Digital Transformation Technology (Digitech) | e-ISSN: 2807-9000
Volume 6, **Number** 1, Maret 2026
https://doi.org/10.47709/digitech.v6i1.7922

Penerapan Convolutional Neural Network untuk Klasifikasi Jenis Suara Vokal Paduan Suara Berdasarkan Fitur Akustik

Arif Harbani[1*], F.R. Dwi Febriantoro[2], Joko Sarjanoko[3], Syafira Amatur Rahmi[4]
[1,2,3,4]Universitas Binaniaga Indonesia, Indonesia.
[1]arifharbani@unbin.ac.id, [2]dwifebriantoro@unbin.ac.id, [3]joko@unbin.ac.id, [4]syafirafiraafir@gmail.com

open access
Histori Artikel:
Diajukan: 29 Januari 2026
Disetujui: 18 Februari 2026
Dipublikasi: 20 Februari 2026
Kata Kunci:
Convolutional Neural Network ; MobileNetV2; MFCC; Paduan Suara; Klasifikasi Vokal .

Abstrak
Paduan suara merupakan entitas seni vokal kompleks yang mengandalkan keselarasan antara kategori vokal (Sopran, Alto, Tenor, Bass) untuk mencapai harmoni optimal. Namun, klasifikasi suara yang dilakukan secara manual oleh pelatih seringkali terhambat oleh subjektivitas perseptual dan inefisiensi waktu. Penelitian ini bertujuan untuk mengimplementasikan pendekatan *Machine Learning* berbasis *Convolutional Neural Network* (CNN) dengan arsitektur MobileNetV2 guna mengotomatisasi klasifikasi vokal secara objektif. Metodologi yang digunakan adalah *Research and Development* (R&D) dengan ekstraksi fitur akustik *Mel-Frequency Cepstral Coefficients* (MFCC). Sinyal audio diproses dengan sampling rate 22.050 Hz dan dikonversi menjadi citra spektrogram 224x224 piksel untuk memenuhi standar input MobileNetV2. Hasil eksperimen pada dataset vokal wanita (51 Sopran, 44 Alto) menunjukkan tingkat akurasi sebesar 78,1%, dengan nilai Precision 85%, Recall 64,2%, dan F1-Score 73,2%. Efisiensi komputasi MobileNetV2 melalui *Inverted Residual Blocks* dan *Linear Bottlenecks* (Sandler et al., 2018) memungkinkan inferensi cepat pada backend Flask. Evaluasi kebergunaan melalui kuesioner PSSUQ (Lewis, 1995) menghasilkan skor kepuasan keseluruhan sebesar 83,56%, yang menempatkan sistem dalam kategori "Sangat Efektif". Meskipun terdapat tantangan pada nilai *Recall* akibat kemiripan fitur spektral pada zona transisi vokal, sistem ini terbukti mampu mentransformasi paradigma klasifikasi dari berbasis intuisi *(intuition-driven*) menjadi berbasis data (*data-driven)*, yang secara signifikan mereduksi waktu persiapan komposisi paduan suara.

PENDAHULUAN

Paduan suara merupakan salah satu bentuk ekspresi seni vokal yang paling kompleks, di mana sekumpulan penyanyi dengan karakteristik suara yang berbeda-beda dipadukan untuk menciptakan harmoni musikal yang kaya dan seimbang. Dalam struktur paduan suara modern, pembagian jenis suara umumnya diklasifikasikan ke dalam empat kategori utama yang dikenal sebagai SATB, yaitu Sopran dan Alto untuk kelompok suara wanita, serta Tenor dan Bass untuk kelompok suara pria, berdasarkan rentang frekuensi vokal dan karakteristik timbre masing-masing suara (Kim et al., 2021). Setiap kategori ini memiliki rentang nada (ambitus) dan warna suara (*timbre*) yang spesifik (Liu et al. (2023). Ketepatan dalam menempatkan seorang penyanyi ke dalam kategori yang sesuai bukan hanya masalah teknis, melainkan fondasi utama dalam membentuk kualitas estetika pertunjukan.

