JURNAL SISTEM INFORMASI DAN TEKNOLOGI
(S I N T E K)

Situs Jurnal
https://sintek.stmikku.ac.id/index.php/home

ANALISIS TEKNOLOGI *SPEECH EMOTION RECOGNITION* (SER): PENDEKATAN FITUR AKUSTIK, KLASIFIKASI, KEAMANAN, DAN IMPLEMENTASI PADA SISTEM PORTABEL

Ahmad Roihan[1], Rainy Zein[2], Maulidya Reva Aprianti[3], Fina Nailatul Izzah[4], Aghnia Luthfunnisa[5]

[12345]Prodi Sistem Komputer, Universitas Raharja
Jl. Jenderal Sudirman No.40, Cikokol, Kec. Tangerang, Kota Tangerang

E-mail: ahmad.roihan@raharja.info[1], rainy@raharja.info[2], maulidya@raharja.info[3], finanailatul@raharja.info[4], aghnia.luthfunnisa@raharja.info[5]

ABSTRAK

Speech Emotion Recognition (SER) merupakan teknologi yang bertujuan mengenali kondisi emosional seseorang berdasarkan sinyal suara. Seiring dengan kemajuan *machine learning* dan *deep learning*, akurasi dan efisiensi sistem SER meningkat melalui pemanfaatan fitur akustik seperti MFCC, GFCC, BFCC, *Cochleagram*, dan *Hilbert Spectrum*. Penelitian ini menggunakan pendekatan studi literatur dengan mengumpulkan data sekunder dari sepuluh artikel ilmiah yang relevan dan terpublikasi dalam jurnal nasional maupun internasional bereputasi. Pemilihan artikel didasarkan pada kesesuaiannya dengan topik *Speech Emotion Recognition* (SER). Analisis dilakukan dengan mengkaji teknik ekstraksi fitur, arsitektur model, serta implementasi pada perangkat keras. Hasil kajian menunjukkan bahwa kombinasi *Convolutional Neural Network* (CNN) dengan *attention mechanism* serta penggunaan *auto-encoder* untuk reduksi dimensi secara signifikan meningkatkan performa klasifikasi emosi. Selain itu, penerapan SER pada perangkat portabel seperti Raspberry Pi memperlihatkan potensi besar untuk pemantauan psikologis berbasis suara secara *real-time*. Namun, aspek keamanan, khususnya dalam menghadapi suara sintetis atau palsu, masih menjadi tantangan utama. Pengembangan SER yang akan datang harus mengintegrasikan aspek akurasi teknis, efisiensi komputasi, keamanan digital, serta mempertimbangkan etika dan privasi pengguna.

Kata Kunci: *Speech Emotion Recognition (SER), MFCC, CNN, deep learning*

1. PENDAHULUAN

Emosi merupakan komponen fundamental dalam komunikasi manusia, tidak hanya disampaikan melalui kata-kata, tetapi juga melalui unsur prosodik seperti intonasi, nada, kecepatan bicara, serta ritme vokal. Kemampuan untuk mengenali emosi dalam interaksi verbal memainkan peran penting dalam membangun pemahaman, empati, dan respon yang sesuai dalam komunikasi interpersonal [1]. Dalam konteks interaksi antara manusia dan mesin, kemampuan ini menjadi salah satu tantangan utama yang hendak dipecahkan oleh teknologi kecerdasan buatan, khususnya melalui pendekatan *Speech Emotion Recognition* (SER).

Speech Emotion Recognition (SER) adalah teknologi yang bertujuan untuk mengenali emosi pengguna berdasarkan karakteristik akustik dari suara mereka. Teknologi ini memungkinkan sistem komputer untuk menginterpretasikan aspek emosional dalam komunikasi lisan secara otomatis. Potensi penerapannya mencakup berbagai bidang, seperti layanan pelanggan yang adaptif secara emosional, sistem pembelajaran daring yang responsif, pemantauan kondisi psikologis dalam bidang kesehatan mental, hingga peningkatan sistem keamanan berbasis biometrik vokal [2][3].

Dalam praktiknya, sistem SER mengandalkan proses ekstraksi fitur suara seperti *Mel-Frequency Cepstral Coefficients* (MFCC), pitch, dan intensitas, yang selanjutnya dianalisis menggunakan algoritma pembelajaran mesin maupun pembelajaran mendalam seperti *K-Nearest Neighbor* (KNN), *Random Forest*, *Convolutional Neural Network* (CNN) [4], dan *Long Short-Term Memory* (LSTM).

