Teknologi: Jurnal Ilmiah Sistem Informasi 16 (1) 2026 1-12

ISSN 2527-3671 (*Online*)
ISSN 2087-8893 (*Print*)

Tersedia *online* di www.journal.unipdu.ac.id

Halaman jurnal di www.journal.unipdu.ac.id/index.php/teknologi

Identifikasi Ucapan Disartria Menggunakan Arsitektur Convolutional Neural Network MobileNetV2 dan MFCC

Henry Ardian Irianta [a], Abdul Fadlil [b], Rusydi Umar [c]

[a,c] *Program Studi Magister Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[b] *Program Studi Teknik Informatika, Universitas Siber Muhammadiyah, Yogyakarta, Indonesia*

email: [a,] henryai@sibermu.ac.id*

*Korespondensi

Dikirim 28 Juli 2025; **Direvisi** 05 Desember 2025; **Diterima** 04 Januari 2026; **Diterbitkan** 09 Februari 2026

Abstrak

Seiring meningkatnya kebutuhan *Automatic Speech Recognition* (ASR) pada *edge device*, pengembangan sistem deteksi gangguan bicara menjadi semakin relevan, terutama untuk aplikasi pada alat kesehatan. Deteksi dini disartria memegang peranan krusial dalam intervensi klinis, sehingga diperlukan model klasifikasi yang efisien agar kedepanya dapat diimplementasikan pada edge atau *embedded system* untuk membantu proses diagnosis. Penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi sebuah model *deep learning* yaitu *lightweight Convolutional Neural Network* (LCNN) untuk klasifikasi ucapan disartria dengan memanfaatkan arsitektur *MobileNetV2* melalui pendekatan *transfer learning*. Metode penelitian menggunakan dataset publik UASpeech, di mana fitur akustik diekstraksi menggunakan *Mel-Frequency Cepstral Coefficients* (MFCC) untuk menghasilkan 40 koefisien. Fitur MFCC kemudian divisualisasikan sebagai spektrogram untuk melatih model *MobileNetV2* yang sebelumnya telah dilatih pada dataset berskala besar. Hasil evaluasi terbaik pada data uji menunjukkan performa yang sangat baik dengan pencapaian akurasi sebesar 95%

Kata Kunci: Deteksi Disartria*, ASR, MobileNetV2, Transfer Learning, MFCC, Deep Learning, LCNN*

Dysarthric Speech Identification Using the MobileNetV2 Convolutional Neural Network Architecture and MFCC Features

Abstract

The growing demand for Automatic Speech Recognition (ASR) on edge devices makes the development of speech disorder detection systems increasingly relevant, particularly for healthcare applications. Early detection of dysarthria plays a crucial role in clinical intervention, necessitating efficient classification models that can be implemented on edge or embedded systems to aid in the diagnostic process. This research aims to implement and evaluate a lightweight Convolutional Neural Network (LCNN) deep learning model for dysarthria speech classification by leveraging the MobileNetV2 architecture through a transfer learning approach. The research methodology utilizes the public UASpeech dataset, where acoustic features are extracted using Mel-Frequency Cepstral Coefficients (MFCC) to generate 40 coefficients. These MFCC features are then visualized as spectrograms to train a MobileNetV2 model previously pre-trained on a large-scale dataset. The best evaluation results on the test data demonstrate excellent performance, achieving an accuracy of 95%,

Keywords: *Dysarthic speech Recognition, ASR, MobileNetV2, Transfer Learning, MFCC, Deep Learning, LCNN*

Untuk mengutip artikel inidengan APA Style:
Irianta. H.A., Fadlil. A., Umar. A. (2025). Identifikasi Ucapan Disartria Menggunakan Arsitektur Convolutional Neural Network MobileNetV2 dan MFCC. TEKNOLOGI: Jurnal Ilmiah Sistem Informasi, 16(2)1-12

1. Pendahuluan

Disartria adalah gangguan bicara motorik yang disebabkan oleh kerusakan pada sistem saraf pusat atau perifer, yang mengakibatkan kelemahan, kelumpuhan, atau inkoordinasi otot-otot yang terlibat dalam produksi bicara [1]. Kondisi ini secara signifikan memengaruhi artikulasi, fonasi, resonansi, dan prosodi, sehingga menurunkan kejelasan (*intelligibility*) ucapan dan menghambat kemampuan individu untuk berkomunikasi secara efektif. Dampaknya tidak hanya bersifat fungsional, tetapi juga psikososial, sering kali menyebabkan frustrasi, isolasi sosial, dan penurunan kualitas hidup [2]. Diagnosis konvensional bergantung pada evaluasi klinis oleh ahli patologi wicara, yang bisa bersifat subjektif dan memerlukan

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

waktu. Oleh karena itu, terdapat urgensi untuk mengembangkan alat bantu diagnosis otomatis yang objektif, cepat, dan dapat diakses, terutama yang dapat diimplementasikan pada perangkat portabel untuk pemantauan berkelanjutan [3].

Dalam lima tahun terakhir, riset mengenai *Automatic Speech Recognition* (ASR) telah didominasi oleh model pra-terlatih berskala besar berbasis *Transformer*, seperti Wav2Vec 2.0, yang menunjukkan performa canggih pada berbagai tugas pengenalan ucapan, termasuk untuk ucapan patologis [4]. Namun, model-model ini memiliki kompleksitas komputasi yang sangat tinggi, sehingga menjadi tantangan untuk implementasi pada perangkat dengan sumber daya terbatas. Sebagai respons, muncul tren pengembangan arsitektur yang efisien, seperti *Lightweight Convolutional Neural Networks* (LCNN), yang dirancang untuk menyeimbangkan akurasi dan efisiensi komputasi. Pendekatan LCNN menjadi sangat relevan untuk aplikasi klinis pada *edge devices*, meskipun adaptasinya untuk data disartria yang sangat bervariasi dan terbatas masih menjadi area penelitian aktif [5].

Untuk mengatasi tantangan tersebut, penelitian ini mengusulkan sebuah pendekatan yang menggabungkan ekstraksi fitur akustik yang mapan dengan arsitektur *deep learning* yang efisien. *Mel-Frequency Cepstral Coefficients* (MFCC) dipilih sebagai representasi fitur karena kemampuannya yang terbukti andal dalam menangkap karakteristik spektral ucapan yang relevan dengan persepsi pendengaran manusia [6]. Fitur ini kemudian digunakan untuk melatih arsitektur MobileNetV2, sebuah model CNN yang ringan dan dirancang untuk efisiensi komputasi [7]. Pendekatan *transfer learning* diadopsi dengan memanfaatkan model MobileNetV2 yang telah melalui pra-pelatihan pada dataset visual berskala besar. Strategi ini memungkinkan model untuk memanfaatkan pengetahuan representasi fitur tingkat rendah yang telah dipelajari sebelumnya, sehingga dapat mencapai performa tinggi bahkan dengan dataset ucapan yang relatif terbatas, sekaligus menjadikannya kandidat ideal untuk implementasi pada *embedded system* [8].