Namun, dalam praktiknya, proses klasifikasi suara sering kali menghadapi tantangan signifikan. Kesalahan dalam penempatan jenis suara dapat berdampak fatal, seperti ketidakseimbangan dinamika di mana satu bagian suara mendominasi bagian lainnya, atau yang lebih buruk, risiko cedera pita suara bagi penyanyi yang dipaksa menyanyi di luar jangkauan nyamannya. Selama ini, metode yang digunakan oleh pelatih paduan suara bersifat konvensional dan subjektif. Pelatih harus mendengarkan satu per satu anggota melalui serangkaian tes vokal (*vocal run*) untuk menentukan kategori suara mereka. Meskipun metode ini memiliki nilai artistik, ia sangat bergantung pada kepekaan pendengaran dan kondisi fisik pelatih pada saat itu (Santoso et al., 2023).

Permasalahan utama muncul ketika metode manual ini diterapkan pada kelompok paduan suara dengan jumlah anggota yang besar atau dalam proses audisi massal. Proses ini menjadi tidak efisien secara waktu dan tenaga. Subjektivitas pelatih juga menjadi faktor risiko, faktor kelelahan atau lingkungan akustik ruang audisi dapat mempengaruhi penilaian, sehingga hasil klasifikasi bisa menjadi tidak konsisten. Ketidakefisienan ini berdampak langsung pada manajemen waktu latihan, di mana waktu yang seharusnya digunakan untuk membedah materi lagu justru habis digunakan untuk proses administratif penentuan kategori vokal.

Seiring dengan kemajuan teknologi digital, pengolahan sinyal suara kini dapat dilakukan dengan pendekatan komputasi yang lebih presisi. Suara manusia memiliki karakteristik akustik utama yang meliputi

pitch sebagai representasi frekuensi dasar, timbre sebagai karakteristik spektral yang membedakan warna suara, serta intensitas yang merefleksikan kekuatan energi sinyal suara, yang secara luas digunakan dalam analisis dan klasifikasi sinyal audio modern (Singh et al., 2022). Fitur-fitur inilah yang membedakan antara seorang Sopran yang memiliki suara terang dan tinggi dengan seorang Alto yang memiliki karakteristik suara lebih rendah dan tebal. Namun, menganalisis fitur-fitur ini secara manual melalui pengamatan gelombang suara sangatlah sulit bagi manusia biasa.

Untuk mengatasi kendala tersebut, penelitian ini menerapkan pendekatan deep learning berbasis Convolutional Neural Network (CNN) dengan mentransformasikan sinyal audio ke dalam representasi dua dimensi, seperti spektrogram atau Mel-Frequency Cepstral Coefficients (MFCC), sehingga memungkinkan ekstraksi fitur spasial pada domain waktu–frekuensi dan meningkatkan akurasi klasifikasi suara (Zhang et al., 2023).

Penelitian ini memanfaatkan arsitektur MobileNetV2 sebagai Convolutional Neural Network ringan yang memungkinkan implementasi sistem klasifikasi suara pada platform berbasis web atau perangkat dengan sumber daya terbatas, tanpa mengorbankan akurasi. Pendekatan ini diharapkan mampu menghasilkan proses penentuan jenis suara paduan suara yang lebih cepat, objektif, dan terukur secara ilmiah. Meskipun model berbasis CNN dan RNN telah banyak digunakan dalam pengolahan sinyal audio, penerapannya pada klasifikasi jenis suara vokal paduan suara, khususnya pada domain suara wanita (Sopran dan Alto), masih terbatas. Sebagian besar penelitian sebelumnya berfokus pada klasifikasi audio generik dan belum mempertimbangkan karakteristik khas paduan suara serta integrasi hasil klasifikasi ke dalam sistem pendukung keputusan berbasis web yang praktis bagi pelatih non-teknis.

