwa fitur Log-Mel Spectrogram yang digunakan mampu merepresentasikan karakteristik emoosi secara efektif.

7. Pengujian Model

Hasil pengujian model CNN 2D pada data uji sebanyak 200 sampel menunjukkan performa yang sangat baik dengan akurasi keseluruhan 96,86%.

	precision	recall	f1-score	support
Emosi1	0.98	0.96	0.97	56
Emosi2	0.98	0.98	0.98	52
Emosi3	1.00	0.93	0.96	54
Emosi4	0.92	1.00	0.96	48
Emosi5	0.96	0.98	0.97	45
accuracy			0.97	255
macro avg	0.97	0.97	0.97	255
weighted avg	0.97	0.97	0.97	255



GAMBAR 10
(Confusion Matrix Model CNN2D)

Berdasarkan **Gambar 10** terlihat bahwa sebagian besar prediksi berada pada diagonal utama, menandakan model mampu mengenali tiap emosi dengan sangat baik. Kesalahan prediksi hanya terjadi pada beberapa sampel, khususnya pada kelas emosi yang memiliki pola yang mirip, sehingga menyebabkan sedikit penurunan akurasi pada kelas tersebut.

8. Pra-pemrosesan Audio 10 Menit

Pada tahap ini, seluruh audio diproses melalui teknik segmentasi untuk mengubah sinyal audio kontinu menjadi bentuk data sekuensial dan berhasil dibagi menjadi ±300 segmen tetap per file, sehingga setiap segmen berdurasi 3 detik (tergantung panjang total audio) .

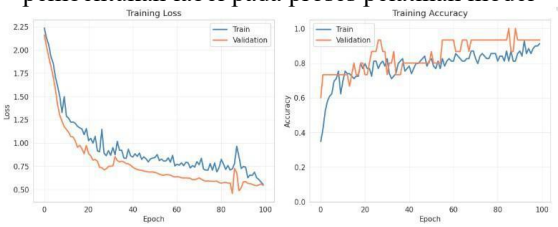
GAMBAR 11
(Wavefrom Audio 10 Menit Dengan Batas Segmen)

9. Emotion Sequence

Setiap segmen audio dianalisis menggunakan CNN 2D *Speech Emotion Recognition* (SER) dan diprediksi ke dalam salah satu dari lima emosi yaitu bahagia, marah, netral, sedih, dan takut. Proses ini menghasilkan urutan emosi (*emotion sequence*) untuk setiap audio berdurasi 10 menit.

10. Permodelan LSTM

Sebelum dilakukan pelatihan model LSTM, data yang digunakan telah melalui proses validasi oleh tenaga ahli (psikolog). Label yang digunakan dalam penelitian ini terdiri dari tiga kategori tingkat stress, yaitu *low*, *medium*, dan *high*. Hasil validasi ini digunakan sebagai dasar dalam pembentukan label pada proses pelatihan model

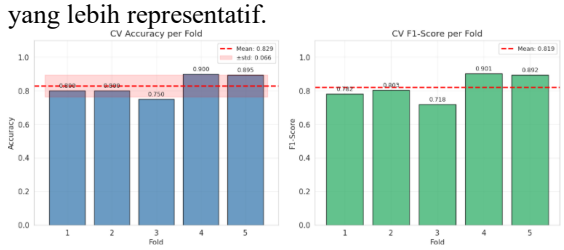


GAMBAR 12
(Grafik perubahan *training loss* dan *validation loss*)

11. Evaluasi Model (K-Fold Cross Validation)

Dalam penelitian ini digunakan nilai k = 5, sehingga dataset dibagi menjadi lima subset.

Proses pelatihan dan pengujian dilakukan sebanyak lima kali dengan kombinasi data yang berbeda pada setiap fold, kemudian hasil evaluasi dirata-ratakan untuk memperoleh performa model yang lebih representatif.

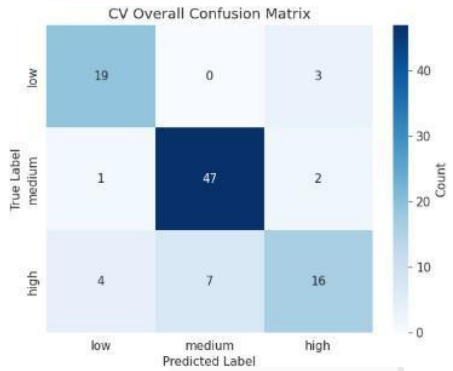


GAMBAR 13
(CV Accuracy dan F1-Score)

Gambar 13 menjelaskan bahwa secara keseluruhan, rata-rata F1-score sebesar 0,819 menunjukkan bahwa model memiliki keseimbangan yang baik antara precision dan recall dalam mendeteksi tingkat stres. Oleh karena itu, model LSTM yang dikembangkan dapat dikatakan memiliki kemampuan generalisasi yang cukup baik dan layak digunakan untuk klasifikasi tingkat stres berbasis suara.

