Journal of Vocational Education and Information Technology
Vol. 7, No. 2, Desember 2026, 607 - 618
E-ISSN: 2722-5305
Terakreditasi **SINTA Peringkat 4**
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Nomor: 230/E/KPT/2024

Ensemble Capsule Network Berbasis *MFCC20* Untuk Klasifikasi Audio Batuk

Desta Rizky Andhika [1], Faisal Muttaqin [2,*], Fetty Tri Anggraeny [3]

[1,2,3] Program Studi Informatika; Universitas Pembangunan Nasional Veteran Jawa Timur:
e-mail: 22081010072@student.upnjatim.ac.id[1], faisalmuttaqin.if@upnjatim.ac.id[2],
fettyanggraeny.if@upnjatim.ac.id[3]

* Korespondensi: e-mail: faisalmuttaqin.if@upnjatim.ac.id,

Diterima: 10-08-2026 ; Review: 13-08-2026; Disetujui: 13-10-2026

Abstrak: Batuk kering dan batuk berdahak memiliki karakteristik akustik yang dapat dimanfaatkan sebagai dasar klasifikasi audio untuk mendukung skrining awal secara non-invasif. Namun, variasi kualitas rekaman, perbedaan pola spektral, serta kemiripan karakteristik antarjenis batuk dapat memengaruhi kinerja model tunggal. Penelitian ini bertujuan mengembangkan dan mengevaluasi pendekatan *ensemble weighted soft voting* berbasis Capsule Network dengan representasi fitur Mel-Frequency Cepstral Coefficients (MFCC) untuk klasifikasi audio batuk kering dan batuk berdahak. Dataset yang digunakan berasal dari data publik yang telah melalui proses kurasi hingga diperoleh subset *balanced-confident* sebanyak 800 audio, terdiri atas 400 audio batuk kering dan 400 audio batuk berdahak. Data dibagi secara stratifikasi menjadi data pelatihan, validasi, dan pengujian. Fitur MFCC13 digunakan pada model M1–M4, sedangkan MFCC20 digunakan pada model M5. Model tunggal MFCC20-CapsNet digunakan sebagai baseline, sedangkan ensemble dibentuk dengan menggabungkan probabilitas prediksi lima model menggunakan *weighted soft voting*. Hasil pengujian menunjukkan bahwa baseline memperoleh *accuracy* dan *balanced accuracy* sebesar 71,67%, sedangkan ensemble memperoleh *accuracy* dan *balanced accuracy* rata-rata sebesar 70,58% dari sepuluh kali pengujian. Ensemble menunjukkan performa yang kompetitif dan memberikan evaluasi yang lebih kuat melalui pengujian berulang, meskipun peningkatan terhadap baseline tunggal belum bersifat mutlak
.
Kata kunci: batuk kering, batuk berdahak, capsule network, MFCC, weighted soft voting ensemble

Abstract: *Dry cough and wet cough have acoustic characteristics that can be utilized as a basis for audio classification to support non-invasive preliminary screening. However, variations in recording quality, differences in spectral patterns, and similarities in characteristics between cough types may affect the performance of a single model. This study aims to develop and evaluate a weighted soft voting ensemble approach based on a Capsule Network using Mel-Frequency Cepstral Coefficients (MFCC) feature representations for the classification of dry and wet cough audio. The dataset was obtained from publicly available data that had undergone a curation process, resulting in a balanced-confident subset of 800 audio recordings, consisting of 400 dry cough recordings and 400 wet cough recordings. The data were stratified into training, validation, and testing sets. MFCC13 features were used in models M1–M4, while MFCC20 features were used in model M5. The single MFCC20-CapsNet model was used as the baseline, while the ensemble was constructed by combining the prediction probabilities of five models using weighted soft voting. The results showed that the baseline achieved an accuracy and balanced accuracy of 71.67%, whereas the ensemble achieved an average accuracy and balanced accuracy of 70.58% across ten evaluation runs. The ensemble demonstrated*

Copyright@2026. LEMBAGA PENGEMBANGAN DAN INOVASI UNDHARI

competitive performance and provided a more robust evaluation through repeated testing, although it did not achieve an absolute improvement over the single-model baseline.
***Keywords**: dry cough, productive cough, capsule network, MFCC, weighted soft voting ensemble.*

1. Pendahuluan

Batuk merupakan respons fisiologis yang berfungsi membersihkan saluran pernapasan dari lendir, partikel asing, maupun rangsangan iritatif. Berdasarkan keberadaan sputum, batuk dapat dibedakan menjadi batuk kering dan batuk berdahak. Batuk kering umumnya tidak disertai produksi sekret, sedangkan batuk berdahak ditandai dengan adanya lendir dari saluran pernapasan. Perbedaan karakteristik tersebut dapat memberikan informasi awal mengenai kondisi respirasi, meskipun tidak dapat digunakan secara langsung sebagai dasar diagnosis medis. Oleh karena itu, analisis audio batuk lebih tepat digunakan sebagai pendekatan pendukung untuk skrining awal yang bersifat non-invasif.[1], [2]

Perkembangan machine learning dan deep learning memungkinkan rekaman batuk dianalisis secara komputasional berdasarkan pola akustik yang terkandung di dalamnya, seperti energi, durasi, distribusi frekuensi, dan perubahan temporal. Beberapa penelitian menunjukkan bahwa audio batuk dapat dimanfaatkan dalam klasifikasi gangguan pernapasan, seperti COVID-19 dan tuberkulosis, menggunakan pendekatan deep learning.[3], [4]. Salah satu fitur yang banyak digunakan dalam pengolahan sinyal suara adalah Mel-Frequency Cepstral Coefficients (MFCC), karena mampu merepresentasikan karakteristik spektral audio berdasarkan skala Mel yang mendekati persepsi pendengaran manusia[5], [6]

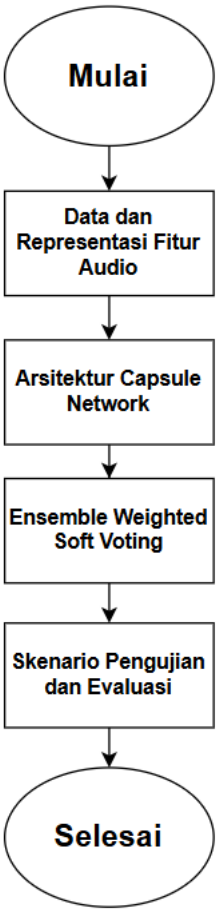
Selain pemilihan fitur, arsitektur model juga berperan penting dalam proses klasifikasi. Capsule Network (CapsNet) merupakan arsitektur deep learning yang mampu mempertahankan hubungan antarfitur melalui mekanisme routing-by-agreement. Beberapa penelitian menunjukkan bahwa CapsNet berpotensi digunakan pada klasifikasi audio medis, termasuk audio batuk.[7], [8] Namun, penggunaan model tunggal masih dipengaruhi oleh konfigurasi fitur dan parameter sehingga performanya belum selalu konsisten.