Di sisi lain, belum banyak studi yang mengevaluasi efektivitas arsitektur *hybrid* CNN–RNN dalam menangkap pola spasial (spektral) dan pola temporal (dinamika vokal) secara simultan pada data vokal paduan suara, serta menguji manfaat praktisnya melalui evaluasi usability berbasis pengguna akhir.

Berdasarkan kesenjangan tersebut, penelitian ini memberikan kontribusi ilmiah sebagai berikut:

1. Kontribusi domain-spesifik, yaitu penerapan awal pendekatan deep learning untuk klasifikasi suara vokal paduan suara wanita (Sopran–Alto) dalam konteks pelatihan paduan suara di pendidikan tinggi di Indonesia.
2. Validasi arsitektur hybrid CNN–RNN, dengan mengombinasikan MobileNetV2 dan Bi-LSTM yang terbukti meningkatkan performa klasifikasi, khususnya dalam menangkap dinamika temporal suara vokal.
3. Pengembangan sistem AI berbasis web yang *usable*, mencakup pemrosesan audio, inferensi model, dan antarmuka pengguna, serta dievaluasi menggunakan PSSUQ.
4. Kontribusi praktis bagi manajemen paduan suara, melalui percepatan proses klasifikasi dan penyediaan dasar objektif berbasis data untuk mendukung keputusan pelatih.

Tujuan utama dari penelitian ini adalah untuk mengevaluasi sejauh mana model CNN dapat mengenali perbedaan fitur akustik antara jenis suara Sopran dan Alto, serta membangun sebuah prototipe sistem yang dapat membantu pelatih paduan suara dalam melakukan klasifikasi secara instan. Hasil penelitian ini diharapkan memberikan kontribusi nyata bagi dunia seni musik dan teknologi informasi, khususnya dalam digitalisasi manajemen paduan suara di Indonesia.

Seiring berkembangnya deep learning dalam pengolahan sinyal audio, berbagai penelitian terkini menunjukkan bahwa pendekatan berbasis Convolutional Neural Network (CNN) semakin dominan untuk tugas klasifikasi suara dan musik. Penelitian terkini membuktikan bahwa pemanfaatan spektrogram dan MFCC sebagai input CNN memberikan performa klasifikasi audio yang lebih unggul dibandingkan pendekatan konvensional berbasis fitur statistik (Zhang et al., 2023). Selain itu, arsitektur CNN ringan seperti MobileNet dan EfficientNet mulai banyak digunakan karena efisiensinya yang memungkinkan implementasi pada sistem real-time dan perangkat dengan sumber daya terbatas (Imam & Suhartono, 2024).

Dalam konteks audio berbasis waktu, sejumlah penelitian mutakhir juga menekankan pentingnya pemodelan dimensi temporal. Kombinasi CNN dengan Recurrent Neural Network (RNN), khususnya Long Short-Term Memory (LSTM) dan Bidirectional LSTM, terbukti mampu meningkatkan performa pada tugas-tugas seperti speech emotion recognition, speaker identification, dan music tagging karena kemampuannya menangkap dinamika sinyal dari waktu ke waktu (Zhang et al., 2023). Pendekatan hybrid CNN–RNN ini dipandang sebagai state of the art untuk pemrosesan audio kompleks yang tidak hanya bergantung pada pola spasial, tetapi juga pada transisi temporal antar frame suara.

Namun demikian, telaah terhadap literatur terbaru menunjukkan bahwa mayoritas penelitian tersebut masih berfokus pada domain audio umum seperti pengenalan ujaran, klasifikasi emosi suara, atau genre musik, dengan dataset berskala besar dan konteks penggunaan yang bersifat generik. Penerapan deep learning untuk klasifikasi jenis suara vokal paduan suara (choral voice classification), khususnya pada suara wanita (Sopran dan Alto), masih relatif jarang dibahas secara spesifik dalam lima tahun terakhir. Penelitian yang ada umumnya berhenti pada evaluasi performa model secara teknis, tanpa mengintegrasikan hasil klasifikasi ke dalam sistem berbasis web yang siap digunakan oleh pelatih sebagai alat bantu pengambilan keputusan (decision support

system).