12. Evaluasi Test Set

Setelah model LSTM dilatih dan dievaluasi menggunakan metode *K-Fold Cross Validation*, dilakukan pengujian akhir menggunakan *test set* yang tidak pernah digunakan selama proses pelatihan.



GAMBAR 14
(Confusion Matrix)

Berdasarkan *confusion matrix* pada **Gambar 14**, dapat dilihat bahwa model mampu mengklasifikasikan kelas *low* dan *medium* dengan sangat baik, dimana seluruh data pada kedua kelas tersebut berhasil diprediksi dengan benar. Namun, pada kelas *high* masih terdapat beberapa kesalahan prediksi, dimana sebagian data diklasifikasikan sebagai kelas *low* dan *medium*.

V. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa pendekatan berbasis *Speech Emotion Recognition* (SER) menggunakan kombinasi model CNN 2D dan LSTM mampu mengidentifikasi pola emosi serta mengklasifikasikan tingkat stres dari data suara dengan baik. Tahap pra-pemrosesan, augmentasi, dan ekstraksi fitur menggunakan Log-Mel *Spectrogram* terbukti efektif dalam menghasilkan representasi data yang konsisten dan informatif. Model CNN 2D menunjukkan performa yang sangat tinggi dalam klasifikasi emosi dengan akurasi mencapai 96,86% pada data uji, yang mengindikasikan kemampuannya dalam menangkap karakteristik frekuensi dan temporal sinyal audio. Selanjutnya, pemodelan LSTM yang memanfaatkan urutan emosi (*emotion sequence*) berhasil menganalisis dinamika emosi secara temporal dan mencapai performa yang cukup baik dalam klasifikasi tingkat stres, dengan akurasi validasi mencapai 93,3% serta akurasi test set sebesar 86,7%. Hasil evaluasi menggunakan K-Fold *Cross Validation* menunjukkan bahwa model memiliki kemampuan generalisasi yang cukup stabil dengan rata-rata akurasi sebesar 82,9% dan *F1-score* sebesar 0,819. Secara keseluruhan, penelitian ini membuktikan bahwa pendekatan CNN–LSTM berbasis sekuens emosi efektif untuk mendeteksi tingkat stres dari data suara, meskipun masih terdapat keterbatasan pada distribusi data, khususnya pada kelas stres tinggi.

REFERENSI

[1] ”Menjaga Kesehatan Mental para penerus bangsa.” Diakses: 13 Mei 2025. [Daring]. Tersedia pada: https://kemkes.go.id/id/rilis-kesehatan/menjaga- kesehatan-mental-para-penerus-bangsa

[2] “Mental disorders.” Diakses: 13 Mei 2025. [Daring]. Tersedia pada: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[3] “Menjaga Kesehatan Mental Para Penerus Bangsa.” Diakses: 13 Mei 2025. [Daring]. Tersedia pada: https://kemkes.go.id/id/rilis-kesehatan/menjaga- kesehatan-mental-para-penerus-bangsa

[4] F. N. Rosita, “Prevalensi dan Asosiasi Antara Depresi, Kecemasan, Stres, dan Kualitas Tidur Pada Mahasiswa Selama Pandemi Covid-19,” *PSIKODIMENSIA*, vol. 20, no. 2, hlm. 131–143, Des 2021, doi:10.24167/psidim.v20i2.3507.

[5] W. Agastya dan Aripin, “Pemetaan Emosi Dominan pada Kalimat Majemuk Bahasa Indonesia Menggunakan Multinomial Naïve Bayes,” *J. Nas. Tek. Elektro Dan Teknol. Inf.*, vol. 9, no. 2, hlm. 171–179, Mei 2020, doi: 10.22146/jnteti.v9i2.157.

[6] P. Ekman dan D. Cordaro, “What is Meant by Calling Emotions Basic,” *Emot. Rev.*, vol. 3, no. 4, hlm. 364–370, Okt 2011, doi: 10.1177/1754073911410740.

[7] S.-T. Pan dan H.-J. Wu, “Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with