Untuk mengurangi ketergantungan terhadap satu model tunggal, penelitian ini menerapkan weighted soft voting yang menggabungkan probabilitas prediksi beberapa Capsule Network. Pendekatan ini memungkinkan model dengan performa validasi lebih baik memberikan kontribusi lebih besar terhadap keputusan akhir. Berdasarkan kajian sebelumnya, penerapan ensemble pada klasifikasi audio masih terbatas, khususnya untuk kombinasi CapsNet dengan MFCC13 dan MFCC20 pada klasifikasi batuk kering dan batuk berdahak. Beberapa penelitian terdahulu telah berhasil menerapkan metode deep learning untuk klasifikasi sinyal audio medis, seperti penggunaan BiLSTM dan CNN berbasis MFCC untuk membedakan batuk kering dan berdahak.[9] Di sisi lain, Capsule Network (CapsNet) juga terbukti unggul dalam mengekstrak fitur spasial audio[10] Meskipun teknik ensemble terbukti dapat meningkatkan stabilitas prediksi audio, enerapan skema Weighted Soft Voting Ensemble yang menggabungkan variasi fitur MFCC13 dan MFCC20 pada arsitektur CapsNet masih belum dieksplorasi secara mendalam. [11]

Penelitian ini menggunakan dataset publik yang telah melalui proses kurasi sehingga diperoleh dataset balanced-confident sebanyak 800 audio yang terdiri atas batuk kering dan batuk berdahak. Selanjutnya dikembangkan weighted soft voting ensemble berbasis lima Capsule Network menggunakan representasi MFCC13 dan MFCC20. Kinerja model dievaluasi terhadap model tunggal menggunakan accuracy, precision, recall, F1-score, specificity, balanced accuracy, macro-F1, dan MCC

2. Metode Penelitian

Penelitian ini dilakukan melalui beberapa tahapan untuk membangun model klasifikasi audio batuk berbasis ensemble Capsule Network. Tahapan dimulai dari kurasi dataset audio batuk hingga diperoleh dataset balanced-confident, kemudian dilanjutkan dengan pembagian data, preprocessing audio, ekstraksi fitur MFCC13 dan MFCC20, normalisasi fitur, serta pelatihan lima model Capsule Network. Model M1–M4 menggunakan fitur MFCC13, sedangkan M5 menggunakan fitur MFCC20. Pada tahap evaluasi, M5 digunakan sebagai baseline model tunggal terbaik dalam satu kali pengujian, sementara pendekatan utama yang dikaji adalah weighted soft voting ensemble yang menggabungkan prediksi M1–M5 dan dievaluasi melalui sepuluh kali pengujian. Alur umum metode penelitian ditunjukkan pada Gambar 1.



Gambar 1.Diagram Alur Metode Penelitian

2.1 Data dan Representasi Fitur Audio

Data yang digunakan pada penelitian ini berupa rekaman audio batuk dari dataset publik yang memiliki informasi label batuk kering (dry) dan batuk berdahak (wet). Penelitian ini hanya menggunakan dua kelas tersebut karena fokus utama diarahkan pada klasifikasi jenis batuk berdasarkan karakteristik audio, bukan pada penentuan diagnosis penyakit pernapasan. Sebelum digunakan dalam proses pelatihan, data melalui tahap kurasi yang meliputi seleksi label, penghapusan data dengan konflik label, verifikasi keberadaan file audio, serta audit kualitas rekaman. Tahap ini dilakukan untuk memperoleh data yang lebih konsisten dan mengurangi potensi bias evaluasi akibat label yang tidak sesuai atau kualitas audio yang tidak memadai. Berdasarkan hasil seleksi, dari 27.550 metadata awal terdapat 2.348 data yang memiliki label dry atau wet. Setelah proses seleksi dan audit, diperoleh 2.188 data terpilih, terdiri atas 1.678 data dry dan 510 data wet. Karena distribusi tersebut tidak seimbang, penelitian ini membentuk dataset akhir balanced-confident sebanyak 800 audio, yang terdiri atas 400 audio batuk kering dan 400 audio batuk berdahak. Penggunaan dataset seimbang bertujuan untuk mengurangi kecenderungan model terhadap kelas mayoritas serta menjaga objektivitas evaluasi model klasifikasi. Cakupan dataset akhir tersebut juga sesuai dengan batasan penelitian yang menggunakan 800 data audio seimbang untuk dua kategori batuk.

Tabel 1. Distribusi Dataset Penelitian

Kelompok Data	Batuk Kering	Batuk Berdahak	Total
Pelatihan	280	280	560
Validasi	60	60	120
Pengujian	60	60	120
Total	400	400	800

Dataset akhir kemudian dibagi menggunakan skema stratified split dengan rasio 70% data pelatihan, 15% data validasi, dan 15% data pengujian. Pembagian secara stratifikasi digunakan agar proporsi kelas batuk kering dan batuk berdahak tetap seimbang pada setiap

subset data. Sebelum digunakan pada proses pelatihan model, seluruh data terlebih dahulu memalui tahap preprocessing untuk menyeragamkan karakteristik masukan sehingga proses pembelajaran dapat berlangsung secara lebih konsisten.[12] Data pelatihan digunakan untuk membangun model, data validasi digunakan untuk membantu pemilihan konfigurasi dan pembentukan bobot ensemble, sedangkan data pengujian digunakan untuk mengevaluasi performa akhir model. Sebelum dilakukan proses pelatihan model, data perlu melalui tahapan pengumpulan, analisis, dan pengolahan agar informasi yang digunakan memiliki kualitas serta konsistensi yang memadai untuk proses pembelajaran model.[13] Dengan pembagian tersebut, diperoleh 560 data pelatihan, 120 data validasi, dan 120 data pengujian.