Berdasarkan latar belakang tersebut, tujuan utama dari penelitian ini adalah merancang, mengimplementasikan, dan mengevaluasi sistem klasifikasi ucapan disartria menggunakan fitur MFCC dan arsitektur MobileNetV2 dengan metode *transfer learning*. Kontribusi utama dari penelitian ini meliputi: (1) demonstrasi efektivitas pendekatan *transfer learning* dari domain visual ke domain audio untuk klasifikasi ucapan patologis; (2) validasi performa arsitektur MobileNetV2 sebagai model yang akurat dan efisien secara komputasi untuk tugas ASR; dan (3) penyediaan kerangka kerja yang dapat menjadi dasar pengembangan alat diagnosis disartria portabel berbasis ASR untuk aplikasi klinis di masa depan.

2. State Of The Art

Penelitian yang dilakukan oleh Vinotha et al. [9] memberikan informasi lebih lanjut tentang kemampuan sistem pengenalan ucapan hibrid yang mereka usulkan, menggunakan kombinasi dari Transformer dan *Connectionist Temporal Classification* (CTC). Hasil penelitian menunjukkan bahwa sistem ini mampu mengenali kata-kata dengan tingkat akurasi yang tinggi dan berhasil mengurangi *Word Error Rate* (WER) dibandingkan dengan sistem pengenalan ucapan konvensional. Dalam konteks ini, WER yang diperoleh lebih rendah dari 20%, menunjukkan efisiensi model dalam menangani variasi ucapan disartria yang kompleks.

Studi terbaru oleh Nguyen et al. [10] mendemonstrasikan kapabilitas model pra-terlatih berskala besar seperti Wav2Vec 2.0. Studi ini mengusulkan jaringan adaptasi sederhana untuk *fine-tuning* wav2vec2 menggunakan fitur fMLLR, yang juga mendukung xvectors. Hasil menunjukkan peningkatan kinerja di semua tingkat keparahan gangguan, dengan WER 57,72% pada disartria berat di dataset UASpeech, serta hasil konsisten pada dataset berbahasa Jerman [10].

Menyoroti tantangan pada data disartria, penelitian oleh Yue et al. [11] menginvestigasi penggunaan spektrum fase dan magnitudo mentah sebagai input. Menggunakan model CNN multi-stream pada dataset TORGO, mereka berhasil mencapai Word Error Rate (WER) sebesar 31,2% untuk ucapan disartria. Hasil ini menunjukkan potensi penggunaan representasi sinyal level rendah untuk menangkap detail patologis, meskipun performanya masih menunjukkan adanya ruang untuk perbaikan pada dataset yang sangat bervariasi ini [11].

Fokus pada arsitektur hibrida yang efisien, Narendra dan Al-Haddad [12] mengusulkan model CNN-LSTM untuk klasifikasi pada dataset UASpeech. Dengan menggabungkan kemampuan CNN dalam ekstraksi fitur spasial dari spektrogram dan LSTM untuk memodelkan dependensi temporal, pendekatan mereka mencapai akurasi sebesar 77,57%. Studi ini menjadi contoh penerapan model yang relatif ringan untuk membedakan ucapan disartria, meskipun akurasinya belum setinggi model-model yang lebih kompleks [12].

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

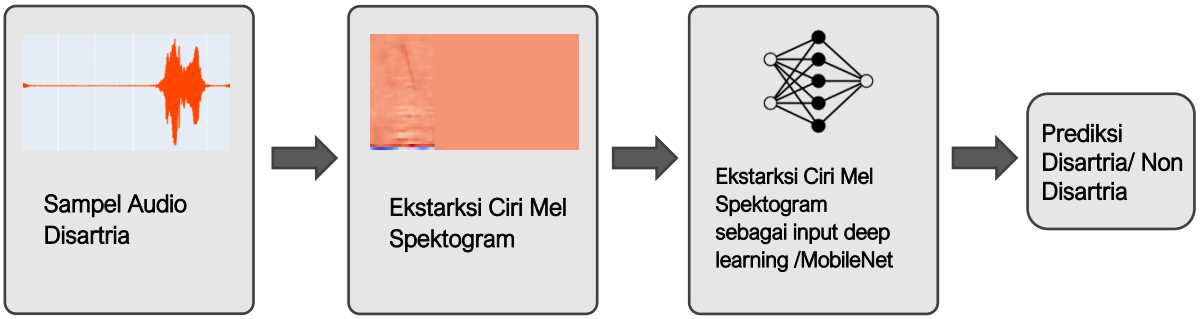
Selain itu, penelitian Lee et al. [13] mencatat bahwa sistem pengenalan ucapan otomatis (ASR) untuk disartria mampu memberikan hasil akurasi yang baik dengan penilaian konsistensi bicara yang memadai. Penelitian ini berfokus pada perbandingan antara ASR dengan evaluasi bicara konvensional oleh profesional (SLP), di mana akurasi sistem ASR dalam mengenali ucapan disartria berada pada kisaran 75% - 80%, memberikan validitas yang kuat untuk aplikasinya dalam analisis disartria oleh tim medis.

Penelitian oleh Fadlil et al. [14] menerapkan MFCC dan algoritma MLP untuk mendeteksi disartria bunyi "R" pada anak-anak. Dengan 16 koefisien MFCC dan 8 neuron tersembunyi, sistem mencapai akurasi 100% pada skenario pelatihan 80:20. Penelitian juga mengevaluasi penggunaan 16, 32, dan 64 koefisien MFCC untuk menganalisis pengaruh kompleksitas fitur terhadap akurasi dan efisiensi komputasi. Hasilnya menunjukkan potensi pengembangan sistem deteksi disartria yang ringan dan akurat untuk mendukung analisis medis berbasis suara.

Penelitian fundamental oleh Piczak [15] secara efektif menunjukkan bahwa arsitektur CNN yang dirancang untuk analisis gambar dapat langsung diterapkan pada spektrogram audio. Menggunakan arsitektur CNN sederhana, ia berhasil mencapai rata-rata akurasi sebesar 73,1% pada dataset klasifikasi suara lingkungan Urbansound8K. Keberhasilan ini membuka jalan bagi adaptasi model visi komputer yang lebih canggih dan efisien, seperti MobileNetV2, untuk tugas-tugas klasifikasi audio dalam kerangka *transfer learning* [15].

3. Metode Penelitian

Kerangka kerja yang diusulkan dalam penelitian ini mengikuti alur proses standar untuk deteksi ucapan disartria berbasis *deep learning*, sebagaimana diilustrasikan dalam **Gambar 1** dan didukung oleh penelitian serupa oleh Sekhar et al. [16]. Proses ini diawali dengan ekstraksi sampel audio dari subjek. Sampel audio tersebut kemudian ditransformasikan menjadi representasi visual 2D dalam bentuk spektrogram Mel, yang mampu menangkap karakteristik waktu-frekuensi dari sinyal. Spektrogram ini selanjutnya dijadikan input untuk model *Transfer Learning-Convolutional Neural Network* (TL-CNN), yang pada akhirnya menghasilkan prediksi kelas, yaitu "Disartria" atau "Non Disartria".