Selain itu, studi-studi terkini juga menunjukkan bahwa aspek usability dan penerimaan pengguna akhir masih sering terabaikan dalam pengembangan sistem AI berbasis audio. Padahal, evaluasi berbasis pengguna seperti Post-Study System Usability Questionnaire (PSSUQ) direkomendasikan dalam penelitian sistem cerdas modern untuk memastikan bahwa solusi yang dihasilkan tidak hanya akurat secara algoritmik, tetapi juga efektif dan mudah digunakan dalam praktik nyata (Santoso et al., 2023).

Berdasarkan posisi *state of the art* tersebut, penelitian ini menempati celah riset yang jelas, yaitu mengintegrasikan pendekatan hybrid CNN–RNN terkini dengan arsitektur ringan MobileNetV2 untuk klasifikasi suara vokal paduan suara wanita, sekaligus mengimplementasikannya dalam sistem berbasis web yang dievaluasi dari sisi kinerja model dan kebergunaan pengguna. Dengan demikian, penelitian ini tidak hanya melanjutkan tren mutakhir deep learning audio, tetapi juga memperluas kontribusinya ke domain spesifik paduan suara dan aspek implementatif yang selama ini masih terbatas.

STUDI LITERATUR

LANDASAN TEORI

Deep Learning dan Representasi Data *Audio Deep learning* merupakan sub-bidang dari machine learning yang menggunakan jaringan syaraf tiruan (*neural networks*) dengan banyak lapisan tersembunyi untuk mempelajari representasi data secara otomatis (Zhang et al., 2023). Dalam konteks klasifikasi suara, *deep learning* menawarkan keunggulan dibandingkan metode statistik tradisional karena kemampuannya untuk mengekstraksi fitur kompleks secara hierarkis (Imam & Suhartono, 2024) langsung dari data mentah. Fenomena ini menggeser paradigma lama yang mengandalkan rekayasa fitur manual (*hand-crafted features*) menuju pemrosesan fitur yang dipelajari secara mandiri oleh model, yang terbukti lebih tangguh terhadap variasi derau (*noise*) dan perbedaan karakteristik vokal individu penyanyi.

Fitur Akustik dan Karakteristik Vokal

Suara manusia dihasilkan oleh getaran pita suara yang dimodulasi oleh saluran vokal. Dalam paduan suara, klasifikasi vokal didasarkan pada parameter fisik yang dapat diukur (Liu et al. (2023):

1. *Pitch* (Tinggi Nada):
   Persepsi pendengaran terhadap frekuensi dasar (f0). Penyanyi Sopran memiliki rentang frekuensi yang lebih tinggi (247 Hz hingga 1046 Hz) dibandingkan Alto (164 Hz hingga 698 Hz).
2. *Timbre* (Warna Suara):
   Kualitas suara yang membedakan dua instrumen atau suara manusia meskipun berada pada nada yang sama. Timbre ditentukan oleh struktur harmonik dan selubung spektral.
3. Intensitas:
   Berkaitan dengan tekanan suara dan energi yang dikeluarkan, yang sering kali direpresentasikan dalam skala logaritmik desibel (dB).

Mel-Frequency Cepstral Coefficients (MFCC)

Karena sinyal audio bersifat satu dimensi dan temporal, diperlukan transformasi untuk mengubahnya menjadi format yang dapat diproses oleh algoritma pengenalan pola gambar. MFCC adalah metode ekstraksi fitur yang merepresentasikan selubung spektral suara dengan meniru cara kerja sistem pendengaran manusia (Singh et al. (2022). Proses ini melibatkan transformasi Fourier untuk mengubah sinyal waktu menjadi frekuensi, diikuti oleh pemetaan ke skala Mel (frekuensi yang dirasakan manusia) dan transformasi kosinus diskrit. Hasil akhirnya adalah citra dua dimensi yang disebut spektrogram MFCC, yang menyimpan informasi unik mengenai tekstur suara penyanyi.

Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) merupakan arsitektur jaringan saraf tiruan yang dirancang untuk memproses data berbentuk grid dua dimensi, seperti citra dan representasi spektral audio, melalui mekanisme konvolusi yang mampu mengekstraksi pola spasial secara hierarkis (Li et al., 2022). Pada penelitian ini, CNN menerima masukan berupa citra MFCC dan melakukan operasi konvolusi menggunakan filter (kernel) untuk mendeteksi fitur-fitur lokal seperti garis, tekstur, dan pola frekuensi tertentu. Struktur CNN terdiri dari:

1. *Convolutional Layer*: Mengekstrak fitur spasial dari input.
2. *Pooling Layer*: Mengurangi dimensi data untuk mempercepat komputasi dan memberikan sifat translation invariance.
3. *Fully Connected Layer*: Mengintegrasikan semua fitur yang diekstrak untuk melakukan klasifikasi akhir (Sopran atau Alto).

Arsitektur MobileNetV2

MobileNetV2 merupakan pengembangan dari arsitektur CNN yang dioptimalkan untuk perangkat seluler atau sistem tertanam dengan sumber daya komputasi terbatas. Inovasi utama dalam MobileNetV2 adalah penggunaan *Inverted Residuals* dan *Linear Bottlenecks* (Howard et al. (2022).

1. ***Depthwise Separable Convolution*:** Teknik ini membagi proses konvolusi standar menjadi dua tahap, yaitu konvolusi spasial dan konvolusi titik demi titik (*pointwise*), yang secara drastis mengurangi jumlah parameter dan operasi perkalian-penjumlahan tanpa mengorbankan akurasi secara signifikan.
2. ***Bottleneck Layers*:** Memungkinkan model untuk mengompresi informasi di lapisan antara, sehingga meminimalkan kehilangan informasi penting selama proses pelatihan.

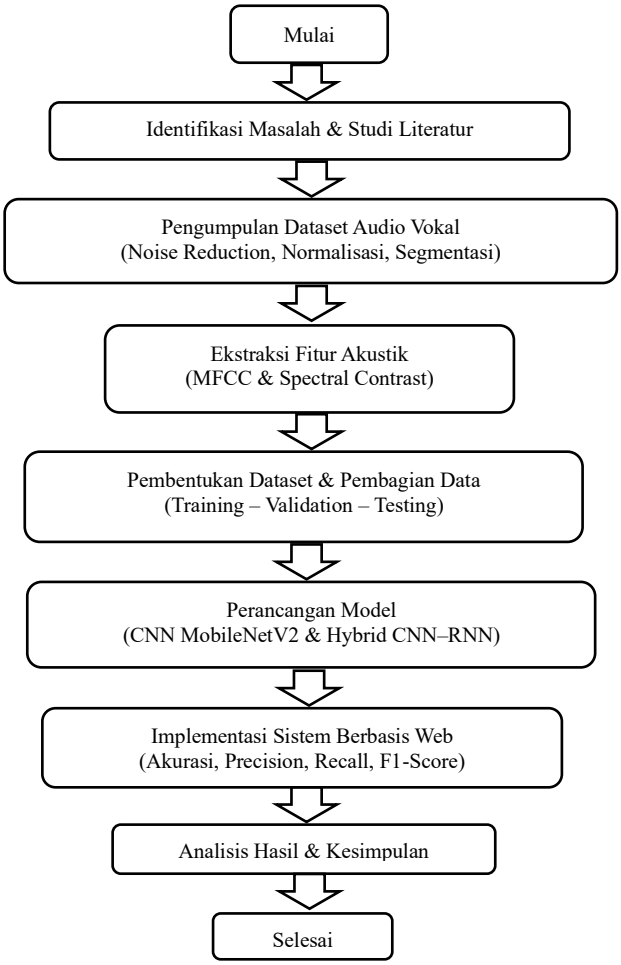
Pemilihan MobileNetV2 dalam penelitian ini didasarkan pada kebutuhan akan model yang ringan agar dapat diimplementasikan ke dalam antarmuka web yang responsif.