Sebelum ekstraksi fitur, setiap audio diproses melalui tahap preprocessing untuk menyeragamkan format sinyal. Tahapan preprocessing meliputi konversi audio ke format WAV apabila diperlukan, perubahan kanal menjadi mono, resampling ke 16 kHz, segmentasi bagian batuk, serta penyeragaman durasi menjadi tiga detik melalui proses padding atau cropping. Tahap ini bertujuan agar seluruh audio memiliki format, frekuensi sampling, dan durasi yang konsisten sebelum digunakan dalam proses ekstraksi fitur. Karena fokus penelitian berada pada pengembangan ensemble Capsule Network, preprocessing ditempatkan sebagai tahap pendukung untuk memastikan kualitas masukan model, bukan sebagai fokus utama pembahasan.

Audio hasil preprocessing kemudian diekstraksi menggunakan Mel-Frequency Cepstral Coefficients (MFCC). MFCC dipilih karena mampu merepresentasikan karakteristik spektral sinyal suara berdasarkan skala Mel yang mendekati persepsi pendengaran manusia, sehingga banyak digunakan dalam pengenalan pola audio dan klasifikasi suara.[5], [6]. Secara umum, proses MFCC mencakup pembagian sinyal audio ke dalam frame pendek, penerapan windowing, transformasi ke domain frekuensi, pemetaan energi ke skala Mel, kompresi logaritmik, dan pembentukan koefisien cepstral. Pada penelitian ini digunakan dua konfigurasi fitur, yaitu MFCC13 dan MFCC20. MFCC13 digunakan pada model M1–M4 untuk menghasilkan representasi fitur yang lebih ringkas, sedangkan MFCC20 digunakan pada model M5 untuk mempertahankan informasi spektral yang lebih detail. Perbedaan jumlah koefisien tersebut dimanfaatkan untuk membentuk variasi representasi fitur pada anggota ensemble.

Fitur MFCC yang dihasilkan kemudian dinormalisasi sebelum digunakan sebagai masukan model Capsule Network. Normalisasi dilakukan untuk menyeragamkan skala nilai antarfitur sehingga proses pelatihan dapat berlangsung lebih stabil dan tidak didominasi oleh fitur dengan rentang nilai tertentu. Hasil akhir dari tahap ini berupa fitur MFCC13 dan MFCC20 yang siap digunakan sebagai masukan bagi lima model Capsule Network. Dengan demikian, tahap data dan representasi fitur audio berperan penting dalam menyediakan masukan yang konsisten sekaligus membentuk variasi fitur yang menjadi dasar penerapan weighted soft voting ensemble pada tahap berikutnya.

2.2 Arsitektur Capsule Network

Arsitektur klasifikasi yang digunakan pada penelitian ini adalah Capsule Network atau CapsNet. CapsNet menggunakan representasi vektor kapsul untuk mempertahankan informasi pola dan hubungan struktural antarfitur, sehingga tidak hanya memperhatikan keberadaan suatu pola, tetapi juga relasi antarbagian fitur masukan. Karakteristik ini relevan untuk pengolahan fitur MFCC karena MFCC merepresentasikan audio dalam bentuk hubungan antara koefisien spektral dan frame waktu. Dengan demikian, fitur MFCC13 dan MFCC20 yang telah dijelaskan pada Subbab 2.1 digunakan sebagai representasi numerik audio batuk yang menjadi masukan bagi model. Beberapa penelitian menunjukkan bahwa CapsNet dapat diterapkan pada klasifikasi audio medis dan audio batuk karena kemampuannya dalam mempertahankan relasi pola fitur yang kompleks.[7], [8]

Pada implementasi penelitian ini, input model berupa tensor MFCC dengan bentuk $(n_mfcc, n_frame, channel)$, yaitu MFCC13 atau MFCC20 dengan satu *channel* fitur dasar. Fitur tersebut diproses menggunakan *Convolutional Layer* berbasis Conv2D untuk menangkap pola lokal antara koefisien MFCC dan frame waktu. Tahap ini dilengkapi *batch normalization*, aktivasi ReLU, *dropout*, dan regularisasi L2 untuk membantu stabilitas pelatihan serta mengurangi risiko *overfitting*. Feature map hasil konvolusi kemudian diubah menjadi *Primary Capsule* melalui *convolution primary capsule* dan proses *reshape* ke bentuk $(jumlah\ kapsul, dimensi\ kapsul)$. Vektor kapsul primer selanjutnya dinormalisasi menggunakan fungsi *squash* agar panjang vektor berada pada rentang 0 hingga mendekati 1, sementara arah vektor tetap mempertahankan informasi fitur.

Tahap berikutnya adalah Class Capsule dengan mekanisme routing-by-agreement. Pada tahap ini, kapsul primer diproyeksikan menjadi prediction vector menuju kapsul kelas, kemudian coupling coefficient diperbarui agar kapsul input memberikan kontribusi lebih besar pada kapsul kelas yang paling sesuai. Klasifikasi ini menggunakan dua kapsul kelas, yaitu batuk kering dan batuk berdahak. Panjang vektor pada setiap kapsul kelas dihitung melalui Length Layer sebagai skor keyakinan model terhadap masing-masing kelas. Model dilatih menggunakan margin loss khas CapsNet karena keluaran model direpresentasikan oleh panjang vektor kapsul, serta optimizer Adam untuk proses pembaruan bobot.

Penelitian ini menggunakan lima konfigurasi CapsNet, yaitu M1 sampai M5, untuk membentuk variasi model penyusun baseline dan ensemble. Model M1–M4 menggunakan MFCC13, sedangkan M5 menggunakan MFCC20. Perbedaan konfigurasi mencakup jumlah routing iterations, dimensi kapsul, nilai dropout, dan stride pada primary capsule. Routing iterations memengaruhi proses pembentukan kesepakatan antar kapsul, dimensi kapsul berkaitan dengan kapasitas representasi fitur, dropout digunakan untuk mengurangi risiko overfitting, sedangkan stride memengaruhi pola ekstraksi fitur. M5 atau kode internal S06 digunakan sebagai baseline terbaik pada pengujian tunggal, sedangkan seluruh model M1–M5 digunakan sebagai penyusun weighted soft voting ensemble. Konfigurasi model ditunjukkan pada Tabel 2.