Seiring dengan perkembangan teknologi, penelitian terbaru juga menekankan pada peningkatan efisiensi sistem, kemampuan adaptasi terhadap kondisi lingkungan yang bising, serta pentingnya penggunaan *dataset* yang seimbang dan representatif untuk menghindari bias klasifikasi [1][5].

Namun demikian, adopsi teknologi SER juga menghadapi berbagai tantangan non-teknis, khususnya dalam aspek keamanan dan etika. Isu seperti manipulasi sinyal suara yang dapat menyesatkan sistem, hingga risiko pelanggaran privasi data suara pengguna, menjadi perhatian yang semakin besar dalam pengembangan sistem SER yang aman dan dapat dipercaya [5]. Oleh karena itu, pemahaman yang komprehensif terhadap aspek teknis dan non-teknis dalam pengembangan teknologi SER sangat diperlukan untuk mendukung implementasi yang tidak hanya efektif, tetapi juga etis dan aman.

Berdasarkan hal tersebut, artikel ini bertujuan untuk mengkaji berbagai pendekatan yang telah digunakan dalam penelitian *Speech Emotion Recognition*, menilai keunggulan dan keterbatasan masing-masing metode, serta mengevaluasi kesiapan teknologi ini untuk diimplementasikan secara luas dalam berbagai sektor. Kajian ini juga akan menyoroti isu-isu kritis yang berkaitan dengan keamanan, etika, dan privasi yang perlu menjadi perhatian utama dalam pengembangan sistem SER di masa depan.

2. LANDASAN TEORI

2.1. *Speech Emotion Recognition* (SER)

Speech Emotion Recognition (SER) merupakan proses otomatis dalam mengenali kondisi emosional seseorang berdasarkan sinyal suara yang diucapkannya. Sistem ini bertujuan agar mesin mampu mendeteksi dan memahami emosi manusia melalui karakteristik vokal yang bersifat non-verbal. Proses SER umumnya terdiri atas tiga tahap utama: (1) ekstraksi fitur dari sinyal suara, (2) pemrosesan dan klasifikasi menggunakan algoritma kecerdasan buatan, serta (3) interpretasi hasil klasifikasi menjadi kategori emosi tertentu seperti marah, senang, sedih, atau netral [1].

Penerapan SER sangat luas, mulai dari layanan pelanggan berbasis AI, pengawasan psikologis di bidang kesehatan mental, hingga sistem keamanan biometrik suara. Namun, keberhasilan sistem ini sangat bergantung pada kualitas fitur yang diekstraksi dan kemampuan model klasifikasi dalam mengenali pola-pola emosional dalam data suara [2].

2.2. Fitur Akustik dalam SER

Fitur akustik adalah representasi numerik dari sinyal suara yang mengandung informasi penting mengenai karakteristik vokal pembicara [6]. Berbagai teknik ekstraksi fitur telah dikembangkan untuk meningkatkan performa sistem SER, di antaranya:

2.2.1. *Mel-Frequency Cepstral Coefficient* (MFCC)

MFCC adalah salah satu fitur yang paling umum digunakan dalam pengenalan ucapan dan emosi karena kemampuannya dalam meniru cara telinga manusia merespons frekuensi suara. MFCC menekankan frekuensi-frekuensi yang relevan secara psikofisik dan mengabaikan detail frekuensi tinggi yang tidak terlalu penting secara perseptual.

Langkah-langkah utama dalam ekstraksi MFCC meliputi:

- *Pre-emphasis*: Menonjolkan frekuensi tinggi untuk menyeimbangkan spektrum suara.
- *Framing* dan *Windowing*: Membagi sinyal menjadi potongan pendek dan mengurangi distorsi di tepi bingkai.
- *Fourier Transform*: Mengubah sinyal ke domain frekuensi.
- *Mel Filter Bank*: Menyaring frekuensi berdasarkan skala mel.
- Logaritma dan DCT: Menghasilkan koefisien MFCC sebagai fitur utama [4].

2.2.2. *Gammatone-Frequency Cepstral Coefficient* (GFCC)

GFCC dikembangkan berdasarkan prinsip kerja koklea dalam sistem pendengaran manusia. Dengan menggunakan *filterbank gammatone*, GFCC meniru lebih akurat respons biologis telinga terhadap sinyal suara. GFCC dikenal lebih tahan terhadap noise, membuatnya lebih andal dalam lingkungan dengan gangguan suara tinggi [2]. Proses akhirnya mirip dengan MFCC, yaitu menggunakan transformasi DCT untuk menghasilkan koefisien akhir [7].