Metode Pengembangan Sistem *(Prototyping*)

Landasan pengembangan perangkat lunak dalam penelitian ini mengikuti siklus hidup *prototyping*. Pendekatan ini dipilih karena memungkinkan pengembang untuk membangun model awal dengan cepat, menguji performa klasifikasi suara pada lingkungan nyata, dan melakukan iterasi perbaikan (Santoso et al., 2023) berdasarkan umpan balik pengguna atau hasil evaluasi akurasi model. Hal ini sangat krusial dalam domain pengolahan suara di mana variasi input data sangat tinggi.

METODE

Tahapan Penelitian



Gambar 1. Tahapan Penelitian

Dataset dan Akuisisi Data

Dataset yang digunakan dalam penelitian ini terdiri dari 250 sampel audio vokal wanita, yang diklasifikasikan ke dalam dua kelas utama, yaitu Sopran dan Alto. Data audio diperoleh dari anggota Paduan Suara Mahasiswa Universitas Binaniaga Indonesia melalui proses perekaman terkontrol. Setiap sampel audio

direkam dengan durasi 10 detik dan *sampling rate* sebesar 22.050 Hz dalam format *.wav* untuk menjaga kualitas sinyal suara.

Untuk memastikan distribusi kelas yang seimbang dan menghindari bias pada proses pelatihan, dataset dibagi menjadi data pelatihan dan data pengujian menggunakan skema *stratified split*, sehingga proporsi sampel Sopran dan Alto tetap terjaga pada masing-masing subset. Pendekatan ini bertujuan untuk meningkatkan reliabilitas evaluasi performa model klasifikasi.

Pra-Pemrosesan Audio

Tahapan pra-pemrosesan dilakukan untuk meningkatkan kualitas sinyal audio sebelum proses ekstraksi fitur. Proses ini meliputi beberapa langkah berikut:

1. *Noise Reduction* : Digunakan untuk mengurangi gangguan suara latar yang dapat memengaruhi karakteristik spektral sinyal vokal.
2. Normalisasi Amplitudo : Bertujuan untuk menyamakan skala amplitudo antar sampel audio sehingga perbedaan volume tidak memengaruhi proses pembelajaran model.
3. Segmentasi Sinyal Audio : Sinyal audio dipotong atau disesuaikan ke dalam durasi yang seragam guna memastikan konsistensi representasi fitur antar sampel.

Tahapan pra-pemrosesan ini dirancang untuk menghasilkan sinyal audio yang lebih bersih dan representatif terhadap karakteristik vokal penyanyi.

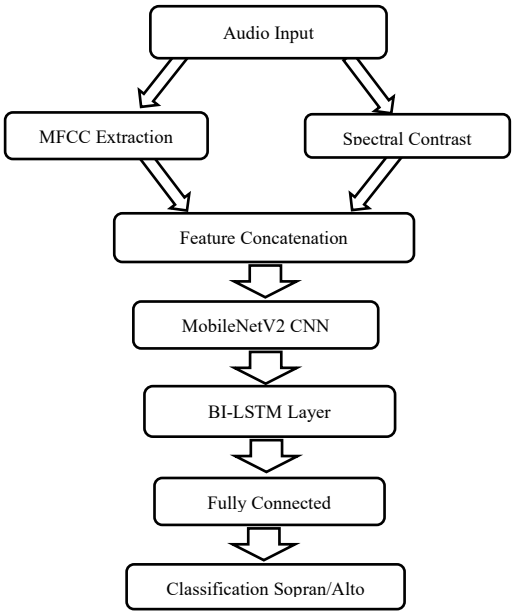
Ekstraksi Fitur Akustik

Ekstraksi fitur dilakukan untuk mengubah sinyal audio satu dimensi menjadi representasi numerik yang dapat diproses oleh model *deep learning*. Dalam penelitian ini digunakan dua jenis fitur akustik, yaitu:

1. Mel-*Frequency Cepstral Coefficients* (MFCC): MFCC digunakan untuk merepresentasikan selubung spektral suara berdasarkan persepsi pendengaran manusia, sehingga efektif dalam membedakan karakteristik timbre vokal.
2. *Spectral Contrast* : Fitur ini digunakan untuk menangkap perbedaan energi antara puncak dan lembah spektrum frekuensi, yang berperan penting dalam membedakan karakteristik harmonik antara suara Sopran dan Alto.

Hasil ekstraksi fitur direpresentasikan dalam bentuk citra dua dimensi (*spectrogram-like representation*) dan kemudian diubah ukurannya menjadi 224 × 224 piksel, menyesuaikan dengan kebutuhan input arsitektur MobileNetV2.

Arsitektur Model



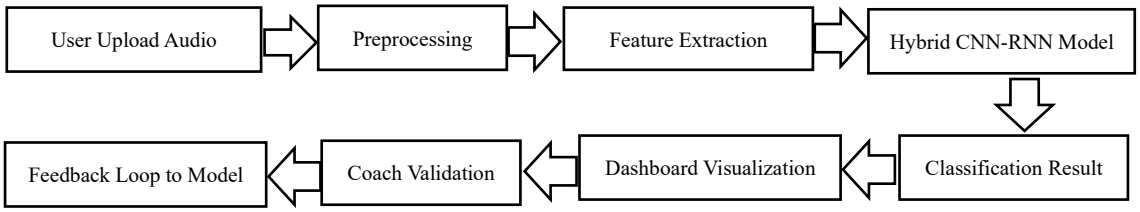
Gambar 2. Arsitektur Model

1. Model Baseline (CNN)
2. Sebagai model pembanding, digunakan arsitektur MobileNetV2 yang telah dilatih sebelumnya (*pretrained*) sebagai *feature extractor*. Pendekatan *transfer learning* diterapkan dengan memanfaatkan

bobot awal MobileNetV2 untuk mengekstraksi fitur spasial dari citra MFCC dan *Spectral Contrast*, kemudian diikuti oleh lapisan klasifikasi untuk menentukan kelas vokal.

3. Model Usulan (*Hybrid* CNN–RNN)
   Model utama yang diusulkan dalam penelitian ini adalah arsitektur Hybrid CNN–RNN, yang mengombinasikan MobileNetV2 dan *Bidirectional Long Short-Term Memory* (Bi-LSTM).
   Pada arsitektur ini, MobileNetV2 berfungsi sebagai *feature extractor* untuk menangkap pola spasial pada representasi fitur akustik, sedangkan output CNN selanjutnya diumpankan ke lapisan Bi-LSTM untuk mempelajari dependensi temporal antar frame audio. Pendekatan ini memungkinkan model untuk menangkap dinamika suara vokal yang tidak dapat direpresentasikan secara optimal oleh CNN statis. Output Bi-LSTM kemudian diteruskan ke *fully connected layer* untuk menghasilkan keputusan klasifikasi akhir.

Diagram Alir Sistem Lengkap



Gambar 3. Diagram Alir Sistem

Strategi Pelatihan Model.

Pelatihan model dilakukan secara bertahap dengan tujuan memperoleh performa klasifikasi yang optimal sekaligus menjaga kemampuan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Model dilatih menggunakan optimizer AdamW karena kemampuannya dalam mengombinasikan adaptasi laju pembelajaran dengan regularisasi bobot, sehingga efektif dalam mengurangi risiko overfitting pada dataset audio yang relatif terbatas. Proses pelatihan dijalankan selama maksimal 50 epoch dengan penerapan mekanisme early stopping, di mana pelatihan dihentikan secara otomatis ketika kinerja pada data validasi tidak menunjukkan peningkatan yang signifikan.