Tabel 2.Konfigurasi Model Capsule Network penyusun baseline dan ensemble

Model	Fitur	Routing	Kapsul	Dropout
M1	MFCC13	3	16	0,40
M2	MFCC13	2	16	0,40
M3	MFCC13	2	8	0,40
M4	MFCC13	2	8	0,35
M5	MFCC20	2	8	0,35

2.3 Ensemble Weighted Soft Voting

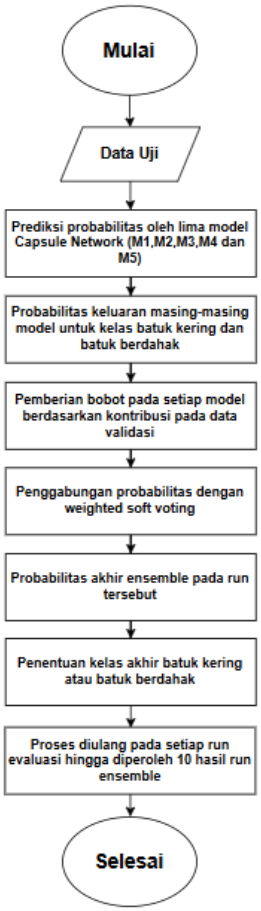
Pendekatan ensemble digunakan untuk menggabungkan keluaran beberapa model sehingga keputusan klasifikasi tidak hanya bergantung pada satu model tunggal. Pada penelitian ini, ensemble diterapkan pada level probabilitas keluaran model, bukan pada level penggabungan fitur. Dengan demikian, setiap model Capsule Network terlebih dahulu menghasilkan probabilitas prediksi terhadap dua kelas, yaitu batuk kering dan batuk berdahak, kemudian probabilitas tersebut digabungkan untuk memperoleh prediksi akhir. Pendekatan soft voting dipilih karena tetap mempertahankan informasi tingkat keyakinan dari setiap model, sedangkan pemberian bobot memungkinkan model dengan kontribusi validasi yang lebih baik memberikan pengaruh lebih besar terhadap keputusan akhir. Strategi ensemble berbasis pembobotan juga telah digunakan dalam penelitian klasifikasi audio untuk memperkuat keputusan model melalui kombinasi beberapa keluaran prediksi.[14] Ensemble pada penelitian ini dibentuk dari lima model Capsule Network, yaitu M1, M2, M3, M4, dan M5. Model M1–M4 menggunakan fitur MFCC13, sedangkan M5 menggunakan fitur MFCC20. Setiap model menghasilkan probabilitas prediksi untuk kedua kelas, kemudian probabilitas tersebut digabungkan secara berbobot. Secara umum, probabilitas akhir kelas $c$untuk data masukan $x$dihitung menggunakan Persamaan (1).

$$P(c \mid x) = \frac{\sum_{i=1}^{5} w_i \, P_i(c \mid x)}{\sum_{i=1}^{5} w_i} \qquad (1)$$

Pada Persamaan (1), $P_i(c \mid x)$ menyatakan probabilitas kelas c yang dihasilkan oleh model ke-i, sedangkan $w_i$menyatakan bobot model ke-i. Kelas akhir ditentukan berdasarkan probabilitas gabungan tertinggi. Dengan mekanisme tersebut, model yang memberikan kontribusi lebih baik pada data validasi dapat diberi bobot lebih besar, sehingga keputusan akhir tidak hanya ditentukan oleh satu model, tetapi oleh kombinasi beberapa model dengan karakteristik yang berbeda. Komposisi model penyusun ensemble mengikuti konfigurasi M1–M5 yang telah ditunjukkan pada Tabel 2.

Weighted soft voting ensemble pada penelitian ini bukan model baru yang dilatih langsung dari awal sebagai satu jaringan tunggal. Ensemble dibentuk setelah masing-masing model penyusun selesai dilatih dan menghasilkan probabilitas prediksi. Pada setiap run evaluasi, lima model M1–M5 dijalankan pada run yang sama, kemudian probabilitas hasil prediksi kelima model tersebut digabungkan untuk membentuk hasil ensemble pada run

tersebut. Ensemble pada penelitian ini dibentuk dari lima model Capsule Network, yaitu M1, M2, M3, M4, dan M5. Model M1–M4 menggunakan representasi MFCC13, sedangkan M5 menggunakan representasi MFCC20. Setiap model menghasilkan probabilitas prediksi terhadap kelas batuk kering dan batuk berdahak. Probabilitas tersebut kemudian digabungkan menggunakan weighted soft voting untuk memperoleh keputusan akhir.Dengan demikian, proses ensemble dilakukan pada tahap fusi keputusan, bukan dengan menggabungkan fitur MFCC maupun parameter jaringan.



Gambar 2. Alur Pembentukan ensemble weighted soft voting pada setiap run evaluasi

Berdasarkan Gambar 2, proses pembentukan ensemble dimulai dengan menjalankan lima model Capsule Network, yaitu M1, M2, M3, M4, dan M5, pada run yang sama. Setiap model menghasilkan probabilitas prediksi terhadap dua kelas, yaitu batuk kering dan batuk berdahak. Probabilitas keluaran masing-masing model kemudian dikalikan dengan bobot yang ditentukan berdasarkan performa pada data validasi. Selanjutnya, seluruh probabilitas berbobot digabungkan menggunakan metode weighted soft voting untuk memperoleh probabilitas akhir ensemble pada run tersebut. Kelas akhir ditentukan berdasarkan probabilitas gabungan terbesar, sehingga setiap data uji diklasifikasikan sebagai batuk kering atau batuk berdahak. Proses ini dilakukan secara berulang pada setiap run evaluasi, mulai dari run pertama hingga run kesepuluh, sehingga diperoleh sepuluh hasil ensemble yang kemudian digunakan untuk menghitung nilai rata-rata dan standar deviasi pada tahap evaluasi.

2.4 Skenario Pengujian dan Evaluasi

Pengujian pada penelitian ini dilakukan melalui dua rancangan utama, yaitu evaluasi model tunggal sebagai baseline dan evaluasi ensemble sebagai pendekatan utama. Model tunggal yang digunakan sebagai baseline adalah Capsule Network berbasis MFCC20. Baseline ini digunakan sebagai pembanding terhadap pendekatan ensemble, sehingga kinerja model tunggal dapat dibandingkan dengan hasil fusi keputusan dari beberapa model. Sementara itu, pendekatan utama menggunakan weighted soft voting ensemble yang menggabungkan lima model Capsule Network, yaitu M1, M2, M3, M4, dan M5.

Vol.7, No. 2, Desember 2026, 607 – 618

Perbedaan antara baseline dan ensemble terletak pada proses evaluasinya. Baseline merupakan model tunggal yang dievaluasi melalui satu kali pengujian, sedangkan ensemble bukan satu jaringan baru yang dilatih dari awal, melainkan metode penggabungan keputusan dari lima model yang telah dilatih. Pada evaluasi ensemble, setiap model penyusun dijalankan sebanyak sepuluh kali menggunakan seed berbeda, yaitu 42 sampai 51. Pada setiap run, model penyusun dari run yang sama digabungkan melalui weighted soft voting, sehingga diperoleh satu keluaran ensemble untuk setiap run evaluasi. Hasil dari sepuluh run kemudian diringkas menggunakan rata-rata dan standar deviasi untuk melihat kestabilan performa terhadap variasi pelatihan.

Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, F1-score, specificity, balanced accuracy, macro-F1, dan Matthews Correlation Coefficient (MCC). Penggunaan beberapa metrik diperlukan agar evaluasi tidak hanya bergantung pada akurasi, tetapi juga mempertimbangkan keseimbangan performa antar kelas dan kecenderungan kesalahan prediksi. Accuracy dan balanced accuracy digunakan untuk menilai performa umum model. Precision dan recall digunakan untuk mengamati kecenderungan kesalahan klasifikasi. F1-score dan macro-F1 digunakan untuk menilai keseimbangan antara precision dan recall, sedangkan specificity digunakan untuk mengukur kemampuan model mengenali kelas negatif. MCC digunakan sebagai ukuran korelasi antara prediksi dan label aktual yang lebih robust pada klasifikasi biner, karena mempertimbangkan seluruh komponen confusion matrix. Penggunaan metrik yang beragam juga sejalan dengan kajian evaluasi klasifikasi yang menekankan bahwa pemilihan metrik perlu disesuaikan dengan karakteristik data dan tujuan evaluasi, bukan hanya mengandalkan akurasi Tunggal.[15] Metrik evaluasi dihitung berdasarkan komponen *confusion matrix*, yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Rumus metrik evaluasi yang digunakan ditunjukkan pada Persamaan (2) sampai Persamaan (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2} \quad (7)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

Macro-F1 dihitung sebagai rata-rata nilai *F1-score* dari setiap kelas, sehingga metrik ini digunakan untuk melihat keseimbangan performa model terhadap kelas batuk kering dan batuk berdahak. Dengan rancangan evaluasi tersebut, *ensemble* diposisikan sebagai pendekatan fusi keputusan yang kompetitif dan lebih kuat secara evaluatif karena diuji secara berulang serta dilaporkan dalam bentuk rata-rata dan standar deviasi. Namun, penelitian ini tidak menempatkan *ensemble* sebagai model yang paling unggul secara mutlak. Grafik perbandingan performa, confusion matrix, dan kurva pelatihan tidak disajikan pada subbab metode, melainkan digunakan pada bagian hasil dan pembahasan untuk mendukung analisis performa model.

3. Hasil dan Pembahasan

3.1 Hasil Evaluasi Model Tunggal dan Ensemble

Evaluasi dilakukan untuk membandingkan performa model tunggal MFCC20-CapsNet sebagai baseline dengan pendekatan ensemble weighted soft voting yang menggabungkan lima model Capsule Network M1–M5. Model tunggal MFCC20-CapsNet dievaluasi melalui satu kali pengujian (single run), sedangkan ensemble diperoleh dari penggabungan probabilitas

prediksi lima model penyusun pada sepuluh kali pengujian. Perbedaan skema evaluasi ini perlu diperhatikan agar interpretasi hasil dilakukan secara proporsional, karena baseline merepresentasikan performa satu model tunggal pada satu kali pengujian, sedangkan ensemble merepresentasikan rata-rata performa dari evaluasi berulang.

Tabel 3.Ringkasan Performa baseline dan ensemble

Pendekatan	Accuracy	Balanced Accuracy	Macro-F1	MCC
Model tunggal MFCC20-CapsNet	71,67%	71,67%	71,60%	0,4355
Ensemble weighted soft voting M1–M5	70,58%	70,58%	70,32%	0,4195

Berdasarkan Tabel 3, model tunggal MFCC20-CapsNet memperoleh *accuracy* dan *balanced accuracy* sebesar 71,67%, *macro-F1* sebesar 71,60%, serta MCC sebesar 0,4355. Sementara itu, ensemble weighted soft voting memperoleh rata-rata *accuracy* dan *balanced accuracy* sebesar 70,58%, *macro-F1* sebesar 70,32%, serta MCC sebesar 0,4195. Model tunggal juga menghasilkan *recall* dan *F1-score* yang lebih tinggi, yaitu 66,67% dan 70,18%, sedangkan ensemble memiliki *precision* dan *specificity* yang sedikit lebih tinggi, masing-masing sebesar 74,28% dan 76,83%. Hasil tersebut menunjukkan bahwa ensemble memiliki performa yang kompetitif, meskipun belum melampaui model tunggal pada sebagian besar metrik evaluasi.

Untuk memperjelas proses pembentukan ensemble, dilakukan pula evaluasi terhadap lima model penyusun pada sepuluh kali pengujian. Hasil evaluasi menunjukkan bahwa model M1 memperoleh balanced accuracy rata-rata 71,50%, M2 sebesar 71,67%, M3 sebesar 71,17%, M4 sebesar 69,08%, dan M5 sebesar 67,75%. Setelah probabilitas dari kelima model tersebut digabungkan melalui weighted soft voting pada setiap run yang sama, ensemble memperoleh balanced accuracy rata-rata sebesar 70,58%.

Tabel 4. Perbandingan rata-rata performa 10 run model penyusun dan ensemble

Model	Fitur	Balanced Accuracy	Macro-F1	MCC
M1	MFCC13	71,50%	71,35%	0,4350
M2	MFCC13	71,67%	71,52%	0,4384
M3	MFCC13	71,17%	70,96%	0,4297
M4	MFCC13	69,08%	68,81%	0,3863
M5	MFCC20	67,75%	66,88%	0,3666

Berdasarkan Tabel 4, beberapa model penyusun, khususnya M1, M2, dan M3, memiliki rata-rata balanced accuracy yang sedikit lebih tinggi dibandingkan ensemble. Oleh karena itu, ensemble tidak dapat diklaim sebagai pendekatan dengan nilai performa tertinggi secara mutlak. Namun, ensemble tetap memiliki nilai metodologis karena berperan sebagai fusi keputusan dari lima model dengan karakteristik berbeda. Dengan demikian, pendekatan ensemble lebih tepat diposisikan sebagai metode fusi keputusan yang kompetitif dan lebih kuat secara evaluatif karena diuji melalui pengujian berulang, meskipun tidak unggul secara mutlak terhadap model tunggal maupun beberapa model penyusun pada seluruh metrik.