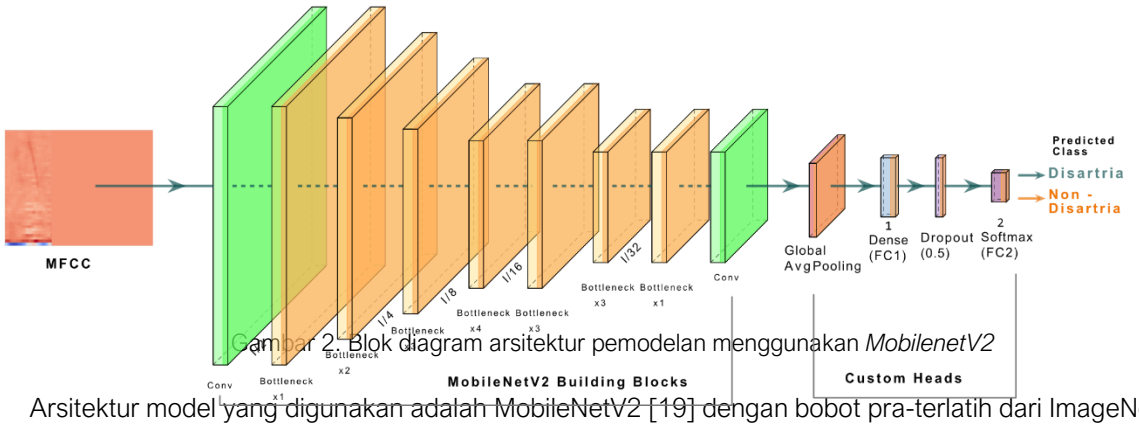


Gambar 1. Diagram alur untuk Deteksi Ucapan Disartria.

Berdasarkan observasi umum terhadap sampel audio penderita disartria, ditemukan bahwa ucapan mereka cenderung lambat, tidak jelas (*slurred*), serak (*raspy*), dan memiliki nada (*pitch*) yang lebih rendah dibandingkan individu normal. Karakteristik-karakteristik ini termanifestasi dalam bentuk pola-pola tertentu di dalam sampel audio. Oleh karena itu, penelitian ini mengusulkan sebuah sistem untuk memprediksi apakah suatu sampel audio berasal dari penderita disartria atau bukan. Rancangan arsitektur yang lebih rinci, mulai dari pra-pemrosesan data, ekstraksi ciri, hingga evaluasi model LCNN, disajikan dalam bentuk blok diagram pada **Gambar 5.**

Oleh karena itu, penelitian ini mengusulkan sebuah sistem untuk memprediksi apakah suatu sampel audio berasal dari penderita disartria atau bukan. Tahapan penelitian dimulai dari penggunaan dataset publik UASpeech [17], yang berisi rekaman dari 15 individu penderita disartria dan 13 individu kontrol yang sehat. Setiap sinyal audio melalui tahap pra-pemrosesan di mana fitur MFCC [18] dengan 40 koefisien diekstraksi, dan panjangnya diseragamkan menjadi 174 frame melalui *padding* atau *truncating*. Dataset yang telah diproses kemudian dibagi dengan rasio 80% untuk data latih dan 20% untuk data uji, menggunakan metode *stratified sampling* untuk menjaga keseimbangan proporsi kelas.

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)



Gambar 2. Blok diagram arsitektur pemodelan menggunakan *MobilenetV2*

Arsitektur model yang digunakan adalah MobileNetV2 [19] dengan bobot pra-terlatih dari ImageNet, di mana lapisan-lapisan dasarnya dibekukan (*frozen*) dan sebuah *head classifier* kustom ditambahkan, yang terdiri dari lapisan GlobalAveragePooling2D, Dense(128, 'relu'), Dropout(0.5), dan Dense(2, 'softmax'). Proses *tuning* dan pelatihan model dilakukan selama 20 *epoch* menggunakan optimizer Adam [17] dan fungsi *loss* sparse_categorical_crossentropy, dengan strategi ModelCheckpoint untuk menyimpan bobot terbaik. Evaluasi performa akhir model didasarkan pada metrik akurasi, presisi, recall, dan F1-score untuk memprediksi dua kelas target: "Disartria" dan "Non Disartria".

Seperti yang dikutip pada Gambar 2, arsitektur ini terdiri dari dua komponen utama: model dasar MobileNetV2 yang berfungsi sebagai ekstraktor fitur, dan *custom head* yang berfungsi sebagai klasifikator. Mekanisme *building block* utama dalam MobileNetV2 adalah modul *inverted residual bottleneck*, yang secara efisien memproses peta fitur melalui serangkaian lapisan konvolusi untuk mengekstraksi representasi tingkat tinggi dari spektrogram input. Dalam pendekatan *transfer learning* ini, seluruh blok *bottleneck* dari model dasar dibekukan (*frozen*). Selanjutnya, *custom head* yang ditambahkan akan mengambil peta fitur yang telah diekstraksi oleh model dasar dan memetakannya ke kelas target. *Custom head* ini terdiri dari beberapa lapisan *fully-connected* yang pada akhirnya menghasilkan prediksi antara kelas "Disartria" dan "Non-Disartria".

3.1. Dataset dan *Preprocessing Dataset*

3.1.1 Dataset

Penelitian ini menggunakan dataset publik UASpeech [17]. Subjek penelitian berasal dari *Rehabilitation Education Center* di University of Illinois at Urbana-Champaign dan melalui kolaborasi dengan *Trace Research and Development Center*, University of Wisconsin-Madison. Setiap blok berisi 255 kata, yang terdiri dari 155 kata yang diulang di setiap blok dan 100 kata tidak umum yang berbeda untuk setiap blok. Dengan demikian, setiap pembicara menghasilkan total 765 rekaman kata terisolasi.

Rangkaian 155 kata yang diulang mencakup 10 angka (misalnya, ‘zero’ hingga ‘nine’), 26 huruf alfabet radio (misalnya, ‘Alpha’, ‘Bravo’, ‘Charlie’), 19 perintah komputer (misalnya, ‘backspace’, ‘delete’, ‘enter’), dan 100 kata umum dari korpus Brown (misalnya, ‘it, is, you’). Sementara itu, 300 kata tidak umum (misalnya, ‘naturalization’, ‘moonshine’, ‘exploit’) dipilih dari novel anak-anak untuk memaksimalkan keragaman sekuens fonem (*biphone*). Struktur ini memungkinkan peneliti untuk melatih dan menguji pengenal ucapan kata-utuh (*whole-word speech recognizers*) menggunakan data berulang, sekaligus menyediakan data yang kaya akan keragaman fonetik. Setelah perekaman, data audio dari 7 kanal disimpan sebagai file wav terpisah, dan nada DTMF digunakan untuk segmentasi otomatis menjadi file kata tunggal. Tabel 1 merangkum karakteristik dari 8 dari total 19 subjek yang telah direkam dan di pilih pada peniliitan ini untuk pembicara disartria, di mana tingkat kejelasan ucapan (*speech intelligibility*) ditentukan berdasarkan tugas transkripsi kata oleh pendengar manusia, sedangkan pembicara non-disartria menggunakan dataset yang sama, dan ucapan yang sama dengan pembicara disartria, dengan *codename* ditandai dengan kata “*controlled*” atau “C” diawal katanya, dalam penelitian ini yang digunakan adalah “CM01”,”CM04”,”CM05”,”CM08,”CF01”,”CF04”,”CF05”,”CF08”.

