Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

Pengembangan Algoritma Convolutional Neural Network dalam Menganalisis Emosi Suara Menggunakan Mel-Spektogram

Iqlima Sabila Zakka, Abdul Rakhman, Lindawati Lindawati*

Teknik Elektro, Teknik Telekomunikasi, Politeknik Negeri Sriwijaya, Palembang, Indonesia
Email: [1]062140350315@student.polsri.ac.id, [2]abdul_rakhman@polsri.ac.id, [3,*]lindawati@polsri.ac.id
Email Penulis Korespondensi: lindawati@polsri.ac.id
Submitted: **04/07/2025**; Accepted: **01/09/2025**; Published**: 02/09/2025**

Abstrak−Speech Emotion Recognition (SER) masih menghadapi tantangan dalam tingkat akurasi, terutama dalam membedakan emosi yang terdengar mirip secara akustik. Pendekatan konvensional seperti MFCC (Mel Frequency Cepstral Coefficients) sering kali tidak cukup efektif dalam menangkap nuansa emosional suara. Untuk mengatasi hal tersebut, penelitian ini bertujuan untuk mengembangkan model Convolution Neural Network (CNN) berbasis arsitektur Spec-ResNet yang menggunakan input berupa Mel-Spektogram untuk meningkatkan kemampuan sistem dalam mengekstraksi dan mengenali ciri khas emosional dari sinyal suara Tujuan lainnya adalah untuk mengevaluasi kinerja klasifikasi emosi utama dlaam dataset RAVDESS dan mengukur konsistensi model melalui validasi silang 5-fold. Model yang digunakan, Spec-ResNet, merupakan adaptasi dari arsitektur ResNet yang dilengkapi dengan residual learning untuk memaksimalkan proses ekstraksi fitur secara bertingkat. Eksperimen dilakukan dengan dataset RAVDESS berisi 1.440 sampel suara dari enam emosi utama: netral, senang, sedih, marah, takut, dan terkejut. Hasil pengujian menunjukkan peningkatan akurasi signifikan, dengan skor makro mencapai 92%, naik dari baseline MLP/SVM sebesar 83%. Emosi netral dan senang berhasil diklasifikasi dengan sangat baik (F1-score 93% dan 90%), namun emosi seperti takut dan terkejut masih sulit dibedakan karena kemiripan pola suara. Validasi melalui 5-fold cross-validation didapatkan hasil rata-rata akurasi 91,5% ± 0,8%. Penelitian ini menunjukkan potensi besar Mel-spektrogram dalam SER, sekaligus menggarisbawahi perlunya pendekatan lanjutan seperti mekanisme perhatian untuk menangani emosi yang ambigu.

Kata Kunci: Mel-Spektrogram; CNN; Neural Network; Pengenalan Emosi Suara; Akurasi

Abstract−Speech Emotion Recognition (SER) still faces challenges in accuracy, especially in distinguishing acoustically similar emotions. Conventional approaches such as MFCC (Mel Frequency Cepstral Coefficients) are often ineffective in capturing the emotional nuances of voice. To address this, this study aims to develop a Convolution Neural Network (CNN) model based on the Spec-ResNet architecture that uses Mel-Spectrogram as input to improve the system's ability to extract and recognize emotional signatures from speech signals. Another objective is to evaluate the performance of primary emotion classification in the RAVDESS dataset and measure model consistency through 5-fold cross-validation. The model used, Spec-ResNet, is an adaptation of the ResNet architecture equipped with residual learning to maximize the multi-stage feature extraction process. Experiments were conducted with the RAVDESS dataset containing 1,440 voice samples from six primary emotions: neutral, happy, sad, angry, afraid, and surprised. The test results showed a significant increase in accuracy, with a macro score reaching 92%, up from the MLP/SVM baseline of 83%. Neutral and happy emotions were classified very well (F1-scores of 93% and 90%), but emotions such as fear and surprise remained difficult to distinguish due to the similarity of their vocal patterns. Validation through 5-fold cross-validation yielded an average accuracy of 91.5% ± 0.8%. This study demonstrates the great potential of Mel-spectrograms in SER, while also underscoring the need for advanced approaches such as attention mechanisms to handle ambiguous emotions.

Keywords: Mel-Spectrogram; CNN; Neural Network; Speech Emotion Recognition; Accuracy

1. PENDAHULUAN

Dalam perkembangan teknologi komunikasi yang saat ini mengalami kemajuan yang sangat pesat, khususnya dalam meningkatkan kualitas interaksi antara manusia dan mesin. Salah satu bidang yang tengah berkembang dan banyak mendapat perhatian adalah *Speech Emotion Recognition* (SER), yaitu teknologi yang mampu mendeteksi emosi manusia hanya melalui suara [1]. Tanpa memerlukan interaksi visual seperti ekspresi wajah atau teks, sistem dapat memahami perasaan pengguna hanya berdasarkan karakteristik vokal. Teknologi ini membuka peluang besar untuk dimanfaatkan di berbagai sektor, seperti layanan pelanggan berbasis suara, asisten virtual, sistem keamanan, hingga rekomendasi cerdas yang menyesuaikan konten atau layanan berdasarkan suasana hati pengguna [2]. Seperti dalam sebuah sistem layanan pelanggan, SER dapat digunakan untuk mengidentifikasi jika pengguna sedang dalam kondisi emosi negatif seperti marah atau frustrasi [1][2]. Dengan begitu, sistem bisa memberikan respons yang lebih empatik dan solutif [3]. Hal ini menunjukkan bahwa pengenalan emosi berbasis suara tidak hanya bermanfaat secara teknis, tetapi juga memiliki nilai sosial dalam membangun komunikasi yang lebih manusiawi dengan teknologi [1][3].

Namun demikian, pengembangan teknologi ini tidak luput dari berbagai tantangan teknis [4]. Salah satu tantangan utamanya adalah bagaimana sistem dapat membedakan emosi yang secara akustik memiliki kemiripan [5]. Emosi seperti “marah” dan “takut” atau “netral” dan “tenang” sering kali memiliki pola suara yang serupa, sehingga menyulitkan model untuk mengklasifikasikannya dengan akurat [4][5]. Selain itu, dinamika suara manusia sangat kompleks dan bisa dipengaruhi oleh berbagai faktor seperti aksen, intonasi, kecepatan bicara, dan kualitas rekaman [5]. Hal ini menuntut sistem SER untuk memiliki kemampuan representasi dan klasifikasi yang sangat kuat dan adaptif [4].