3.2 Analisis Kestabilan Ensemble pada 10 kali pengujian

Analisis kestabilan dilakukan untuk melihat konsistensi performa ensemble weighted soft voting ketika model penyusunnya dijalankan dengan seed berbeda. Sebelum membahas kestabilan ensemble, perlu dijelaskan bahwa setiap model penyusun, yaitu M1, M2, M3, M4, dan M5, terlebih dahulu dievaluasi sebanyak sepuluh kali. Ensemble kemudian dibentuk dengan menggabungkan probabilitas prediksi dari kelima model pada run yang sama. Dengan demikian, kestabilan ensemble perlu dibaca sebagai kelanjutan dari kestabilan model-model penyusunnya, bukan sebagai hasil dari satu model tunggal yang berdiri sendiri.

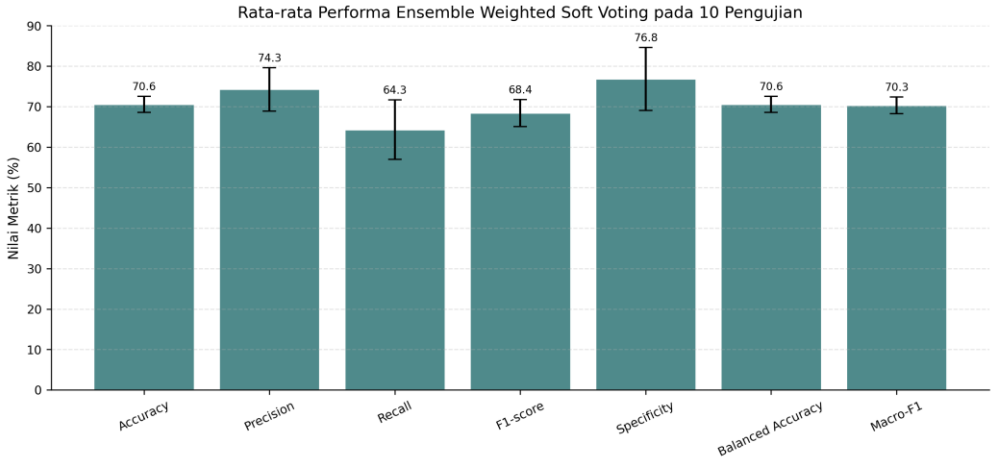
Berdasarkan hasil pada tabel 5, model M2 memperoleh rata-rata balanced accuracy tertinggi sebesar 71,67%, sedangkan M5 memperoleh rata-rata terendah sebesar 67,75%. Ensemble memperoleh rata-rata balanced accuracy sebesar 70,58%. Hasil ini menunjukkan bahwa ensemble memiliki performa kompetitif, tetapi bukan nilai rata-rata tertinggi di antara seluruh model penyusun. Oleh karena itu, kontribusi ensemble lebih tepat diposisikan sebagai pendekatan fusi keputusan yang menggabungkan beberapa model, bukan sebagai pendekatan yang paling unggul secara mutlak pada seluruh metrik. Data perbandingan 10 run menunjukkan bahwa setiap model penyusun memiliki nilai rata-rata, standar deviasi, serta rentang performa yang berbeda sebelum digabungkan melalui weighted soft voting.

Tabel 5.Ringkasan Kestabilan performa ensemble pada 10 kali pengujian

Metrik	Mean	Std	Min	Max
Balanced Accuracy	70,58%	1,97%	67,50%	73,33%
Macro-F1	70,32%	2,08%	67,50%	73,15%

MCC	0,4195	0,0407	0,3500	0,4733
Recall	64,33%	7,34%	46,67%	75,00%
Specificity	76,83%	7,80%	66,67%	91,67%
Threshold	0,5480	0,0278	0,5100	0,6000

Pada Tabel 5 menunjukkan bahwa ensemble diuji sebanyak sepuluh kali menggunakan seed 42 sampai 51. Rata-rata balanced accuracy yang diperoleh adalah 70,58% dengan standar deviasi 1,97%. Nilai minimum balanced accuracy sebesar 67,50%, sedangkan nilai maksimum sebesar 73,33%. Run terbaik berdasarkan balanced accuracy diperoleh pada run ke-3, sementara run terendah terjadi pada run ke-6. Selain itu, threshold prediksi memiliki rata-rata 0,5480 dengan standar deviasi 0,0278, sehingga nilai ambang keputusan relatif stabil pada sepuluh kali pengujian.



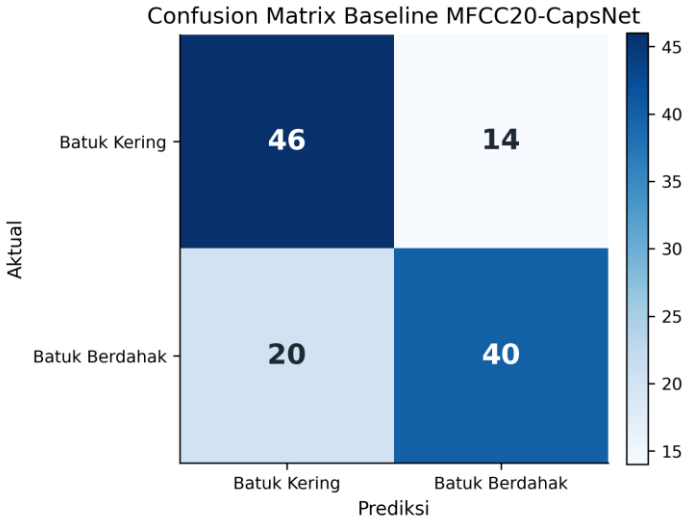
Gambar 5. Rata-rata dan standar deviasi metrik ensemble pada 10 kali pengujian

Berdasarkan Gambar 5, variasi balanced accuracy dan macro-F1 relatif lebih kecil dibandingkan recall dan specificity. Standar deviasi balanced accuracy yang relatif kecil menunjukkan bahwa performa umum ensemble masih terkendali pada evaluasi berulang. Namun, fluktuasi pada recall dan specificity menunjukkan bahwa ensemble belum sepenuhnya menghilangkan variasi akibat proses pelatihan model penyusun. Kondisi ini dapat terjadi karena setiap run menggunakan seed berbeda, sehingga proses inisialisasi, pembelajaran bobot, dan keluaran probabilitas model penyusun dapat menghasilkan pola prediksi yang tidak sepenuhnya identik.