2.2.3. *Bark-Frequency Cepstral Coefficient* (BFCC)

BFCC juga meniru persepsi pendengaran manusia, namun dengan skala Bark yang dianggap lebih sesuai dalam beberapa konteks, seperti aplikasi dengan keterbatasan komputasi atau analisis sinyal terenkripsi. BFCC menunjukkan performa serupa dengan MFCC, namun menawarkan kelebihan dalam efisiensi pemrosesan dan ketahanan terhadap distorsi tertentu [1].

2.2.4. *Cochleagram dan Hilbert Spectrum*

Dua pendekatan lain dalam representasi sinyal suara adalah *cochleagram* dan *Hilbert spectrum*. *Cochleagram* menghasilkan representasi waktu-

frekuensi yang menyerupai cara koklea memproses suara, dengan filterbank gammatone yang memberikan resolusi frekuensi tinggi pada area penting pendengaran manusia [3].

Sementara itu, *Hilbert Spectrum* diperoleh melalui *Hilbert-Huang Transform* (HHT), yang memungkinkan analisis dinamis dari sinyal nonlinier dan nonstasioner. Kombinasi *cochleagram* dan *Hilbert Spectrum* terbukti efektif dalam menangkap karakteristik emosional suara secara *real-time*, memberikan keunggulan dalam klasifikasi emosi berbasis waktu [5].

2.3. Algoritma Klasifikasi

Setelah fitur diekstraksi dari sinyal suara, proses berikutnya adalah klasifikasi emosi, yang melibatkan sejumlah algoritma machine learning maupun deep learning:

2.3.1. *K-Nearest Neighbor* (KNN)

KNN adalah metode klasifikasi berbasis jarak yang sederhana namun efektif. Algoritma ini mengelompokkan data baru berdasarkan mayoritas label dari sejumlah tetangga terdekatnya dalam ruang fitur. Kelebihannya adalah kemudahan implementasi dan interpretasi, namun performanya menurun jika *dataset* memiliki dimensi tinggi [8].

2.3.2. *Random Forest Classifier* (RFC)

RFC adalah teknik ensemble learning yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi prediksi. RFC bekerja dengan membangun beberapa decision tree dari subset acak data pelatihan dan menggunakan metode voting untuk menentukan hasil klasifikasi. Algoritma ini terkenal stabil dan tahan terhadap overfitting [2].

2.3.3. *Convolutional Neural Network* (CNN)

CNN adalah jenis *deep neural network* yang efektif untuk menganalisis data spasial seperti gambar atau representasi waktu-frekuensi dari sinyal suara (misalnya, *spectrogram* atau *cochleagram*). Dalam SER, CNN digunakan untuk mengenali pola visual dalam representasi spektral suara, memberikan hasil akurasi yang tinggi [3].

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

LSTM merupakan arsitektur dari *Recurrent Neural Network* (RNN) yang dirancang untuk menangani dependensi jangka panjang dalam data sekuensial. Dalam konteks SER, LSTM unggul dalam memahami dinamika temporal suara, sehingga lebih sensitif terhadap perubahan emosi seiring waktu [9][10].

2.3.5. *Auto-Encoder*

Auto-Encoder digunakan sebagai metode reduksi dimensi dan ekstraksi fitur *nonlinier*. Dalam SER, *Auto-Encoder* dapat dilatih untuk mengkodekan sinyal suara ke bentuk representasi yang lebih padat namun tetap informatif. Fitur-fitur ini kemudian dapat digunakan untuk klasifikasi emosi, atau sebagai tahap awal dalam arsitektur *deep learning* lainnya [5].

3. METODOLOGI PENELITIAN

3.1. Metode Pengumpulan Data

Penelitian ini menggunakan pendekatan studi literatur dengan mengumpulkan data sekunder dari sepuluh artikel ilmiah yang relevan dan terpublikasi dalam jurnal nasional maupun internasional bereputasi. Pemilihan artikel didasarkan pada kesesuaiannya dengan topik *Speech Emotion Recognition* (SER), khususnya yang membahas aspek teknis seperti teknik ekstraksi fitur suara, algoritma klasifikasi, performa akurasi, serta isu keamanan sistem.