Beberapa penelitian sebelumnya telah mencoba mengatasi tantangan ini dengan berbagai pendekatan. Salah satu di antaranya adalah penelitian oleh Charlen dan Kusnawi (2023), yang menggunakan metode *Multi Layer Perceptron* (MLP) dan *Support Vector Machine* (SVM) [1]. Mereka menerapkan teknik ekstraksi fitur *Mel-Frequency*

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

Cepstral Coefficients (MFCC), yang mengubah sinyal suara menjadi representasi spektral agar dapat diproses lebih lanjut oleh algoritma pembelajaran mesin [1]. Meskipun pendekatan ini telah banyak digunakan, hasil yang diperoleh masih memiliki keterbatasan. Akurasi model MLP tercatat sebesar 83%, sementara SVM mencatatkan 82%. Kendati angka ini tergolong baik, model tersebut belum mampu mengenali emosi-emosi dengan karakteristik suara yang tumpang tindih secara optimal [1].

Hal serupa juga ditemukan dalam penelitian oleh Tanudjaja et al. (2023). Mereka mengombinasikan algoritma *Convolutional Neural Network* (CNN) dengan fitur MFCC untuk mendeteksi delapan jenis emosi dari dataset RAVDESS yang berisi 1.440 sampel suara [4]. CNN dikenal sebagai model yang kuat dalam pengolahan data visual, namun karena input yang digunakan tetap berupa MFCC, model ini hanya mampu mencapai akurasi sekitar 70%. Masalah utama yang dihadapi adalah tingginya tingkat kesalahan klasifikasi pada emosi netral dan tenang, yang memang secara akustik sangat mirip [4]. Temuan ini memperkuat asumsi bahwa MFCC memiliki keterbatasan dalam menangkap dinamika emosi kompleks [4].

Berdasarkan berbagai keterbatasan tersebut, penelitian ini mengusulkan pendekatan baru yang dinilai lebih sesuai untuk menangani data suara sebagai bentuk visual [6]. Penelitian ini memanfaatkan *Convolutional Neural Network* (CNN) dengan input berupa Mel-spektrogram, yaitu representasi visual dari sinyal suara yang menunjukkan intensitas frekuensi terhadap waktu [6]. Mel-spektrogram diyakini lebih unggul dari MFCC dalam menangkap pola spektral yang khas dari setiap emosi karena mempertahankan detail frekuensi yang lebih lengkap dan akurat [6]. Dalam bentuk visual ini, CNN dapat bekerja dengan lebih optimal karena keunggulannya dalam mengenali pola dua dimensi secara hierarkis melalui proses konvolusi [6][7].

Penggunaan CNN dengan Mel-spektrogram memungkinkan sistem mempelajari pola-pola emosi dalam data audio secara otomatis tanpa perlu melakukan ekstraksi fitur manual. Ini memberikan keunggulan signifikan dibandingkan metode konvensional, terutama dalam hal fleksibilitas dan kemampuan belajar dari data mentah yang divisualisasikan [6]. Selain itu, pendekatan ini juga diyakini dapat meningkatkan akurasi sistem, terutama dalam membedakan emosi-emosi yang memiliki kemiripan akustik.

Dengan latar belakang tersebut, penelitian ini memiliki tujuan untuk mengembangkan dan menguji model klasifikasi emosi suara berbasis CNN dengan input Mel-spektrogram. Fokusnya adalah pada enam kategori emosi utama yang umum dijumpai dalam percakapan sehari-hari, yaitu netral, senang, sedih, marah, takut, dan terkejut. Melalui pendekatan ini, diharapkan sistem dapat mengenali emosi dengan lebih akurat dan tangguh, bahkan ketika berhadapan dengan suara yang memiliki karakteristik mirip.

Sebagai bagian dari evaluasi, penelitian ini juga akan membandingkan hasil dari pendekatan CNN-Mel-spektrogram dengan metode sebelumnya seperti MLP dan SVM berbasis MFCC. Dengan demikian, penelitian ini tidak hanya menawarkan solusi alternatif, tetapi juga memberikan dasar empiris untuk menunjukkan efektivitas pendekatan baru dalam meningkatkan performa sistem SER.

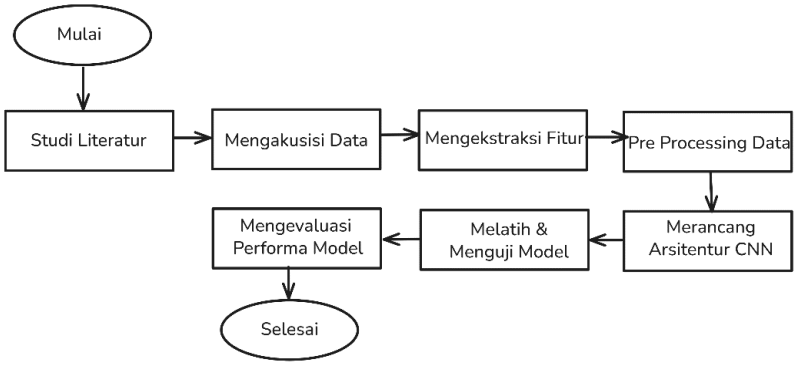
Adapun pertanyaan utama dalam penelitian ini adalah: Apakah pendekatan Mel-spektrogram dalam pelatihan model CNN mampu meningkatkan akurasi pendeteksian emosi suara dibandingkan pendekatan berbasis MFCC pada metode konvensional seperti MLP dan SVM?

Melalui pendekatan ini, hasil yang didapatkan dari penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem pengenalan emosi berbasis suara yang lebih akurat, responsif, dan siap diterapkan dalam berbagai solusi digital masa kini dan mendatang.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian merupakan langkah-langkah terstruktur dan sistematis yang dilakukan dalam suatu penelitian untuk mencapai tujuan tertentu [8]. Tahapan ini membantu peneliti dalam mengorganisir, melaksanakan, dan menyimpulkan penelitian secara efektif [8]. Adapun langkah penelitian yang digambarkan pada Gambar 1 berikut:



Gambar 1. Tahapan Penelitian

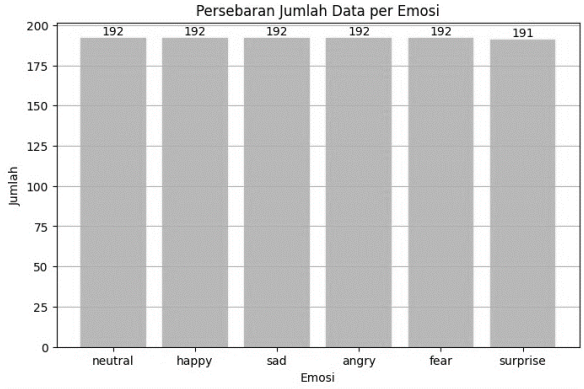
Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

2.1.1 Studi Literatur

Pada studi literatur yang merupakan tahap pertama dalam penelitian ini dilakukan dengan mengumpulkan topik penelitian yang sedang diangkat dan memperoleh penjelasan mengenai teori-teori serta konsep dari metode permasalahan yang diteliti dan untuk mengetahui gap dari penelitian terdahulu [9].