Data Augmentation," *Electronics*, vol. 12, no. 11, hlm. 2436, Mei 2023, doi: 10.3390/electronics12112436.
[8] E. Tzinis dan A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," dalam *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, TX: IEEE, Okt 2017, hlm. 190–195. doi: 10.1109/ACII.2017.8273599.
[9] G. Airlangga, "Stress Detection Using Hybrid Deep Learning Models with Ensemble Approaches," vol. 6, no. 1, 2024.
[10] M. Maryamah, N. J. K. Pradiptamurty, H. K. Shafro, M. S. Al Qurtubi, G. A. Tambahjong, dan Q. Almaulidiyah, "Speech Emotion Recognition (SER) dengan Metode Bidirectional LSTM," *Pros. Semin. Nas. SAINS DATA*, vol. 3, no. 1, hlm. 153–161, Nov 2023, doi: 10.33005/senada.v3i1.105.
[11] D. Oleh, "Identifikasi Emosi Melalui Suara Menggunakan Support Vector Machine Dan Convolutional Neural Network".
[12] Rendi Nurcahyo dan Mohammad Iqbal, "Pengenalan Emosi Pembicara Menggunakan Convolutional Neural Networks," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 6, no. 1, hlm. 115–122, Feb 2022, doi: 10.29207/resti.v6i1.3726.
[13] G. B. Samanta dan S. Mandala, "Classification Analysis using CNN and LSTM on Wheezing Sounds".
[14] Y. K. Aini, T. B. Santoso, dan T. Dutono, "Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia," *J. Komput. Terap.*, vol. 7, no. 1, hlm. 143–152, Jun 2021, doi: 10.35143/jkt.v7i1.4623.
[15] A. Stevenson, Ed., "Oxford Dictionary of English," Oxford University Press, 2010. Diakses: 11 Mei 2025. [Daring]. Tersedia pada: https://www.oxfordreference.com/display/10.1093/acref/9780199571123.001. 0001/acref-9780199571123
[16] L. A. Sroufe, *Emotional Development: The Organization of Emotional Life in the Early Years*. dalam Cambridge Studies in Social and Emotional Development. Cambridge: Cambridge University Press, 1996. doi: 10.1017/CBO9780511527661.
[17] E. P. Sarafino dan T. W. Smith, *Health Psychology: Biopsychosocial Interactions*. John Wiley & Sons, 2014.
[18] M. El Ayadi, M. S. Kamel, dan F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, hlm. 572–587, Mar 2011, doi: 10.1016/j.patcog.2010.09.020.
[19] K. N. Haque, "What is Convolutional Neural Network — CNN (Deep Learning)," Medium. Diakses: 12 Mei 2025. [Daring]. Tersedia pada: https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn- deep-learning-b3921bdd82d5
[20] S. Sakib, N. Ahmed, A. J. Kabir, dan H. Ahmed, "An Overview of Convolutional Neural Network: Its Architecture and Applications," 20 November 2018, *MATHEMATICS & COMPUTER SCIENCE*. doi: 10.20944/preprints201811.0546.v1.
[21] A. Hibatullah Dan I. Maliki, "Penerapan Metode Convolutional Neural Network Pada Pengenalan Pola Citra Sandi Rumput".
[22] "Apa Itu Long Short Term Memory Network (LSTM)?," Algoritma. Diakses: 13 Mei 2025. [Daring]. Tersedia pada: https://algorit.ma/blog/lstm- network-adalah-2022/
[23] "4. Recurrent Neural Networks - Neural networks and deep learning [Book]." Diakses: 12 Mei 2025. [Daring]. Tersedia pada: https://www.oreilly.com/library/view/neural-networks- and/9781492037354/ch04.html
[24] V. Gallan, "LSTM (Long Short Term Memory)," Bina Nusantara IT Division. Diakses: 13 Mei 2025. [Daring]. Tersedia pada: https://medium.com/bina-nusantara-it-division/lstm-long-short-term-memory-d29779e2ebf8
[25] S. Y. Yusdiantoro dan T. B. Sasongko, "Implementasi Algoritma MFCC dan CNN dalam Klasifikasi Makna Tangisan Bayi," *Indones. J. Comput. Sci.*, vol. 12, no. 4, Agu 2023, doi: 10.33022/ijcs.v12i4.3243.
[26] H. Heriyanto, S. Hartati, Dan A. E. Putra, "Ekstraksi Ciri Mel Frequency Cepstral Coefficient (Mfcc) Dan Rerata Coefficient Untuk Pengecekan Bacaan Al-Qur'an," *Telematika*, vol. 15, no. 2, hlm. 99, Okt 2018, doi: 10.31315/telematika.v15i2.3123.
[27] "pitch - Estimate fundamental frequency of audio signal - MATLAB." Diakses: 18 Mei 2025. [Daring]. Tersedia pada: https://ww2.mathworks.cn/help/audio/ref/pitch.html
[28] L. A. Septiandi, E. M. Yuniarno, dan A. Zaini, "Deteksi Kedipan dengan Metode CNN dan Percentage of Eyelid Closure (PERCLOS)," *J. Tek. ITS*, vol. 10, no. 1, hlm. A56–A63, Agu 2021, doi: 10.12962/j23373539.v10i1.61174.
[29] V. Sareen dan S. K.R., "Speech Emotion Recognition using Mel Spectrogram and Convolutional Neural Networks (CNN)," *Procedia Comput. Sci.*, vol. 258, hlm. 3693–3702, 2025, doi: 10.1016/j.procs.2025.04.624.
[30] J. Zhao, X. Mao, dan L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process.*

Control, vol. 47, hlm. 312–323, Jan 2019, doi: 10.1016/j.bspc.2018.08.035.
[31] F. Makhmudov, A. Kutlimuratov, dan Y.-I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," *Appl. Sci.*, vol. 14, no. 23, hlm. 11342, Des 2024, doi: 10.3390/app142311342.