Adapun kriteria pemilihan artikel yang berfokus pada pengembangan metode SER berbasis *machine learning* atau *deep learning*, dan mencakup pembahasan tentang keunggulan teknis dan tantangan implementasi sistem SER. Sumber-sumber yang digunakan meliputi jurnal yang dapat diakses.

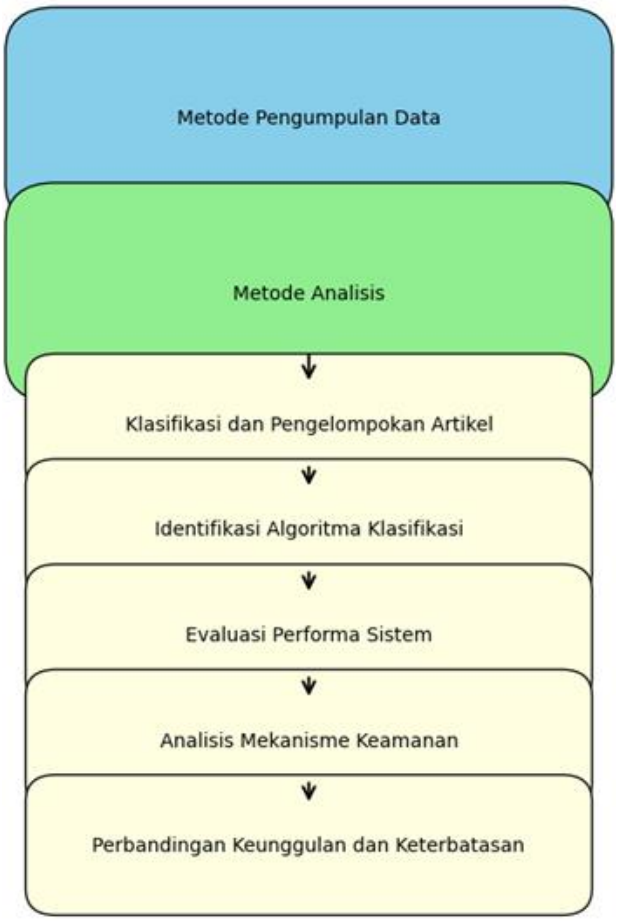
3.2. Metode Analisis

Data yang diperoleh dianalisis secara kualitatif komparatif, dengan tujuan untuk mengidentifikasi dan membandingkan pendekatan teknis dalam sistem SER yang digunakan pada masing-masing studi. Proses analisis dilakukan dengan langkah-langkah sebagai berikut:

- Klasifikasi dan pengelompokan artikel berdasarkan teknik ekstraksi fitur yang digunakan (misalnya MFCC, GFCC, BFCC, *Cochleagram*, dan *Hilbert Spectrum*).
- Identifikasi algoritma klasifikasi yang diterapkan dalam tiap studi, seperti *K-Nearest Neighbor* (KNN), *Random Forest Classifier* (RFC), *Convolutional Neural Network* (CNN), *Long Short-Term Memory* (LSTM), dan *Auto-Encoder*.
- Evaluasi performa sistem, termasuk akurasi, ketahanan terhadap *noise*, dan efisiensi komputasi.
- Analisis mekanisme keamanan, meliputi potensi kerentanan, pendekatan mitigasi risiko manipulasi suara, dan perlindungan privasi data.
- Perbandingan keunggulan dan keterbatasan dari masing-masing pendekatan teknis serta kesesuaian penggunaannya dalam skenario nyata.

Hasil analisis ini diharapkan dapat memberikan pemahaman yang lebih mendalam

mengenai tren dan praktik terbaik dalam pengembangan teknologi SER, serta menjadi landasan dalam pengambilan keputusan untuk implementasi sistem yang aman, akurat, dan efisien.



Gambar 1. Metode Penelitian

4. HASIL DAN PEMBAHASAN

4.1. Perkembangan Teknologi dan Performa Sistem Pengenalan Emosi (SER)

Perkembangan teknologi dalam bidang *Speech Emotion Recognition* (SER) menunjukkan kemajuan signifikan, khususnya dalam peningkatan akurasi, efisiensi, serta kemampuan adaptasi sistem terhadap variabilitas suara manusia. Salah satu inovasi penting yang muncul dalam beberapa tahun terakhir adalah integrasi *Convolutional Neural Network* (CNN) dengan *attention mechanism*. Pendekatan ini memungkinkan model untuk secara dinamis memberi bobot lebih pada segmen-segmen sinyal suara yang paling relevan terhadap ekspresi emosional, sehingga meningkatkan kualitas klasifikasi [3].