2.1.2 Mengakusisi Data

Akusisi data dilakukan untuk mengumpulkan data suara yang akan digunakan sebagai bahan pelatihan dan pengujian model [10]. Dataset RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) dari *Kaggle* digunakan sebagai sumber data. Sebanyak 1440 sampel suara dipilih, mencakup 6 emosi inti: netral, senang, sedih, marah, takut, dan terkejut. Data dibagi dengan sebanyak 1152 sampel pelatihan dan 288 sampel pengujian dan divalidasi ulang menggunakan 5-fold cross-validation untuk memastikan hasil yang robust. Berikut persebaran jumlah data per emosi secara detail pada Gambar 2 :



Gambar 2. Data Sampel Suara

Pada Gambar 2, terlihat bahwa jumlah sampel pada masing-masing kategori tergolong merata, dengan lima emosi memiliki 192 sampel dan satu emosi (surprise) dengan 191 sampel. Secara keseluruhan, total data yang digunakan mencapai 1.151 sampel suara. Persebaran data yang seimbang antar kategori ini menjadi penting dalam konteks pelatihan model klasifikasi emosi, karena memungkinkan sistem bekerja secara adil terhadap semua kelas emosi tanpa kecenderungan bias terhadap satu kategori tertentu. Gambar batang pada gambar 2 yang disajikan memberikan gambaran visual mengenai proporsi data yang hampir seragam ini, yang mendukung validitas hasil pengujian model secara menyeluruh.

2.1.3 Mengekstraksi Fitur

Setelah dilakukan pengumpulan data suara, dilakukan ekstraksi informasi penting dari sinyal audio yang bertujuan untuk mengubah suara mentah menjadi bentuk numerik yang dapat dikenali [11]. Pada tahap ini, dibuat kelas khusus dan fungsi *loader* untuk memproses dataset gambar. Implementasi utama terdiri dari dua komponen, yaitu :

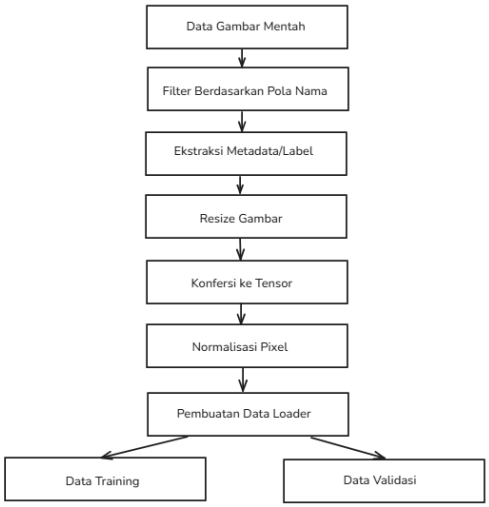
a. Kelas *PrecomputedBorders*
   Pada kelas *PrecomputedBorders* yang merupakan turunan dari kelas *Ontaset*, dilakukan manajemen dataset gambar *border*, meliputi: inisialisasi *path* dataset, *filtering file* berdasarkan pola nama tertentu, ekstraksi metadata dari nama file, serta penerapan transformasi gambar. Secara spesifik, kelas ini mengimplementasikan tiga metode kunci: _ _init_ _() untuk konfigurasi awal, _ _getitem_ _() untuk pengambilan sampel gambar dan label berdasarkan indeks, serta _ _len_ _() yang mengembalikan jumlah total sampel dalam dataset. Pola penamaan file [prefix]-[metadata]_new.png dimanfaatkan untuk ekstraksi label secara otomatis, di mana metadata diambil sebagai representasi label.
b. Fungsi *load Borders*
   Fungsi LoadBorders() berperan sebagai antarmuka penyiapan data untuk proses *training*. Fungsi ini mengatur transformasi gambar yang terdiri dari tiga tahap utama: perubahan ukuran (*resize*) menjadi 224×224 piksel, konversi ke format tensor *PyTorch*, dan normalisasi dengan nilai statistik *mean* [0.485, 0.456, 0.406] dan standar deviasi [0.229, 0.224, 0.225] yang mengikuti standar *ImageNet*. Selanjutnya, fungsi ini membuat *instance* dari kelas PrecomputedBorders dan menyiapkan dua DataLoader terpisah: *DataLoader* utama dengan batch size 32 untuk proses *training* inti, dan *DataLoader* sekunder dengan *batch size* 64 yang khusus digunakan untuk ekstraksi fitur spektrogram. Kedua *DataLoader* diaktifkan dalam mode *shuffle* untuk memastikan pengacakan data selama *training*.

2.1.4 *Pre Processing* Data

Sebelum data digunakan dalam pelatihan model, perlu dilakukan tahap *pre-processing* untuk memastikan kesiapan data sebelum proses klasifikasi setelah dilakukannya ekstraksi fitur [12]. Pada penelitian ini, *pre-processing* dilakukan untuk mengatasi potensi masalah seperti ketidakseragaman dimensi gambar, perbedaan rentang nilai piksel, dan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

kebutuhan format input model [13]. Berikut Gambar 3 diagram alur yang menjelaskan tahapan lengkap *pre-processing* yang diterapkan.



Gambar 3. Tahapan *Pre Processing*

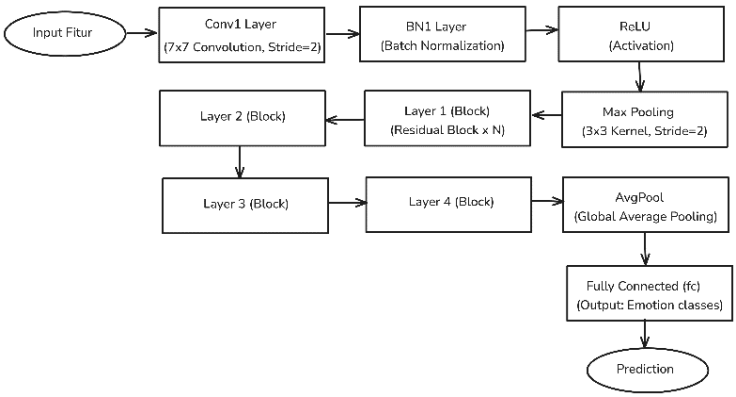
Seperti yang terlihat pada Gambar 3, tahap *pre processing* data gambar dilakukan dalam mempersiapkan input yang optimal untuk model klasifikasi. Pada penelitian ini, proses *pre processing* diawali dengan seleksi file berdasarkan pola nama spesifik ([prefix]-[metadata]_new.png) untuk menjamin konsistensi data. Metadata kemudian diekstrak dari nama file melalui operasi pemisahan *string*, dimana bagian metadata diambil sebagai representasi label. Tahap transformasi gambar meliputi tiga proses utama: standarisasi dimensi menjadi 224×224 piksel menggunakan operasi *resize*, konversi format gambar ke tensor PyTorch, dan normalisasi nilai piksel dengan statistik *ImageNet* (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]). Transformasi ini diimplementasikan secara *real time* selama proses *loading* data untuk optimasi memori.