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Tabel 1. Ringkasan Karakteristik Pembicara Disartria pada Dataset UASpeech [17]

Pembicara	Usia	Kejelasan Ucapan (%)	Diagnosis Disartria
M01	>18	sangat rendah (10%)	Spastic
M04	>18	sangat rendah (2%)	Spastic
M05	21	sedang (58%)	Spastic
M08	28	-	Spastic
F01	>18	low (19%)	Spastic
F04	18	mid (62%)	Athetoid
F05	22	high (95%)	Spastic
F08	28	-	Spastic
CM01 – CF08	-	Non-disartria	-

Spesifikasi teknis dari data audio yang digunakan dalam penelitian ini, setelah melalui tahap pra-pemrosesan, dirangkum dalam **Tabel 2.**

Tabel 2. Spesifikasi Teknis Audio Dataset

Parameter Teknis	INFORMASI
Format File	*.wav
Sample Rate	16.000 Hz
Bit Depth	16 Bit
Channel	1/Mono
Byterate	32 Kbyte/s

Untuk tujuan penelitian ini, data dari pembicara disartria dan individu non-disartria dikelompokkan menjadi dua kelas utama. Tabel 3 menyajikan rincian distribusi dataset yang digunakan dalam penelitian ini.

Tabel 3. Jumlah dataset ucapan kata per-kelas

Kelas	Jumlah Penutur	Total Data	Data Latih (80%)	Data Uji (20%)
Disartria	8	5.605	4.484	1.121
Non-Disartria	8	5.605	4.484	1.121
Jumlah	**16**	**11.210**	**8.968**	**2.242**

Tabel 3 tersebut menunjukkan jumlah total sampel audio untuk setiap kelas ('Disartria' dan 'Non-Disartria'), jumlah penutur yang berkontribusi pada setiap kelas, serta pembagian data menjadi set data latih dan data uji dengan rasio 80:20. Distribusi ini dirancang untuk memastikan bahwa dataset seimbang, yang krusial untuk mencegah bias pada model selama proses pelatihan.

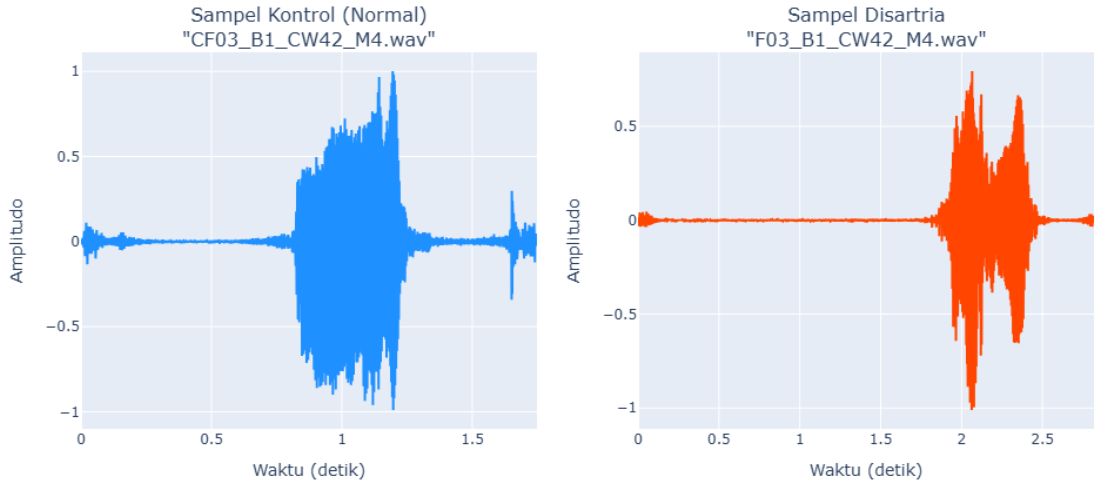
3.1.2 Pra-pemrosesan dan Ekstraksi Fitur

Sinyal audio mentah direpresentasikan sebagai grafik gelombang amplitudo terhadap waktu (*waveform*), seperti yang diilustrasikan pada Gambar 2. Dari visualisasi ini, dapat diamati perbedaan karakteristik antara ucapan kontrol (normal) dan disartria. Untuk memungkinkan model *deep learning* menganalisis pola-pola ini secara efektif, sinyal satu dimensi tersebut ditransformasikan menjadi representasi fitur dua dimensi. Dalam penelitian ini, transformasi tersebut menghasilkan *Mel-Frequency Cepstral Coefficients* (MFCC), yang secara visual dapat dilihat pada **Gambar 4**. Representasi MFCC ini

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

mengubah sinyal waktu menjadi sebuah 'gambar' waktu-frekuensi, di mana sumbu vertikal merepresentasikan koefisien-koefisien MFCC dan sumbu horizontal merepresentasikan waktu dalam frame, sehingga pola-pola spektral menjadi jelas dan dapat dipelajari oleh model *Convolutional Neural Network* (CNN), khususnya arsitektur *Lightweight* CNN (LCNN) seperti MobileNetV2 yang digunakan dalam penelitian ini.

Tahap pra-pemrosesan dan ekstraksi fitur bertujuan untuk mengubah sinyal audio mentah menjadi representasi numerik yang informatif dan seragam, yang sesuai untuk dianalisis oleh model *Convolutional Neural Network* (CNN). Proses ini dimulai dengan standardisasi sinyal, di mana seluruh rekaman audio dikonversi menjadi format mono-channel dan laju samplingnya diseragamkan menjadi 16 kHz. Selanjutnya, dilakukan ekstraksi fitur menggunakan metode *Mel-Frequency Cepstral Coefficients* (MFCC) [18]. Sinyal audio dianalisis dalam segmen-segmen pendek yang tumpang tindih (*overlapping frames*), di mana dari setiap segmen diekstraksi sebanyak 40 koefisien MFCC. Pemilihan 40 koefisien ini bertujuan untuk menangkap informasi detail mengenai amplop spektral sinyal, yang secara efektif merepresentasikan karakteristik artikulasi dan fonetik ucapan. Mengingat durasi setiap rekaman kata yang bervariasi, panjang sekuens fitur yang dihasilkan tidak seragam. Untuk mengatasi hal ini, dilakukan penyeragaman dimensi temporal menjadi 174 frame. Sekuens fitur yang melebihi panjang ini akan dipotong (*truncating*), sementara sekuens yang lebih pendek akan ditambahkan dengan nilai nol (*zero-padding*) hingga mencapai Panjang yang telah ditentukan. Hasil akhir dari tahap ini adalah sebuah set data fitur dalam bentuk tensor tiga dimensi (jumlah_sampel, 40, 174), yang siap untuk dijadikan input bagi arsitektur model yang diusulkan.