Lebih lanjut, kombinasi *cochleagram* dan *Hilbert Spectrum* digunakan untuk memperkaya representasi akustik dari sinyal suara. *Cochleagram* meniru respon koklea manusia terhadap gelombang suara melalui *filterbank gammatone*, sedangkan *Hilbert Spectrum* yang dihasilkan melalui *Hilbert-Huang Transform* menyediakan analisis waktu-frekuensi secara instan terhadap amplitudo dan fase sinyal. Kombinasi keduanya memberi sistem kemampuan untuk menangkap karakteristik emosional secara lebih detail [5].

Dalam upaya meningkatkan efisiensi, *auto-encoder* digunakan sebagai teknik reduksi dimensi. *Auto-encoder* memungkinkan sistem menyederhanakan representasi data masukan tanpa mengorbankan informasi penting. Pendekatan ini tidak hanya mempercepat proses inferensi, tetapi juga mengurangi beban komputasi, menjadikannya sangat berguna untuk aplikasi SER pada perangkat dengan sumber daya terbatas [1].

4.2. Evaluasi Berdasarkan Gender dan Variasi Data

Performa sistem SER tidak hanya ditentukan oleh kompleksitas algoritma, tetapi juga oleh keberagaman dan keseimbangan data yang digunakan dalam pelatihan model. Salah satu temuan penting dalam penelitian ini adalah kecenderungan *overfitting* pada data suara perempuan saat menggunakan kombinasi MFCC dan *K-Nearest Neighbor* (KNN). Hal ini menunjukkan bahwa model terlalu mengandalkan pola spesifik dari subset data tertentu, yang mengurangi kemampuannya dalam melakukan generalisasi terhadap data baru.

Masalah ini mengindikasikan pentingnya penerapan teknik augmentasi data, seperti *pitch shifting*, *time stretching*, dan penambahan *noise* buatan untuk meningkatkan keberagaman data. Selain itu, distribusi data yang seimbang antara jenis kelamin dan ekspresi emosi juga menjadi krusial guna menciptakan sistem yang adil dan inklusif, terutama dalam konteks penggunaan nyata di masyarakat luas.

4.3. Implementasi Sistem Portabel

Implementasi sistem SER pada perangkat portabel seperti Raspberry Pi menunjukkan potensi besar dalam pengembangan teknologi berbasis suara untuk aplikasi sehari-hari. Salah satu pendekatan yang berhasil diterapkan adalah kombinasi GFCC (*Gammatone-Frequency Cepstral Coefficients*) dengan *Random Forest Classifier* (RFC), yang digunakan untuk skrining kondisi psikologis melalui aplikasi berbasis Android. GFCC dipilih karena ketahanannya terhadap *noise* dan kemiripannya dengan mekanisme pendengaran biologis manusia [2].

Pada pengujian lanjutan, penggunaan BFCC (*Bark-Frequency Cepstral Coefficients*) terbukti memberikan hasil yang lebih baik dalam lingkungan

bising, terutama dalam mendeteksi emosi berintensitas rendah seperti kesedihan. BFCC memiliki keunggulan dalam menangkap sinyal dengan rentang frekuensi rendah yang cenderung diabaikan oleh fitur seperti MFCC. Temuan ini menegaskan bahwa sistem SER portabel dapat dioptimalkan untuk bidang seperti pemantauan kesehatan mental, khususnya di wilayah yang memiliki akses terbatas terhadap fasilitas kesehatan konvensional.

4.4. Keamanan Sistem SER

Meskipun sistem SER telah menunjukkan performa teknis yang menjanjikan, aspek keamanan masih menjadi tantangan besar. Hasil pengujian menunjukkan bahwa sebagian besar sistem SER yang tersedia saat ini hanya mampu mendeteksi sekitar 30% dari serangan berbasis suara palsu (*spoofing*). Kondisi ini menunjukkan tingkat kerentanan yang tinggi terhadap manipulasi suara, termasuk suara sintetis hasil generasi AI [5].