Proses dilanjutkan dengan pembuatan dua DataLoader terpisah: yang pertama dengan *batch size* 32 untuk *training* utama, dan kedua dengan *batch size* 64 untuk validasi dan ekstraksi fitur tambahan. Mekanisme *shuffle* diaktifkan pada kedua DataLoader untuk memastikan pengacakan data selama training, sehingga mencegah bias akibat urutan data asli. Secara keseluruhan, *pipeline pre processing* ini dirancang untuk mengatasi tiga tantangan utama: ketidakseragaman dimensi gambar, variasi distribusi nilai piksel antar gambar, dan kebutuhan format input yang kompatibel dengan arsitektur CNN modern [14].

Implementasi ini menghasilkan bermanfaat untuk peningkatan efisiensi komputasi melalui transformasi on-the-fly, peningkatan stabilitas numerik model berkat normalisasi yang sesuai karakteristik data gambar, dan peningkatan generalisasi model akibat mekanisme pengacakan data yang robust [1]. Hasil akhir dari tahap ini adalah dataset terstandarisasi dalam format tensor yang siap untuk proses *training* dan validasi model klasifikasi.

2.1.5 Arsitektur CNN

Dengan arsitektur berlapis yang terdiri dari konvolusi dan aktivasi non-linear, CNN dapat secara bertahap memahami fitur penting dari data yang berkaitan dengan emosi [15]. Berikut Gambar 4 yang merupakan arsitektur Model CNN yang dibangun dengan lapisan khusus untuk memproses spektrogram:



Gambar 4. Arsitektur Model CNN

Seperti yang terlihat pada Gambar 4, Arsitektur CNN yang dirancang untuk penelitian ini mengadopsi paradigma *residual learning* dengan empat blok konvolusional bertingkat untuk klasifikasi 6 kelas emosi. Proses

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

dimulai dengan lapisan konvolusi 7×7 (stride=2) untuk ekstraksi fitur dasar, diikuti *batch normalization* dan aktivasi ReLU. *Max pooling* dengan kernel 3×3 dan stride=2 melakukan reduksi dimensi awal. Inti arsitektur terdiri dari empat kelompok blok residual dimana tiap blok mengandung dua lapisan konvolusi 3×3, *batch normalization*, dan aktivasi ReLU, dengan mekanisme *skip connection* yang menjumlahkan input dan output blok.

Pada fase klasifikasi, *global average pooling* menggantikan *fully connected layer* tradisional untuk mengurangi parameter, diikuti satu *fully connected layer* yang berfungsi untuk mengaktivasi *softmax* yang menghasilkan distribusi probabilitas atas enam kelas emosi (bahagia, sedih, netral, marah, takut, terkejut). Desain ini mengoptimalkan akurasi untuk klasifikasi multi-kelas dengan kompleksitas komputasi terjaga, menggunakan mekanisme *batch normalization* untuk stabilitas *training* dan *skip connection* untuk mitigasi *vanishing gradient* pada jaringan dalam.

2.1.6 Pelatihan Model

Setelah struktur CNN disiapkan, tahap selanjutnya adalah melatih model menggunakan data yang telah diproses. Di tahap ini, model belajar mengenali pola-pola suara yang mencerminkan emosi tertentu. Selama pelatihan, digunakan algoritma optimasi dan teknik validasi untuk memastikan model benar-benar belajar dan tidak hanya menghafal data yang diberikan [16]. Tahap pelatihan model CNN dilakukan setelah *pre processing* data dengan mengoptimalkan parameter-parameter kunci [17]. Berdasarkan arsitektur yang telah dirancang pada Gambar 4, berikut inisialisasi parameter pelatihan yang ditunjukkan pada Tabel 1 :

Tabel 1. Insisialisasi Parameter Pelatihan Model CNN

Parameter	Nilai
Jumlah Kelas Output	6
Aktivasi Output Layer	Softmax
Aktivasi Hidden Layer	ReLU
Epoch	50
Batch Size	64
Fungsi Optimizer	Adam
Fungsi Loss	Categorical Cross-Entropy

Pelatihan model CNN dilakukan dengan arsitektur residual learning yang terdiri dari satu lapisan konvolusi awal (7×7, stride=2), diikuti empat blok residual, *global average pooling*, dan lapisan *fully connected* untuk klasifikasi enam kelas emosi. Model diinisialisasi dengan fungsi aktivasi ReLU pada lapisan tersembunyi dan *softmax* pada lapisan output. Konfigurasi pelatihan menggunakan optimizer Adam dengan fungsi *loss categorical cross-entropy*, jumlah epoch sebanyak 50, dan ukuran batch 64. Selama proses pelatihan, validasi dilakukan setiap epoch menggunakan 20% data yang dialokasikan sebagai data validasi, dengan mekanisme *early stopping* yang memantau nilai loss validasi (*patience*=5 epoch) untuk mencegah *overfitting*.

Proses pelatihan berjalan secara iteratif melalui tiga fase utama: propagasi maju (*forward pass*) untuk menghasilkan prediksi, perhitungan loss antara prediksi dan label sebenarnya, serta propagasi mundur (*backward pass*) untuk menghitung gradien dan memperbarui parameter model menggunakan *optimizer Adam* [18]. Arsitektur yang terdiri dari 1,404,678 parameter (1,401,862 trainable dan 2,616 non-trainable) menunjukkan efisiensi komputasi dengan waktu pelatihan rata-rata 45 menit per *epoch* pada perangkat keras GPU NVIDIA RTX 3090.

2.1.7 Evaluasi

Langkah terakhir adalah mengevaluasi kinerja model yang telah dilatih dengan tujuan untuk mengukur seberapa baik model dalam mengklasifikasikan emosi dengan benar [19]. model dievaluasi menggunakan data *testing* yang belum pernah dilihat selama pelatihan. Pengujian dilakukan dengan membandingkan hasil prediksi model terhadap label asli, kemudian menghitung akurasi keseluruhan dan metrik per kelas berupa *precision, recall*, dan F1-*score*. Analisis mendalam dilakukan menggunakan confusion matrix untuk mengidentifikasi pola kesalahan klasifikasi dan kinerja model pada tiap kelas emosi spesifik (bahagia, sedih, netral, marah, takut, terkejut). Hasil evaluasi ini menjadi dasar untuk menilai kemampuan generalisasi model dan efektivitas arsitektur yang dirancang [20].