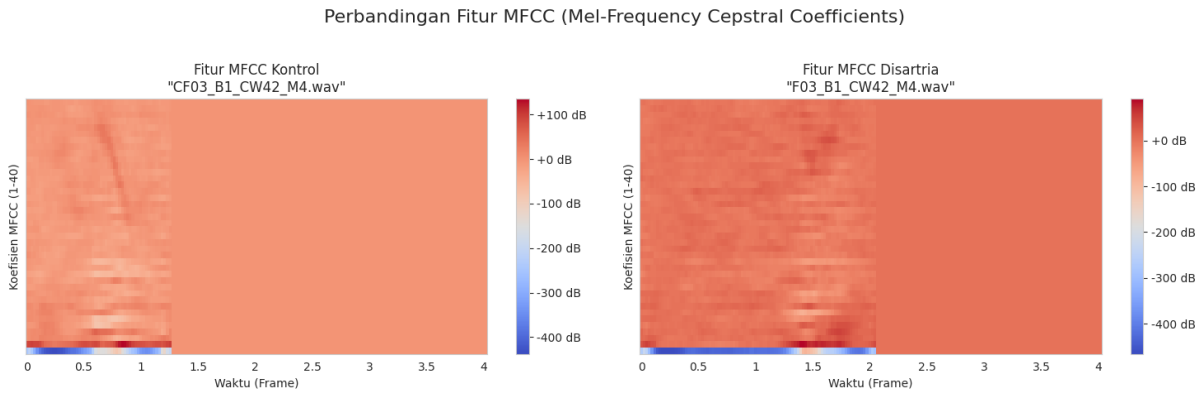


Gambar 3. Perbandingan *Waveplot* Pasangan Sampel Audio, Kiri: Sampel Kontrol (Non-Disartria), Kanan: Sampel Disartria

Visualisasi hasil pra-pemrosesan ditampilkan dalam dua bentuk: **gelombang sinyal audio (*waveform*)** dan **MFCC**. **Gambar 3** menyajikan perbandingan *waveplot* pasangan sampel audio antara ucapan normal (kontrol) dan disartria. Gambar sebelah kiri menampilkan sinyal dari sampel kontrol berkas **"CF03_B1_CW42_M4.wav"**, dengan distribusi amplitudo yang relatif stabil dan terstruktur. Sementara itu, gambar sebelah kanan menampilkan sinyal dari sampel disartria berkas **"F03_B1_CW42_M4.wav"**, dengan amplitudo yang lebih bervariasi dan durasi sinyal yang lebih panjang, mengindikasikan ketidakteraturan artikulasi suara akibat gangguan motorik bicara. Visualisasi ini digunakan sebagai alat eksplorasi awal untuk memahami karakteristik sinyal, tetapi tidak digunakan sebagai fitur input dalam proses pelatihan model.

Sebaliknya, **Gambar 4** menampilkan hasil transformasi sinyal audio menjadi fitur MFCC, yaitu representasi waktu-frekuensi dua dimensi yang digunakan secara langsung sebagai masukan dalam pelatihan model *deep learning*. Gambar sebelah kiri menunjukkan fitur MFCC dari sampel kontrol **"CF03_B1_CW42_M4.wav"**, dengan distribusi spektral yang padat di awal waktu dan pola yang konsisten. Gambar sebelah kanan menunjukkan MFCC dari sampel disartria **"F03_B1_CW42_M4.wav"**, yang memperlihatkan pola energi yang tersebar dan fluktuatif, mencerminkan gangguan dalam kontur fonetik. Transformasi ini dilakukan melalui serangkaian tahapan otomatis yang mencakup konversi sinyal, ekstraksi MFCC sebanyak 40 koefisien, dan penyeragaman panjang sekuens menjadi 174 frame.

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

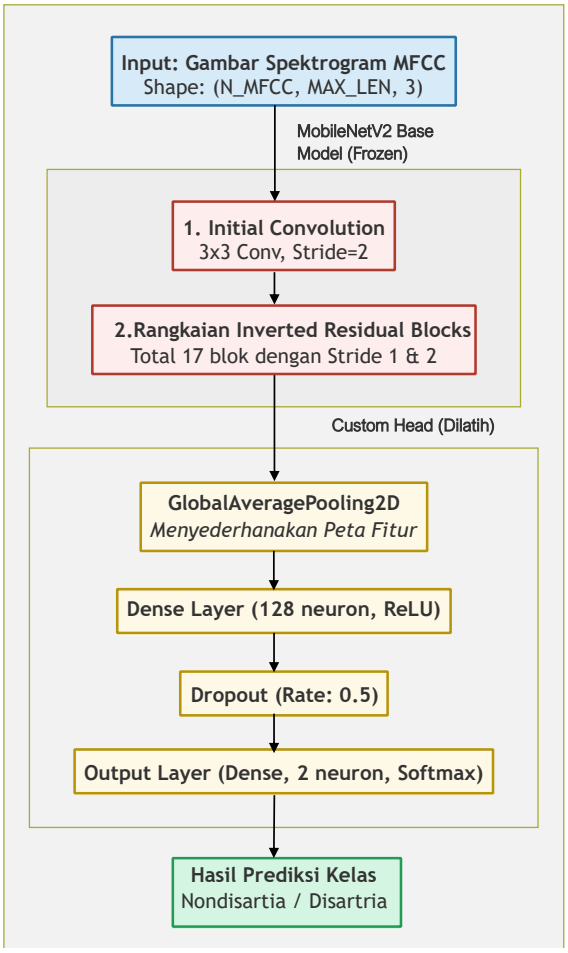


Gambar 4. Perbandingan *MFCC* Pasangan Sampel Audio, Kiri: Sampel Kontrol (Non-Disartria) Kanan: Sampel Disartria

Hasil akhirnya adalah tensor berukuran tetap 40×174x3 yang digunakan sebagai input model LCNN (MobileNetV2).

3.2 Arsitektur Pemodelan *MobilenetV2*

Arsitektur model yang diusulkan dalam penelitian ini berfokus pada efisiensi komputasi dengan mengadopsi pendekatan *Lightweight Convolutional Neural Network*. Secara spesifik, model MobileNetV2 [7] dipilih sebagai arsitektur dasar.



Gambar 5. Arsitektur pemodelan *MobilenetV2*

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Berbeda dengan model CNN konvensional seperti VGG atau ResNet yang memiliki jutaan parameter dan memerlukan sumber daya komputasi yang besar, MobileNetV2 dirancang untuk perangkat berdaya rendah, menjadikannya kandidat ideal untuk implementasi pada *edge devices* atau sistem tertanam di bidang kesehatan. Pendekatan ini dibangun dengan memanfaatkan strategi *transfer learning* berbasis ekstraksi ciri. Alur pemrosesan data dalam arsitektur yang diusulkan dikutip pada **Gambar 3**. Proses dimulai dengan input berupa gambar spektrogram MFCC berdimensi (N_MFCC, MAX_LEN, 3). Input ini pertama kali melewati model dasar MobileNetV2 yang bobotnya telah dibekukan (*frozen*). Strategi pembekuan ini merupakan inti dari pendekatan *transfer learning* yang digunakan, di mana model dasar berfungsi sebagai ekstraktor fitur tetap. Bobot pra-terlatih dari ImageNet, yang telah belajar mengenali pola-pola visual umum, dipertahankan tanpa pembaruan. Hal ini bertujuan untuk mencegah *catastrophic forgetting* dan memungkinkan model untuk menggeneralisasi dengan baik meskipun dilatih pada dataset yang relatif kecil [20-22]. Output dari model dasar ini kemudian diteruskan ke *head classifier* kustom yang lapisannya dilatih secara spesifik untuk tugas ini [23]. *Head classifier* inilah satu-satunya bagian dari model yang bobotnya diperbarui selama pelatihan, memungkinkannya untuk belajar memetakan fitur-fitur yang diekstraksi oleh MobileNetV2 ke kelas target. *Head classifier* ini menyederhanakan peta fitur menggunakan GlobalAveragePooling2D, lalu memprosesnya melalui lapisan Dense (128 neuron, ReLU) dan Dropout (rate 0.5) untuk regularisasi. Tahap akhir adalah lapisan output Dense (2 neuron, Softmax) yang menghasilkan probabilitas untuk setiap kelas, sehingga memberikan hasil prediksi akhir: "Nondisartria" atau "Disartria".