Kerentanan tersebut dapat membahayakan integritas sistem, terutama dalam aplikasi yang berkaitan dengan autentikasi suara atau diagnosis kondisi psikologis. Oleh karena itu, diperlukan integrasi sistem SER dengan teknologi keamanan siber, seperti *voice liveness detection*, yang mampu membedakan antara suara asli dan suara hasil rekaman atau sintesis. Selain itu, deteksi suara sintetis berbasis spectral anomalies menjadi salah satu arah riset yang menjanjikan untuk meningkatkan ketahanan sistem terhadap serangan.

Tabel 1. Perbandingan Teknologi dalam Sistem SER

Fitur/Metode	Kelebihan	Kekurangan
CNN + *Attention Mechanism*	Meningkatkan fokus pada bagian penting sinyal → akurasi tinggi	Memerlukan *dataset* besar dan daya komputasi tinggi
Cochleagram + Hilbert	Representasi suara lebih kompleks dan mendalam	Proses ekstraksi berat dan komputasi intensif
Auto-encoder	Reduksi dimensi → klasifikasi lebih cepat	Risiko kehilangan informasi jika tidak dilatih optimal
FCC + KNN	Mudah diterapkan, cocok untuk data sederhana	Cenderung *overfitting*, khususnya pada data suara wanita
GFCC + RFC (Raspberry Pi)	Stabil, cocok untuk perangkat *low-resource*	Akurasi lebih rendah dibanding metode *deep learning*
FCC	Tahan terhadap *noise*, efektif untuk emosi dengan intensitas rendah	Belum umum digunakan, penelitian masih terbatas

5. KESIMPULAN

Secara keseluruhan, perkembangan teknologi *Speech Emotion Recognition* (SER) menunjukkan kemajuan signifikan dari segi akurasi, efisiensi, dan adaptabilitas, terutama dengan penerapan CNN bermekanisme perhatian, spektrum *cochleagram-Hilbert*, serta *auto-encoder* dalam perangkat portabel seperti Raspberry Pi. Meski demikian, tantangan serius masih ditemukan dalam aspek keamanan, khususnya terhadap serangan suara sintetis (*spoofing*), serta isu etika dan privasi data pengguna yang rentan terekspos pada penerapan di bidang kesehatan dan pemantauan emosional. Oleh karena itu, pengembangan SER ke depan perlu dilakukan secara holistik dengan membangun *dataset* yang seimbang dan representatif, mengembangkan sistem pendeteksi suara palsu yang terintegrasi dengan autentikasi emosional, serta mengedepankan prinsip etika dan perlindungan data. Pengujian di lingkungan nyata dan kolaborasi lintas disiplin juga mutlak diperlukan guna memastikan sistem yang dihasilkan tidak hanya unggul secara teknis, tetapi juga aman, adil, dan bermanfaat secara sosial.

DAFTAR PUSTAKA

[1] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning—A systematic review," *Intelligent Systems with Applications*, vol. 20, p. 200266, 2023.

[2] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech emotion recognition through hybrid features and convolutional neural network," *Applied Sciences*, vol. 13, no. 8, p. 4750, 2023.

[3] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech emotion recognition using convolutional neural networks with attention mechanism," *Electronics*, vol. 12, no. 20, p. 4376, 2023.

[4] D. Rafiqo, Y. Suyanto, and C. Atmaji, "Klasifikasi Suara Paru-Paru Berdasarkan Ciri MFCC," *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, vol. 12, no. 1, pp. 1–12, 2022.

[5] I. Gurowiec and N. Nissim, "Speech emotion recognition systems and their security aspects," *Artificial Intelligence Review*, vol. 57, no. 6, p. 148, 2024.

[6] M. B. Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features," in *IEEE Access*, vol. 8, pp. 221640-221653, 2020, doi: 10.1109/ACCESS.2020.3043201

[7] B. H. Prasetio, L. O. A. Hazmar, D. Syauqy, and E. R. Widasari, "Gammatone-Frequency Cepstral Coefficients Based Fear Emotion Level Recognition System," *Revista Mexicana de Ingeniería Biomédica*, vol. 45, no. 2, pp. 6–22, 2024.

[8] A. A. Yusuf, S. K. Wijaya, and P. Prajitno, "EEG-based human emotion recognition using k-NN machine learning," in *AIP Conference Proceedings*, vol. 2168, no. 1, AIP Publishing, Nov. 2019.

[9] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence LSTM architecture," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6474–6478.

[10] M. N. Dar, M. U. Akram, S. G. Khawaja, and A. N. Pujari, "CNN and LSTM-based emotion charting using physiological signals," *Sensors*, vol. 20, no. 16, p. 4551, 2020.