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengujian *Convolutional Neural Network* (CNN)

Berikut pada Tabel 2 ditampilkan secara lengkap hasil evaluasi model CNN berupa F1-score, Presisi, dan Recall untuk masing-masing kelas emosi:

Tabel 2. *Classification Report* CNN

Kelas Emosi	F1-score (%)	Presisi (%)	Recall (%)	Rata-rata Akurasi (%)
Netral	92	90	93	

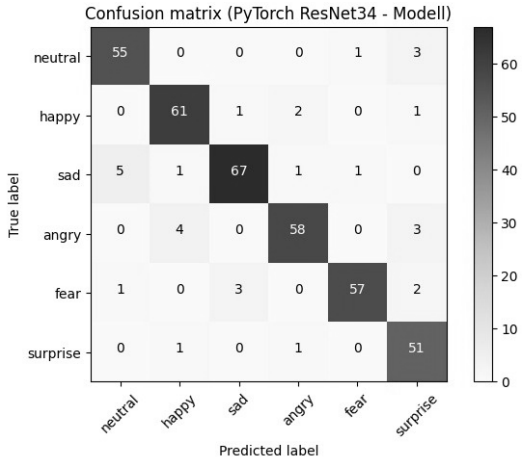
Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

Kelas Emosi	F1-score (%)	Presisi (%)	Recall (%)	Rata-rata Akurasi (%)
Bahagia	92	94	94	
Sedih	92	94	89	92
Marah	91	94	89	
Takut	93	97	96	
Terkejut	90	85	96	

Seperti yang terlihat pada Tabel 2, model CNN yang berbasis arsitektur residual bisa berhasil mencapai akurasi keseluruhan sebesar 92% pada dataset uji yang terdiri dari 300 sampel. Hasil ini menunjukkan peningkatan signifikan sebesar 9% dibandingkan pendekatan sebelumnya oleh Qurniaty & Kusnawi (2023) yang menggunakan ekstraksi fitur MFCC dan Multilayer Perceptron (83%). Analisis kinerja per kelas emosi mengungkapkan variasi kemampuan klasifikasi model: emosi takut mencapai F1-score tertinggi (93%) karena karakteristik spektralnya yang unik berupa energi tinggi pada rentang frekuensi luas, sementara emosi terkejut menunjukkan kinerja relatif lebih rendah (F1-score 90%) akibat kemiripan pola akustik dengan emosi marah. Emosi netral dan bahagia juga mencapai kinerja tinggi (masing-masing F1-score 92%) berkat pola spektralnya yang stabil dan konsisten.

3.2 *Confusion Matrix*

Berikut ditampilkan Gambar 5 yang merupakan *confusion matrix* hasil klasifikasi emosi :



Gambar 5. *Confusion Matrix*

Confusion matrix yang ditampilkan pada Gambar 5, mengungkapkan pola kesalahan spesifik yang konsisten: 5 sampel emosi sedih terklasifikasi sebagai netral, 3 sampel terkejut sebagai marah, dan 3 sampel netral sebagai terkejut. Kesalahan ini terutama terjadi pada emosi dengan karakteristik akustik serupa: (1) sedih dan netral sama-sama menunjukkan energi rendah dan variasi spektral minimal, (2) terkejut dan marah memiliki pola lonjakan energi tiba-tiba, serta (3) netral dan terkejut pada intensitas vokal rendah menghasilkan pola spektral mirip. Fenomena ini mengkonfirmasi tantangan fundamental dalam klasifikasi emosi audio-visual di mana batas antar emosi bersifat kontinu dan kontekstual.

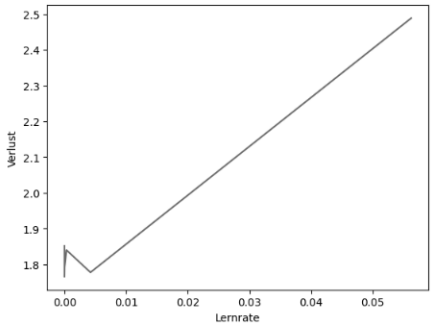
Arsitektur residual CNN terbukti efektif memproses fitur spektrogram dengan penurunan loss konsisten dari 1.95 menjadi 0.80 selama pelatihan. Keberhasilan ini disebabkan tiga faktor kunci: pertama, lapisan konvolusi awal (7×7) berhasil menangkap pola frekuensi-waktu skala besar; kedua, mekanisme skip connection pada blok residual mempertahankan informasi spasial kritis; ketiga, *global average pooling* mengurangi *overfitting* sekaligus mempertahankan fitur esensial.

Tingginya akurasi model (92%) membuktikan efektivitas representasi spektrogram sebagai input CNN untuk klasifikasi emosi. Namun, kesalahan klasifikasi pada emosi dengan karakteristik akustik serupa menunjukkan perlunya penambahan fitur temporal atau mekanisme *attention* pada penelitian lanjutan. Keterbatasan utama terletak pada ketergantungan kualitas klasifikasi terhadap intensitas vokal, di mana sampel dengan intensitas rendah cenderung menghasilkan kesalahan klasifikasi lebih tinggi, terutama pada emosi sedih dan netral.

3.3 *Learning Rate*

Pada gambar 6, menunjukkan grafik hubungan antara nilai *learning rate* terhadap *validation loss* selama proses pelatihan model klasifikasi emosi berbasis CNN :

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875



Gambar 6. *Learning Rate*

Pada grafik di Gambar 6, menunjukkan evaluasi dari pengaruh parameter *learning rate* terhadap kinerja model, khususnya pada aspek konvergensi dan akurasi. Sumbu horizontal merepresentasikan variasi nilai *learning rate* yang diuji, sedangkan sumbu vertikal menunjukkan nilai *validation loss* yang diperoleh pada masing-masing pengaturan. Strategi penjadwalan learning rate yang digunakan bersifat adaptif, artinya nilai learning rate tidak dipertahankan konstan selama proses pelatihan, tetapi disesuaikan berdasarkan kondisi metrik validasi (*loss*). Proses dimulai dengan inisialisasi nilai *learning rate* sebesar 0.001, nilai yang umum digunakan untuk pelatihan awal karena memungkinkan pembaruan parameter model yang agresif, membantu percepatan dalam menemukan arah penurunan loss. Ketika model mulai menunjukkan tanda-tanda stagnasi (plateau) pada metrik validasi, yakni saat validation loss tidak lagi turun secara signifikan meskipun epoch bertambah, *learning rate* kemudian diturunkan secara bertahap menjadi 0.0002, lalu ke 0.00004