Secara keseluruhan, hasil evaluasi berulang menunjukkan bahwa ensemble weighted soft voting memiliki kestabilan yang cukup baik pada metrik umum seperti balanced accuracy dan macro-F1. Meskipun demikian, variasi pada recall dan specificity menunjukkan bahwa performa pada masing-masing kelas masih dapat berubah antar run. Dengan demikian, evaluasi sepuluh kali pengujian memberikan gambaran performa yang lebih representatif dibandingkan pengujian tunggal, karena mampu memperlihatkan rata-rata performa, rentang variasi, serta kestabilan ensemble. Detail setiap run menunjukkan adanya perbedaan nilai metrik, confusion matrix, dan threshold pada masing-masing seed

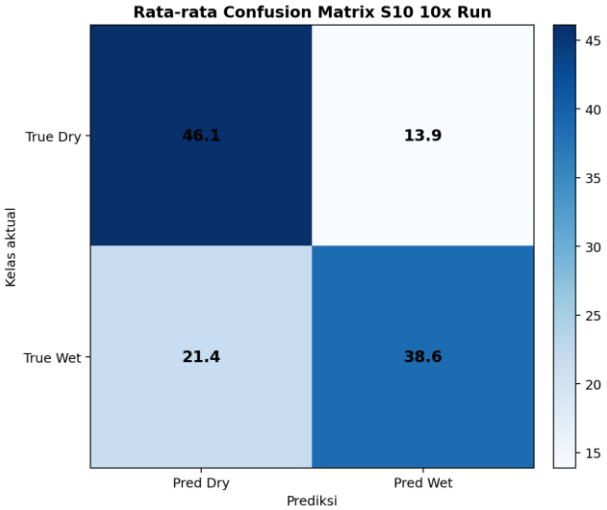
3.3 Analisis Confusion Matrix dan Karakteristik Kesalahan Klasifikasi

Analisis kesalahan klasifikasi dilakukan menggunakan confusion matrix untuk melihat pola prediksi benar dan salah pada kelas batuk kering dan batuk berdahak. Pada baseline model tunggal MFCC20-CapsNet, hasil pengujian menunjukkan 46 data batuk kering diklasifikasikan dengan benar sebagai batuk kering, sedangkan 14 data batuk kering salah diklasifikasikan sebagai batuk berdahak. Pada kelas batuk berdahak, terdapat 40 data yang diklasifikasikan dengan benar sebagai batuk berdahak, sedangkan 20 data lainnya salah diklasifikasikan sebagai batuk kering.



Gambar 6. Confusion Matrix baseline MFCC20-CapsNet pada data uji

Berdasarkan Gambar 6, baseline MFCC20-CapsNet memiliki nilai specificity sebesar 76,67% dan recall sebesar 66,67%. Nilai tersebut menunjukkan bahwa model relatif lebih baik dalam mengenali kelas batuk kering dibandingkan mendeteksi seluruh sampel batuk berdahak. Kesalahan false negative menunjukkan data batuk berdahak yang diprediksi sebagai batuk kering, sedangkan false positive menunjukkan data batuk kering yang diprediksi sebagai batuk berdahak. Dalam konteks klasifikasi audio batuk, pola kesalahan ini mengindikasikan bahwa sebagian karakteristik audio batuk berdahak masih dapat menyerupai batuk kering pada representasi fitur MFCC, terutama apabila energi, durasi, atau pola spektral batuk tidak menunjukkan perbedaan yang tegas.



Gambar 7. Rata-rata confusion matrix ensemble weighted soft voting M1-M5 pada 10 kali pengujian

Pada Gambar 7 menunjukkan rata-rata confusion matrix ensemble dari sepuluh kali pengujian. Secara rata-rata, ensemble menghasilkan 46,10 data batuk kering yang diklasifikasikan benar sebagai batuk kering dan 13,90 data batuk kering yang salah diklasifikasikan sebagai batuk berdahak. Pada kelas batuk berdahak, ensemble menghasilkan 38,60 data yang diklasifikasikan benar sebagai batuk berdahak dan 21,40 data yang salah diklasifikasikan sebagai batuk kering. Nilai ini merupakan agregat dari sepuluh confusion matrix, sehingga tidak merepresentasikan satu run tunggal.

Tabel 6. Agregat komponen confusion matrix ensemble pada 10 kali pengujian

Komponen	Mean per Run	Std per Run	Min	Max
TN	46,10	4,44	40	55
FP	13,90	4,44	5	20
FN	21,40	4,18	15	32
TP	38,60	4,18	28	45

Vol.7, No. 2, Desember 2026, 607 – 618

Berdasarkan Tabel 6, ensemble cenderung mempertahankan kemampuan yang cukup baik dalam mengenali kelas batuk kering, terlihat dari rata-rata TN yang lebih tinggi dibandingkan FP. Namun, nilai FN yang masih cukup besar menunjukkan bahwa deteksi terhadap kelas batuk berdahak masih menjadi area yang perlu diperbaiki. Pola ini sejalan dengan nilai recall ensemble yang lebih rendah dibandingkan specificity, sehingga dapat disimpulkan bahwa sebagian sampel batuk berdahak masih lebih sering bergeser ke kelas batuk kering. Secara ilmiah, kondisi ini dapat terjadi karena karakteristik akustik batuk berdahak tidak selalu muncul secara konsisten pada rekaman audio, sehingga sebagian pola spektralnya masih tumpang tindih dengan batuk kering.

Secara keseluruhan, analisis confusion matrix menunjukkan bahwa baik baseline maupun ensemble masih menghadapi tantangan utama pada pemisahan kelas batuk berdahak. Ensemble memberikan keputusan yang lebih representatif karena diperoleh dari gabungan lima model dan dirata-ratakan pada sepuluh kali pengujian, tetapi pendekatan tersebut belum sepenuhnya menghilangkan kesalahan klasifikasi pada kelas batuk berdahak. Oleh karena itu, peningkatan kualitas representasi fitur, strategi threshold, atau pendekatan fusi keputusan lanjutan masih berpotensi dilakukan untuk memperbaiki sensitivitas model terhadap batuk berdahak