Model yang dihasilkan memiliki total 2.422.210 parameter, namun hanya 164.226 parameter (sekitar 641 KB) yang dapat dilatih (*trainable*), sementara 2.257.984 parameter (sekitar 8.61 MB) dari model dasar tetap beku. Ukuran total model yang relatif kecil (sekitar 9.24 MB) dan jumlah parameter terlatih yang sangat minim menunjukkan bahwa model ini sangat efisien dan sangat memungkinkan untuk diterapkan pada *embedded system* di masa depan.

Pelatihan dan evaluasi model diawali dengan pembagian dataset menjadi data latih (80%) dan data uji (20%) menggunakan strategi *stratified sampling* untuk memastikan proporsi kelas "Disartria" dan "Kontrol" tetap seimbang di kedua set. Model kemudian dikompilasi menggunakan optimizer Adam [24-27] dengan laju pembelajaran (*learning rate*) awal 0.001 dan fungsi *loss sparse_categorical_crossentropy*. Proses pelatihan dijalankan dengan ukuran *batch* 32 selama 10 *epoch*, dan sebuah *callback* ModelCheckpoint diaktifkan untuk memonitor val_accuracy serta secara otomatis menyimpan hanya bobot model yang menghasilkan akurasi terbaik pada data validasi. Performa model dievaluasi pada data uji menggunakan *confusion matrix*, di mana dari matriks ini dihitung beberapa metrik standar: **Akurasi**, **Presisi**, **Recall**, dan **F1-score**.

4. Hasil dan Pembahasan

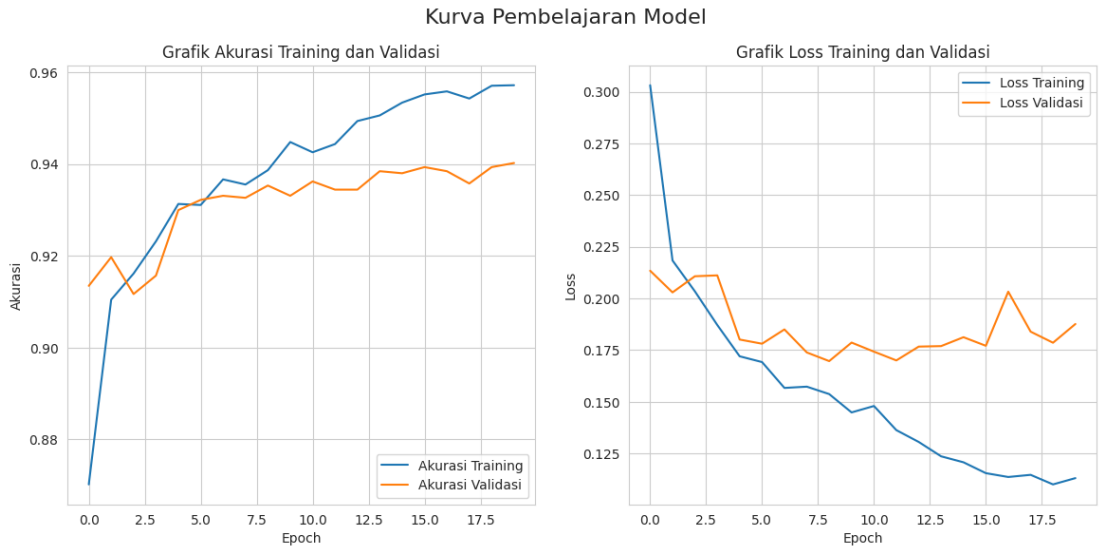
Proses pelatihan model selama 20 epoch terdokumentasi dalam **Tabel 1** yang menampilkan nilai **akurasi**, **loss**, **val_accuracy**, dan **val_loss** untuk setiap epoch.

Tabel 4. Hasil Akurasi Dan Loss pada *training* pemodelan

Epoch	Accuracy	Loss	Val_Accuracy	Val_Loss
0	0.8702	0.3032	0.9135	0.2134
2	0.9161	0.2034	0.9117	0.2107
4	0.9182	0.1872	0.9150	0.1987
6	0.9311	0.1696	0.9322	0.1781
8	0.9355	0.1573	0.9326	0.1739
10	0.9448	0.1448	0.9331	0.1756
12	0.9505	0.1262	0.9366	0.1684
14	0.9518	0.1205	0.9340	0.1764
16	0.9562	0.1136	0.9384	0.1780
18	0.9571	0.1100	0.9393	0.1786

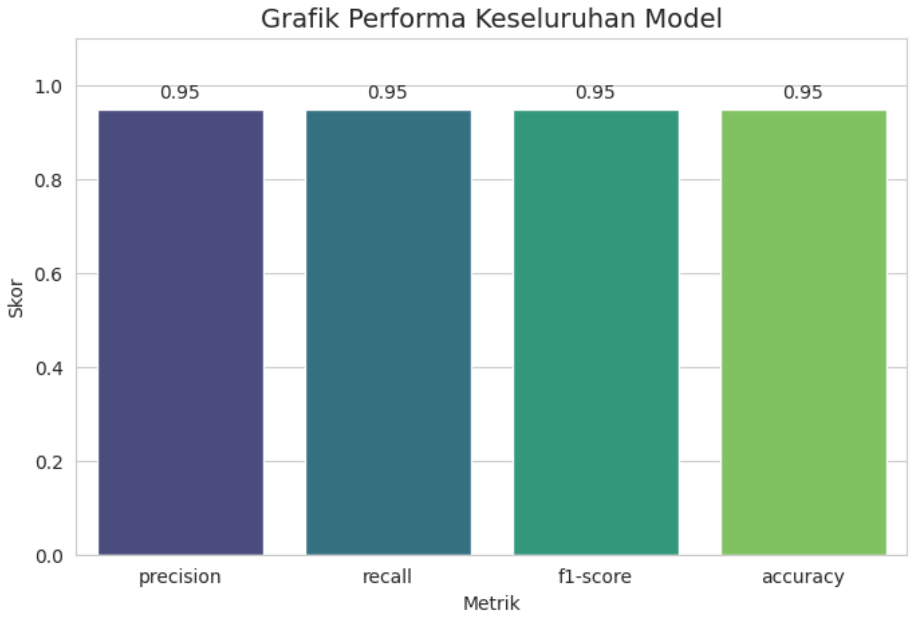
ISSN 2502-3357 (*Online*)
H.A. Irianta. dkk./TEKNOLOGI 16 (2) 2026 1-12 ISSN 2503-0477 (*Print*)

Dari tabel tersebut, terlihat bahwa **akurasi pelatihan meningkat dari 0.8702 pada epoch pertama menjadi 0.9572 pada epoch ke-20**, sementara **akurasi validasi bergerak stabil di kisaran 0.9135 hingga 0.9402**. Tren ini menunjukkan bahwa model mampu mempelajari pola dari data pelatihan secara progresif dan mempertahankan performa yang baik pada data validasi. Selain itu, nilai **loss pelatihan menurun dari 0.3032 menjadi 0.1100**, sedangkan **loss validasi bergerak fluktuatif di kisaran 0.1739 hingga 0.2134**. Perbedaan antara loss pelatihan dan loss validasi pada epoch akhir menunjukkan **indikasi awal overfitting ringan**, meskipun masih dalam batas wajar.