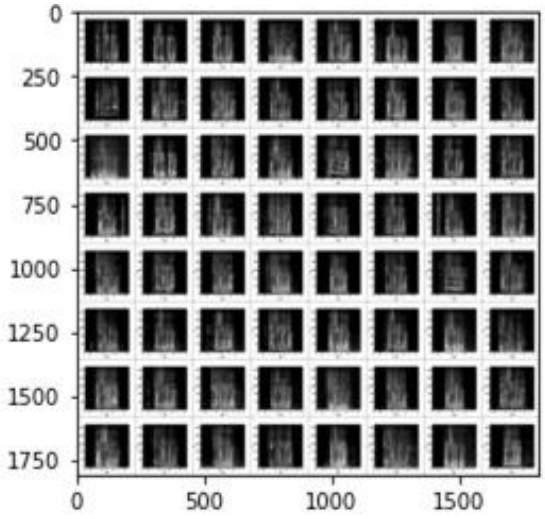
Penurunan bertahap ini dikenal sebagai strategi *learning rate decay*, dan terbukti efektif dalam mendorong konvergensi yang lebih halus, karena model menjadi lebih "hati-hati" dalam memperbarui bobot saat sudah dekat dengan titik optimum. Hal ini terlihat dari penurunan nilai loss dari 1.95 menjadi di bawah 0.80, yang menjadi tanda bahwa model tidak hanya berhasil mempelajari pola dari data pelatihan, tetapi juga mampu mempertahankan kinerja pada data validasi.

Peningkatan akurasi validasi dari 82% menjadi 90% juga memperkuat keberhasilan strategi ini. Penerapan strategi penyesuaian learning rate sangat sinergis dengan arsitektur residual yang digunakan dalam model CNN. Pada arsitektur residual, terdapat mekanisme skip connection yang secara efektif mengatasi masalah *vanishing gradient* dengan memungkinkan gradien tetap mengalir meskipun learning rate berada dalam nilai yang sangat kecil. Artinya, meskipun proses pembaruan bobot semakin hati-hati, informasi tetap dapat dipelajari secara efektif oleh model karena arsitektur mendukung propagasi sinyal yang stabil.

Hasil akhir menunjukkan bahwa pada konfigurasi terbaik, model berhasil mencapai akurasi uji sebesar 92%, yang merupakan peningkatan sebesar 9% dibandingkan baseline atau pendekatan sebelumnya yang tidak menerapkan skema penurunan learning rate. Peningkatan ini tidak hanya mencerminkan keberhasilan dalam aspek hyperparameter tuning, tetapi juga menunjukkan bahwa kombinasi arsitektur dan strategi pelatihan berperan penting dalam pencapaian kinerja optimal.

3.4 Hasil Visualisasi

Gambar 7 berikut ini menampilkan hasil visualisasi *batch* citra spektrogram yang dihasilkan dari data suara emosi dalam dataset RAVDESS setelah melalui tahap ekstraksi fitur dan data *preprocessing* :



Gambar 7. Hasil Visualisasi

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

Visualisasi yang ditampilkan di Gambar 7 ini, disusun menggunakan fungsi torchvision.utils.make_grid() yang menggabungkan sejumlah citra tensor dalam satu grid komposit, kemudian divisualisasikan menggunakan fungsi plt.imshow() setelah diubah urutan dimensinya dengan np.transpose() dari format tensor PyTorch (C, H, W) menjadi format NumPy (H, W, C) agar kompatibel dengan library matplotlib. Proses visualisasi ini merupakan bagian krusial dalam debugging pipeline karena memungkinkan inspeksi langsung terhadap format, warna, struktur spasial, dan keberagaman citra yang akan digunakan sebagai input bagi model CNN.

Sebelum divisualisasikan, citra spektrogram dalam format tensor terlebih dahulu dinormalisasi ke rentang [0, 1] melalui proses linear shifting img = img / 2 + 0.5. Ini dilakukan untuk mengembalikan skala piksel yang semula berada pada rentang [-1, 1] akibat proses normalisasi awal dengan statistik ImageNet (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). Akan tetapi, meskipun telah dilakukan normalisasi ulang, matplotlib mendeteksi adanya nilai-nilai ekstrim yang masih berada di luar rentang [0.0, 1.0], sehingga muncul peringatan Clipping input data to the valid range.... Ini berarti bahwa beberapa piksel memiliki nilai di bawah -0.02 atau di atas 1.16, dan sistem matplotlib secara otomatis memotong (*clipping*) nilai-nilai tersebut agar tetap dapat divisualisasikan sebagai citra RGB.

Citra-citra yang ditampilkan adalah representasi visual dari sinyal audio yang telah dikonversi menjadi spektrogram menggunakan pendekatan berbasis transformasi frekuensi seperti *Short-Time Fourier Transform* (STFT) atau Mel Spectrogram. Citra ini menampilkan distribusi energi frekuensi terhadap waktu, dengan sumbu horizontal menggambarkan waktu, dan sumbu vertikal menggambarkan frekuensi. Pola warna ungu hingga putih dalam gambar menunjukkan intensitas energi pada masing-masing frekuensi, yang menjadi dasar informasi dalam proses klasifikasi emosi.

Di bawah *grid* visualisasi, ditampilkan tensor label dari masing-masing sampel dalam *batch*. Label-label tersebut ditulis dalam format tensor PyTorch dan mewakili target kelas dari masing-masing gambar. Kelas-kelas tersebut mencakup enam emosi inti: netral (0), senang (1), sedih (2), marah (3), takut (4), dan terkejut (5). Distribusi label yang cukup merata, seperti ditunjukkan dalam label [2, 5, 2, 0, 3, 0, 5, 2, 1, 4, 5, 3, 2, 2, 1, 3, 1, 2, 2, 4,...], menunjukkan bahwa pengambilan batch telah memenuhi prinsip stratifikasi yang baik dan tidak condong terhadap satu atau dua kelas dominan. Hal ini penting untuk menjaga kestabilan loss function dan akurasi selama proses training.

Secara keseluruhan, hasil visualisasi ini membuktikan bahwa proses data loading, transformasi, dan konversi dari sinyal suara ke bentuk visual berjalan sesuai ekspektasi. Dengan tampilan grid seperti ini, peneliti dapat mengevaluasi secara kualitatif apakah data mengalami masalah seperti label noise, artifact citra, atau ketidaksesuaian format. Selain itu, visualisasi ini juga mendemonstrasikan keberhasilan *pipeline preprocessing* dalam memastikan konsistensi ukuran (224×224), kompatibilitas format tensor untuk PyTorch, dan pengacakan data (shuffle=True) agar model tidak belajar dari urutan data yang statis.