3.4 Pembahasan dan Keterbatasan

Hasil penelitian menunjukkan bahwa model tunggal MFCC20-CapsNet masih memperoleh performa sedikit lebih tinggi dibandingkan ensemble weighted soft voting M1–M5 pada sebagian besar metrik evaluasi. Meskipun demikian, selisih balanced accuracy, macro-F1, dan MCC antara kedua pendekatan relatif kecil. Di sisi lain, ensemble menghasilkan precision dan specificity yang sedikit lebih tinggi serta menunjukkan performa yang lebih konsisten karena diperoleh dari rata-rata sepuluh kali pengujian. Hasil ini menunjukkan bahwa penggabungan beberapa model mampu meningkatkan kestabilan prediksi, meskipun belum memberikan peningkatan performa yang signifikan dibandingkan model tunggal. Analisis confusion matrix menunjukkan bahwa kedua pendekatan lebih baik dalam mengenali batuk kering dibandingkan batuk berdahak. Hal ini ditunjukkan oleh nilai specificity yang lebih tinggi daripada recall, serta jumlah false negative yang masih lebih besar dibandingkan false positive. Kondisi tersebut mengindikasikan bahwa sebagian sampel batuk berdahak masih diprediksi sebagai batuk kering, yang kemungkinan dipengaruhi oleh kemiripan karakteristik spektral kedua jenis batuk pada representasi fitur MFCC. Penelitian ini masih memiliki beberapa keterbatasan, yaitu penggunaan dataset yang terbatas pada dua kelas batuk dengan jumlah data yang relatif kecil serta evaluasi yang hanya menggunakan arsitektur Capsule Network berbasis MFCC13 dan MFCC20. Oleh karena itu, penelitian selanjutnya dapat mengeksplorasi kombinasi representasi fitur lain, strategi ensemble yang berbeda, atau penambahan variasi dataset agar kemampuan generalisasi model terhadap karakteristik audio batuk yang lebih beragam dapat ditingkatkan.

4. Kesimpulan

Penelitian ini mengembangkan pendekatan ensemble weighted soft voting berbasis Capsule Network dengan representasi fitur MFCC13 dan MFCC20 untuk klasifikasi audio batuk kering dan batuk berdahak. Dataset yang digunakan merupakan subset balanced-confident sebanyak 800 audio yang terdiri atas 400 audio batuk kering dan 400 audio batuk berdahak. Hasil penelitian menunjukkan bahwa model tunggal MFCC20-CapsNet memperoleh performa sedikit lebih tinggi pada pengujian single run dengan accuracy dan balanced accuracy sebesar 71,67%, sedangkan ensemble weighted soft voting M1–M5 memperoleh accuracy dan balanced accuracy rata-rata sebesar 70,58% dari sepuluh kali pengujian. Hasil tersebut menunjukkan bahwa ensemble belum memberikan peningkatan performa mutlak terhadap baseline tunggal, tetapi tetap menunjukkan performa yang kompetitif dan memberikan evaluasi yang lebih kuat melalui pengujian berulang. Analisis kesalahan klasifikasi menunjukkan bahwa model masih menghadapi tantangan dalam mengenali sebagian sampel batuk berdahak yang memiliki karakteristik akustik tumpang tindih dengan batuk kering. Penelitian selanjutnya dapat diarahkan pada penggunaan dataset yang lebih luas dan beragam, validasi eksternal pada dataset berbeda, eksplorasi fitur audio tambahan, serta pengembangan strategi fusi keputusan atau threshold yang lebih adaptif untuk meningkatkan sensitivitas terhadap kelas batuk berdahak.

Referensi

[1] A. H. Morice *et al.*, “ERS guidelines on the diagnosis and treatment of chronic cough in adults and children,” *Eur Respir J*, vol. 55, no. 1, p. 1901136, Jan. 2020, doi: 10.1183/13993003.01136-2019.
[2] H. Wiltingh, J. M. Marchant, and V. Goyal, “Cough in Protracted Bacterial Bronchitis and Bronchiectasis,” *JCM*, vol. 13, no. 11, p. 3305, Jun. 2024, doi: 10.3390/jcm13113305.
[3] M. Loey and S. Mirjalili, “COVID-19 cough sound symptoms classification from scalogram image representation using deep learning models,” *Computers in Biology and Medicine*, vol. 139, p. 105020, Dec. 2021, doi: 10.1016/j.compbiomed.2021.105020.
[4] M. Pahar et al., “Automatic Tuberculosis and COVID-19 cough classification using deep learning,” in 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), Jul. 2022, pp. 1–9. doi: 10.1109/ICECET55527.2022.9873469.
[5] B. Kanata, S. M. A. Sasongko, and M. A. Ali, “Enhancing Music Genres Classification with MFCC and CNN”.
[6] M. Septiyanto, E. B. Susanto, and D. Sugianti, “Improving Indonesian Speech Emotion Classification Using Mfcc And Bilstm With Audio Augmentation,” jiko, vol. 8, no. 3, pp. 244–251, Dec. 2025, doi: 10.33387/jiko.v8i3.10820.
[7] A. Muhammad, M. A. Arserim, and Ö. Türk, “Compare the classification performances of convolutional neural networks and capsule networks on the Coswara dataset,” *Dicle University Journal of Engineering*, vol. 14, no. 2, pp. 265–271, 2023, doi: 10.24012/dumf.1270429.
[8] S. J. S. Rajasekar et al., “Detection of tuberculosis using cough audio analysis: a deep learning approach with capsule networks,” Discov Artif Intell, vol. 4, no. 1, p. 77, Nov. 2024, doi: 10.1007/s44163-024-00179-4.
[9] B. T. Balamurali et al., “Deep Neural Network-Based Respiratory Pathology Classification Using Cough Sounds,” Sensors, vol. 21, no. 16, p. 5555, Aug. 2021, doi: 10.3390/s21165555.
[10] P. Chen, J. Zhang, and A. Mashhadi, “Advanced music classification using a combination of capsule neural network by upgraded ideal gas molecular movement algorithm,” Sci Rep, vol. 14, no. 1, p. 30863, Dec. 2024, doi: 10.1038/s41598-024-81700-8.
[11] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, “An Ensemble of Convolutional Neural Networks for Audio Classification,” Applied Sciences, vol. 11, no. 13, p. 5796, Jun. 2021, doi: 10.3390/app11135796.
[12] M. J. D. Deke, F. Muttaqin, and F. T. Anggraeny, “Implementasi Pix2Pix Conditional GAN dan LMS Color Space untuk Koreksi Warna pada Penderita Buta Warna,” vol. 6, 2026.
[13] R. Saputra, S. Murtia, and N. Susanti, “Sistem Pakar Diagnosa Penyakit Ikan Menggunakan Metode Forward Chaining,” vol. 6, no. 2, 2025.
[14] M. Awais, A. Bhuva, D. Bhuva, S. Fatima, and T. Sadiq, “Optimized DEC: An effective cough detection framework using optimal weighted Features-aided deep Ensemble classifier for COVID-19,” Biomedical Signal Processing and Control, vol. 86, p. 105026, Sep. 2023, doi: 10.1016/j.bspc.2023.105026.
[15] L. Ferrer, “Analysis and Comparison of Classification Metrics,” Sep. 20, 2023, arXiv: arXiv:2209.05355. doi: 10.48550/arXiv.2209.05355.