Gambar 6. Kurva Pembelajaran Model Kiri: Akurasi Training dan Validasi, Kanan: Loss Training dan Validasi

Visualisasi hasil pelatihan model ditampilkan pada **Gambar 6**, yang memperlihatkan kurva akurasi (sebelah kiri) dan loss (sebelah kanan) sepanjang proses pelatihan.

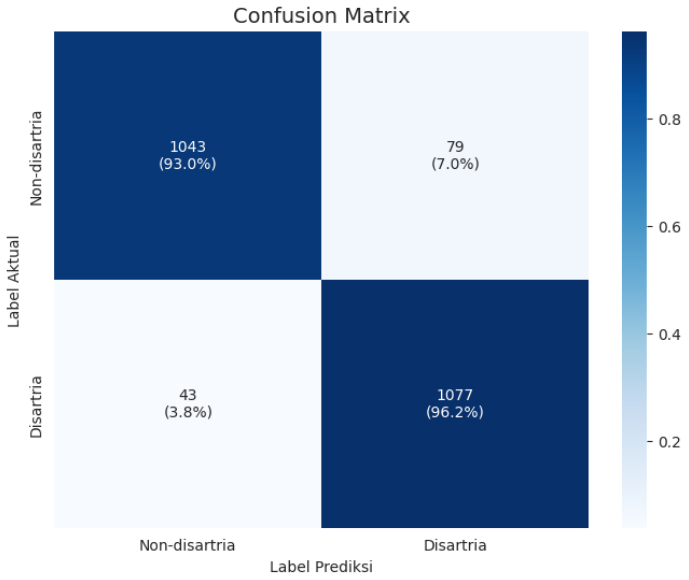


Gambar 7. Grafik Laporan Klasifikasi per Kelas Visualisasi skor precision, recall, dan F1-score untuk masing-masing kelas

Kurva tersebut menunjukkan bahwa model berhasil mempelajari pola data secara efektif dan menunjukkan kestabilan performa, dengan perbedaan yang relatif kecil antara hasil pada data pelatihan dan validasi. Evaluasi model pada data uji dilakukan dengan menghitung metrik **precision**, **recall**, dan **F1-score** untuk

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

masing-masing kelas. Berdasarkan hasil yang ditampilkan pada **Gambar 7**, model menunjukkan kinerja yang sangat solid dan seimbang di semua metrik evaluasi utama. Model mencapai **akurasi keseluruhan sebesar 95%**, yang menunjukkan kemampuannya yang sangat baik dalam membedakan antara ucapan disartria dan non-disartria. Lebih lanjut, nilai **presisi**, **recall**, dan **F1-score** secara konsisten berada di angka 95%. Nilai F1-score yang tinggi mengonfirmasi bahwa model memiliki keseimbangan yang sangat baik antara presisi (kemampuan untuk tidak salah melabeli) dan recall (kemampuan untuk menemukan semua sampel positif), dan tidak bias terhadap salah satu kelas. Konsistensi skor di semua metrik ini menegaskan bahwa model yang diusulkan tidak hanya akurat, tetapi juga andal dan *robust*.



Gambar 8. *Confusion Matrix* Model Klasifikasi.

Analisis lanjutan dilakukan dengan menggunakan **confusion matrix** seperti ditunjukkan pada **Gambar 8**. Matrix ini menggambarkan jumlah prediksi benar dan salah yang dilakukan oleh model terhadap dua kelas: *Non-disartria* dan *Disartria*.
Dari total data uji, matriks menunjukkan:

- **1043** sampel Non-disartria diklasifikasikan dengan benar, yang berarti model menghasilkan *true negative rate* (TNR) atau spesifisitas sebesar **93.0%** untuk kelas ini.
- **79** sampel Non-disartria salah diklasifikasikan sebagai Disartria, mewakili *false positive rate* (FPR) sebesar **7.0%**.

Sementara itu, untuk kelas Disartria:

- **1077** sampel Disartria berhasil diklasifikasikan dengan tepat, menunjukkan *true positive rate* (TPR) atau sensitivitas sebesar **96.2%**.
- **43** sampel Disartria diklasifikasikan secara keliru sebagai Non-disartria, dengan *false negative rate* (FNR) sebesar **3.8%**.

Proporsi kesalahan yang relatif rendah dan hampir seimbang antara kedua jenis kesalahan (FP dan FN) menunjukkan bahwa model memiliki kinerja prediksi yang seimbang dan tidak menunjukkan kecenderungan dominan terhadap salah satu label. Keseimbangan ini merupakan aspek krusial dalam aplikasi diagnosis gangguan bicara, di mana ketidakseimbangan klasifikasi dapat mengakibatkan risiko kesalahan diagnosis yang merugikan, khususnya pada kasus-kasus klinis yang sensitif.

5. Kesimpulan

Penelitian ini bertujuan untuk mengembangkan model klasifikasi ucapan guna membedakan antara sampel disartria dan non-disartria dengan memanfaatkan representasi fitur Mel-Frequency Cepstral Coefficients (MFCC) sebagai input dan menerapkan pendekatan *transfer learning* menggunakan arsitektur MobileNetV2. Hasil pelatihan menunjukkan bahwa model

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

berhasil mencapai akurasi validasi akhir sebesar 95%, dengan performa klasifikasi yang seimbang pada kedua kelas berdasarkan metrik precision, recall, dan F1-score. Hal ini menunjukkan bahwa pendekatan transfer learning dengan MobileNetV2 dapat secara efektif digunakan dalam tugas klasifikasi gangguan bicara berbasis audio, bahkan dengan dataset berukuran terbatas dan arsitektur ringan.

Untuk penelitian selanjutnya, disarankan agar dilakukan fine-tuning pada beberapa lapisan dalam arsitektur MobileNetV2 guna mengeksplorasi potensi peningkatan performa yang lebih optimal. Selain itu, penting untuk mempertimbangkan penggunaan arsitektur pra-terlatih alternatif seperti EfficientNet, ResNet, atau model berbasis transformer untuk dibandingkan kinerjanya dalam tugas klasifikasi disartria. Penelitian lanjutan juga dapat difokuskan pada upaya menurunkan jumlah parameter model serta ukuran file model agar lebih efisien, baik dari sisi komputasi maupun *deployment* di perangkat *edge*. Hal ini dapat dilakukan melalui pendekatan seperti *pruning, quantization,* atau modifikasi arsitektur CNN ringan yang disesuaikan, dengan tetap mempertahankan akurasi prediksi yang tinggi. Selain itu, penggunaan dataset yang lebih besar dan beragam sangat direkomendasikan untuk memperkuat kemampuan generalisasi model, terutama dalam konteks aplikasi klinis di dunia nyata yang lebih kompleks.