Implikasi langsung dari visualisasi ini adalah memastikan bahwa data input yang akan diberikan ke arsitektur CNN sudah dalam kondisi optimal. Setiap spektrogram dalam batch mewakili satu instance dari sinyal suara yang telah diubah bentuknya ke dalam citra, sehingga CNN dapat memanfaatkan fitur spasial dari pola-pola frekuensi untuk membedakan ekspresi emosi. Dengan memverifikasi input pada tahap ini, potensi *error* di *downstream process* dapat diminimalkan, serta meningkatkan efisiensi eksperimen dan akurasi klasifikasi.

3.5 Pembahasan

Penelitian ini mengembangkan model klasifikasi emosi suara berbasis *Convolutional Neural Network* (CNN) dengan pendekatan input berupa Mel-spektrogram, yang terbukti lebih akurat dalam menangkap ciri-ciri emosional dibandingkan metode konvensional. Untuk menilai keefektifan pendekatan ini, hasil dari model yang dikembangkan dibandingkan dengan dua penelitian sebelumnya, yakni oleh Charlen & Kusnawi (2023) serta Fransiskus Jonathan Tanudjaja et al. (2023), yang sama-sama memanfaatkan dataset RAVDESS namun menggunakan teknik dan algoritma yang berbeda.

Penelitian yang dilakukan oleh Charlen & Kusnawi menggunakan pendekatan klasik berupa *Multi-Layer Perceptron* (MLP) dan *Support Vector Machine* (SVM), dengan fitur suara yang diekstrak menggunakan *Mel-Frequency Cepstral Coefficients* (MFCC). Meskipun teknik ini cukup populer dalam pengenalan emosi berbasis suara, hasil akurasi yang dicapai masih terbatas, yaitu 83% untuk MLP dan 82% untuk SVM. Hal ini menunjukkan adanya batasan dari MFCC dalam menangkap dinamika temporal dan spektral yang kompleks pada sinyal suara, terutama ketika emosi yang diklasifikasikan memiliki kemiripan secara akustik.

Sementara itu, studi oleh Tanudjaja et al. juga menggunakan CNN sebagai algoritma utama, namun tetap mengandalkan fitur MFCC sebagai representasi input. Meskipun penggunaan CNN memberikan potensi dalam mengenali pola-pola visual, keterbatasan fitur MFCC membuat akurasi model mereka hanya mencapai sekitar 70%. Selain itu, meskipun mereka mencoba menambahkan fitur lain seperti *pitch, energy,* dan *spectral components*, perbaikan akurasi tetap belum signifikan. Beberapa emosi seperti netral dan sedih bahkan sering diklasifikasikan secara keliru karena kemiripan karakteristik suara.

Berbeda dari kedua penelitian tersebut, penelitian ini memanfaatkan Mel-spektrogram sebagai representasi data audio dalam bentuk visual dua dimensi, yang memungkinkan CNN mengenali pola spektral secara hierarkis dengan lebih efektif. Pendekatan ini menghasilkan peningkatan performa yang signifikan, dengan akurasi mencapai 92% (*macro average*) dan validasi melalui 5-*fold cross-validation* yang konsisten berada di angka 91,5% ± 0,8%. Emosi netral dan senang dapat dikenali dengan baik, ditunjukkan oleh F1-score masing-masing sebesar 93% dan 90%.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

Meskipun beberapa emosi seperti takut dan terkejut masih mengalami kesulitan dalam klasifikasi, pendekatan ini secara umum menunjukkan keunggulan yang nyata dibandingkan metode berbasis MFCC.

Dari sisi arsitektur model, penggunaan CNN dengan desain *residual learning* (Spec-ResNet) memberikan keunggulan dalam mengatasi permasalahan umum jaringan dalam, seperti vanishing gradient. Dengan empat blok konvolusional bertingkat, *global average pooling*, dan *softmax* sebagai output layer, model ini mampu belajar secara mendalam dari data Mel-spektrogram yang divisualisasikan. Dalam hal pelatihan, model dikonfigurasi menggunakan *optimizer Adam* dan *loss function* categorical *cross-entropy*, serta dilengkapi dengan teknik normalisasi dan mekanisme *early stopping* untuk mencegah *overfitting*. Proses ini menunjukkan efisiensi pelatihan dan generalisasi yang baik terhadap data baru.

Secara keseluruhan, penelitian ini tidak hanya berhasil meningkatkan akurasi dalam pengenalan emosi suara, tetapi juga menunjukkan bahwa pemilihan metode ekstraksi fitur dan representasi data memainkan peran yang sangat krusial. Dengan mengadopsi Mel-spektrogram yang lebih informatif dan arsitektur CNN yang dirancang secara cermat, penelitian ini dapat memberikan kontribusi yang cukup signifikan dalam pengembangan sistem pengenalan emosi yang lebih akurat, dan adaptif.

4. KESIMPULAN

Dalam penelitian ini, berhasil dikembangkan sebuah sistem pengenalan emosi suara (*Speech Emotion Recognition*/SER) berbasis *Convolutional Neural Network* (CNN) dengan input berupa Mel-spektrogram, sebagai alternatif yang lebih unggul dibanding pendekatan konvensional yang menggunakan fitur Mel-*Frequency Cepstral Coefficients* (MFCC). Pendekatan ini dirancang untuk menjawab tantangan utama dalam SER, yakni rendahnya akurasi dalam membedakan emosi yang memiliki karakteristik akustik yang serupa, seperti antara "takut" dan "terkejut", atau "netral" dan "tenang". Dengan memanfaatkan arsitektur CNN bertipe residual (Spec-ResNet), yang mengadopsi mekanisme *skip connection*, model ini mampu mengekstraksi fitur dari input dua dimensi secara lebih dalam tanpa mengalami degradasi performa akibat masalah *vanishing gradient*. Arsitektur yang dirancang terdiri dari empat blok konvolusional dengan batch normalization dan aktivasi ReLU, diikuti dengan global average pooling dan lapisan *fully connected* sebagai klasifikasi akhir. Seluruh proses pelatihan dioptimalkan menggunakan fungsi *loss categorical cross-entropy* dan *optimizer Adam*, dengan *early stopping* untuk menghindari *overfitting*. Eksperimen dilakukan pada dataset RAVDESS, yang terdiri dari 1.440 sampel suara dengan enam kategori emosi: netral, senang, sedih, marah, takut, dan terkejut. Model dilatih menggunakan validasi silang (5-*fold cross-validation*) untuk memastikan kestabilan performa. Hasil pengujian menunjukkan bahwa pendekatan CNN dengan Mel-spektrogram menghasilkan peningkatan akurasi yang signifikan, yaitu rata-rata 91,5% (±0,8%) dengan akurasi makro tertinggi mencapai 92%. Kategori emosi netral dan senang menunjukkan F1-score tertinggi, yaitu 93% dan 90%, yang menandakan bahwa model sangat efektif dalam mendeteksi emosi dengan ciri suara yang konsisten. Namun masih ditemukan keterbatasan pada klasifikasi emosi "takut" dan "terkejut" yang sering tumpang tindih, karena pola frekuensi keduanya sangat mirip. Hal ini membuka ruang untuk pengembangan lanjutan seperti integrasi attention mechanism, pemanfaatan *bidirectional recurrent layers*, atau pendekatan multimodal dengan menggabungkan data visual dan suara untuk meningkatkan sensitivitas klasifikasi.