6. Kontribusi Penelitian

H. A. Irianta: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, dan Writing – original draft. A. Fadlil: Supervision, Validation, dan Writing – review & editing. R. Umar: Supervision, Validation, dan Writing – review & editing.

7. Kontribusi Penelitian

Penulis menyatakan tidak ada konflik kepentingan.

8. Referensi

[1] J. R. Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, 3rd ed. St. Louis, MO, USA: Elsevier, 2013.
[2] K. M. Yorkston, D. R. Beukelman, E. A. Strand, and M. Hakel, Management of Motor Speech Disorders in Children and Adults, 3rd ed. Austin, TX, USA: Pro-Ed, 2010.
[3] J. C. Vásquez-Correa, T. Arias-Vergara, and J. R. Orozco-Arroyave, "Pathological speech processing: State-of-the-art, current challenges, and future trends," Signal Process., vol. 157, pp. 193-214, 2019.
[4] M. K. Baskar et al., “Speaker adaptation for Wav2vec2 based dysarthric ASR,” Interspeech 2022, pp. 3403–3407, Sep. 2022, doi: 10.21437/interspeech.2022-10896.
[5] B. Riyanta, H. A. Irianta, and B. P. Kamiel, “Development of Speech Command Control Based TinyML System for Post-Stroke Dysarthria Therapy Device,” Journal of Robotics and Control (JRC), vol. 4, no. 4, pp. 466–478, Jul. 2023, doi: 10.18196/jrc.v4i4.15918.
[6]S. Mehra, V. Ranga, and R. Agarwal, “A deep learning approach to dysarthric utterance classification with BiLSTM-GRU, speech cue filtering, and log mel spectrograms,” The Journal of Supercomputing, vol. 80, no. 10, pp. 14520–14547, Mar. 2024, doi: 10.1007/s11227-024-06015-x.
[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510-4520.
[8] A. Schuh, M. Zöhrer, F. B. Pokorny, and B. Pernkopf, "Speech-based diagnosis of neurological diseases using transfer learning from big data," in Proc. Interspeech, 2019, pp. 4484-4488.
[9] V. R, H. D, and V. Anand, “Empowering Dysarthric Communication: Hybrid Transformer-CTC based Speech Recognition System,” Dec. 2023, doi: 10.21203/rs.3.rs-3736965/v1.

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)
[10] M. K. Baskar et al., “Speaker adaptation for Wav2vec2 based dysarthric ASR,” Interspeech 2022, pp. 3403–3407, Sep. 2022, doi: 10.21437/interspeech.2022-10896.
[11] S. Kim, A. Gholami, A. Shaw, and S. Lee, "Squeezeformer: An Efficient Transformer for Automatic Speech Recognition," in *Proc. Interspeech*, 2022, pp. 4118-4122.
[12] Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571-575.
[13] S. H. Lee, M. Kim, H. G. Seo, B.-M. Oh, G. Lee, and J.-H. Leigh, “Assessment of Dysarthria Using One-Word Speech Recognition with Hidden Markov Models,” Journal of Korean Medical Science, vol. 34, no. 13, 2019, doi: 10.3346/jkms.2019.34.e108.
[14] A. Fadlil, L. Perdana, A. Pujiyanta, H. man, H. I. Karim Fathurrahman, and M. M. Jogo Samodro, “Implementation of Dysarthria Identification Using MFCC and Multilayer Perceptron Algorithm,” International Journal of Electrical and Electronics Engineering, vol. 12, no. 1, pp. 32–46, Jan. 2025, doi: 10.14445/23488379/ijeee-v12i1p105.
[15] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2015, pp. 1-6.
[16] S. R. Mani Sekhar, G. Kashyap, A. Bhansali, A. A. A., and K. Singh, “Dysarthric-speech detection using transfer learning with convolutional neural networks,” ICT Express, vol. 8, no. 1, pp. 61–64, Mar. 2022, doi: 10.1016/j.icte.2021.07.004.
[17] H. Kim *et al*., "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1745–1748.
[18] A. A. Ali, M. A. El-Gamal, and M. I. El-Adawy, "A deep learning approach for spectrogram and MFCC features," *IEEE Access*, vol. 9, pp. 109876-109889, 2021.
[19] Y. Gulzar, “Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique,” Sustainability, vol. 15, no. 3, p. 1906, Jan. 2023, doi: 10.3390/su15031906.
[20] A. Schuh, M. Zöhrer, F. B. Pokorny, and B. Pernkopf, "Speech-based diagnosis of neurological diseases using transfer learning from big data," in *Proc. Interspeech*, 2019, pp. 4484-4488.
[21] T. Sugiharto, Saparudin, and W. F. A. Maki, “Indonesian Cued Speech Transliterate System Using Convolutional Neural Network MobileNet,” 2024 Ninth International Conference on Informatics and Computing (ICIC), pp. 1–7, Oct. 2024, doi: 10.1109/icic64337.2024.10957117.
[22] T. S. Winanto, C. Rozikin, and A. Jamaludin, “Analisa Performa Arsitektur Transfer Learning Untuk Mengindentifikasi Penyakit Daun Pada Tanaman Pangan,” Journal of Applied Informatics and Computing, vol. 7, no. 1, pp. 68–81, Jul. 2023, doi: 10.30871/jaic.v7i1.5991.
[23] N. wangsa Kencana, R. Umar, and Murinto, “Implementasi Transfer Learning Untuk Klasifikasi Jenis Ras Ayam Menggunakan Arsitektur MobileNetV2,” Jurnal Informatika Polinema, vol. 11, no. 2, pp. 147–154, Feb. 2025, doi: 10.33795/jip.v11i2.6469.
[24] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.
[25] M. I. Fathur Rozi, N. O. Adiwijaya, and D. I. Swasono, “Identifikasi Kinerja Arsitektur Transfer Learning Vgg16, Resnet-50, Dan Inception-V3 Dalam Pengklasifikasian Citra Penyakit Daun Tomat,” Jurnal Riset Rekayasa Elektro, vol. 5, no. 2, p. 145, Dec. 2023, doi: 10.30595/jrre.v5i2.18050.
[26] M. I. Fathur Rozi, N. O. Adiwijaya, and D. I. Swasono, “Identifikasi Kinerja Arsitektur Transfer Learning Vgg16, Resnet-50, Dan Inception-V3 Dalam Pengklasifikasian Citra Penyakit Daun Tomat,” Jurnal Riset Rekayasa Elektro, vol. 5, no. 2, p. 145, Dec. 2023, doi: 10.30595/jrre.v5i2.18050.
[27] E. B. Sudewo, M. Kunta Biddinika, R. Umar, and A. Fadlil, “Evaluating the Impact of Optimizer Hyperparameters on ResNet in Hanacaraka Character Recognition,” Preservation, Digital Technology & Culture, Feb. 2025, doi: 10.1515/pdtc-2024-0061.
Feedback: support@crossref.org