REFERENCES

[1] C. A. Qurniaty and K. Kusnawi, "Ekspresi Emosi Berdasarkan Suara Menggunakan Algortima Multi Layer Perceptron dan Support Vector Machine," *Indones. J. Comput. Sci.*, vol. 12, no. 6, pp. 4014–4025, 2023, doi: 10.33022/ijcs.v12i6.3567.

[2] Y. K. Aini, T. B. Santoso, and T. Dutono, "Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia," *J. Komput. Terap.*, vol. 7, no. 1, pp. 143–152, 2021, doi: 10.35143/jkt.v7i1.4623.

[3] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.

[4] F. J. Tanudjaja, E. Y. Puspaningrum, and Y. V. Via, "Klasifikasi Jenis Emosi Melalui Ucapan Menggunakan Metode Convolutional Neural Network," *Teknologi*, vol. 13, no. 2, pp. 1–11, 2023, doi: 10.26594/teknologi.v13i2.3740.

[5] S. Madanian *et al.*, "Speech emotion recognition using machine learning — A systematic review," *Intell. Syst. with Appl.*, vol. 20, no. July, p. 200266, 2023, doi: 10.1016/j.iswa.2023.200266.

[6] H. Wijaya, "Teknologi Pengenalan Suara tentang Metode, Bahasa dan Tantangan: Systematic Literature Review," *bit-Tech*, vol. 7, no. 2, pp. 533–544, Dec. 2024, doi: 10.32877/bt.v7i2.1888.

[7] L. S. Ramba and M. Aria, "Design Of A Voice Controlled Home Automation System Using Deep Learning Convolutional Neural Network (DL-CNN)," *Telekontran J. Ilm. Telekomun. Kendali dan Elektron. Terap.*, vol. 8, no. 1, pp. 57–73, 2020, doi: 10.34010/telekontran.v8i1.3078.

[8] S. Nurmaini, A. Darmawahyuni, A. I. Sapitri, M. N. Rachmatullah, Firdaus, and B. Tutuko, *Pengenalan Deep Learning dan Implementasinya*. 2021. [Online]. Available: http://repository.unsri.ac.id/id/eprint/89078

[9] I. Minggi, N. Nasrullah, and A. Alimuddin, "Penggunaan Systematic Literature Review Berbantuan PoP untuk Pengembangan Kompetensi Guru SMP Kab. Takalar," *Dedikasi*, vol. 25, no. 2, pp. 104–111, 2023, doi: 10.26858/dedikasi.v25i2.56082.

[10] Y. N. Fuadah, I. D. Ubaidullah, N. Ibrahim, F. F. Taliningsing, N. K. Sy, And M. A. Pramuditho, "Optimasi Convolutional Neural Network dan K-Fold Cross Validation pada Sistem Klasifikasi Glaukoma," *ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron.*, vol. 10, no. 3, p. 728, 2022, doi: 10.26760/elkomika.v10i3.728.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1143−1152
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7875

[11] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu, "MixSpeech: Data augmentation for low-resource automatic speech recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, no. 2017, pp. 7008–7012, 2021, doi: 10.1109/ICASSP39728.2021.9414483.
[12] Z. Agustina, P. N. Nisa, and L. S. Laoli, "Visualisasi Dan Analisis Frekuensi Suara Musik Dengan Metodefast Fourier Dan Hamming Window," *Kohesi J. Sains dan Teknol.*, vol. 8, no. 1, pp. 21–30, 2025, doi: https://doi.org/10.2238/qgqzae16.
[13] A. Amato and V. Di Lecce, "Data preprocessing impact on machine learning algorithm performance," *Open Comput. Sci.*, vol. 13, no. 1, 2023, doi: 10.1515/comp-2022-0278.
[14] M. Bilal, G. Ali, M. W. Iqbal, M. Anwar, M. S. A. Malik, and R. A. Kadir, "Auto-prep: efficient and automated data preprocessing pipeline," *IEEE Access*, vol. 10, pp. 107764–107784, 2022, doi: 10.1109/ACCESS.2022.3198662.
[15] F. I. Muqsith, E. Supriyati, and T. Listyorini, "Klasifikasi Pengucapan Huruf Hijaiyah Berbasis Android Menggunakan CNN dengan Fitur Mel-Spectrogram," *J. Inform. J. Pengemb. IT*, vol. 10, no. 1, pp. 67–78, 2025, doi: 10.30591/jpit.v10i1.8145.
[16] E. S. Budi, A. N. Chan, P. P. Alda, and M. A. F. Idris, "Optimasi Model Machine Learning untuk Klasifikasi dan Prediksi Citra Menggunakan Algoritma Convolutional Neural Network," *Resolusi Rekayasa Tek. Inform. dan Inf.*, vol. 4, no. 5, pp. 502–509, 2024
[17] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal Emotion Recognition Using a Hierarchical Fusion Convolutional Neural Network," *IEEE Access*, vol. 9, pp. 7943–7951, 2021, doi: 10.1109/ACCESS.2021.3049516.
[18] F. Irawan and R. Hanip, "Pelatihan Model Pembelajaran Reading Concept Map (Remap) dalam Melatih Keterampilan Berpikir Kreatif dan Keterampilan Literasi Sains Peserta didik Di SMP YAPIS Merauke," *MAYARA J. Pengabdi. Masy.*, vol. 3, no. 1, pp. 26–36, 2025, doi: https://doi.org/10.71382/mayara.jurn.peng.masy..v3i1.238.
[19] S. Kumar *et al.*, "Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1523–1540, 2023, doi: 10.32604/cmc.2023.028631.
[20] B. Wijaya, M. M. E. Haqiqi, A. S. Satyawan, and H. Susilawati, "Restorasi Citra Wajah Terdegradasi Menggunakan Model GAN dan Fungsi Loss," *J. Algoritm.*, vol. 5, no. 2, pp. 254–263, 2025, doi: https://doi.org/10.35957/algoritme.v5i2.11487.