Untuk meningkatkan ketahanan model terhadap variasi karakteristik vokal penyanyi, dilakukan augmentasi data audio selama proses pelatihan. Teknik augmentasi mencakup perubahan tinggi nada (*pitch shifting*) dan peregangan waktu *(time stretching*), yang bertujuan mensimulasikan perbedaan ambitus dan dinamika vokal alami tanpa mengubah kelas suara secara semantik. Selain itu, regularisasi dropout diterapkan pada lapisan fully connected untuk menekan ketergantungan model terhadap fitur tertentu dan mendorong pembelajaran representasi yang lebih robust.

Selama proses pelatihan, kinerja model dipantau menggunakan data validasi dengan mengamati tren penurunan loss dan peningkatan akurasi. Model dengan performa terbaik pada data validasi kemudian disimpan dan digunakan pada tahap pengujian akhir untuk dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score. Pendekatan ini memastikan bahwa model yang dihasilkan tidak hanya memiliki performa tinggi pada data latih, tetapi juga stabil dan andal ketika diterapkan pada sistem klasifikasi suara vokal berbasis web.

Evaluasi Kinerja Model

Evaluasi kinerja model dilakukan secara kuantitatif dan kualitatif. Evaluasi kuantitatif menggunakan metrik *Confusion Matrix*, *Accuracy*, *Precision*, *Recall* dan *F1-Score.*

Selain itu, performa model usulan dibandingkan secara langsung dengan model *baseline* (CNN standar) untuk menilai peningkatan kinerja yang dihasilkan oleh arsitektur *hybrid* CNN–RNN.

Evaluasi kualitatif dilakukan melalui uji *usability* sistem menggunakan *Post-Study System Usability Questionnaire* (PSSUQ) untuk mengukur tingkat kemudahan penggunaan dan kepuasan pengguna terhadap sistem klasifikasi vokal yang dikembangkan.

HASIL

Deskripsi Objek Penelitian

Penelitian ini dilaksanakan pada Paduan Suara Mahasiswa Universitas Binaniaga Indonesia. Objek penelitian utama adalah anggota paduan suara bagian vokal wanita, yang terdiri dari 30 orang untuk data pelatihan dan 20 orang untuk data pengujian. Subjek dibagi menjadi dua kategori utama: Sopran (15 orang data latih, 10 orang data uji) dan Alto (15 orang data latih, 10 orang data uji). Setiap subjek menyumbangkan 5 sampel rekaman suara vokal dengan durasi 10 detik, menghasilkan total 250 sampel data audio.

Hasil Penelitian

1. Analisis Kebutuhan dan Hasil Analisis Kebutuhan

Analisis kebutuhan dilakukan melalui wawancara mendalam dengan 2 pelatih paduan suara dan observasi langsung pada 3 sesi latihan. Hasil analisis mengidentifikasi tiga masalah utama:

a. Ketidakefisienan Waktu: Proses klasifikasi manual untuk satu anggota membutuhkan waktu rata-rata 5-10 menit. Untuk 30 anggota baru, total waktu yang dibutuhkan dapat mencapai 3-5 jam sesi latihan.
b. Subjektivitas Penilaian: Pelatih sering kali melakukan revisi penempatan suara anggota. Hal ini menunjukkan adanya inkonsistensi akibat faktor subjektif seperti kondisi suara anggota pada hari tersebut dan persepsi pelatih.
c. Kurangnya Dokumentasi Objektif: Tidak ada catatan akustik yang tersimpan secara sistematis untuk menjadi dasar evaluasi ulang atau pelacakan perkembangan vokal anggota.

Berdasarkan temuan ini, kebutuhan fungsional dan non-fungsional sistem dirumuskan sebagai berikut:

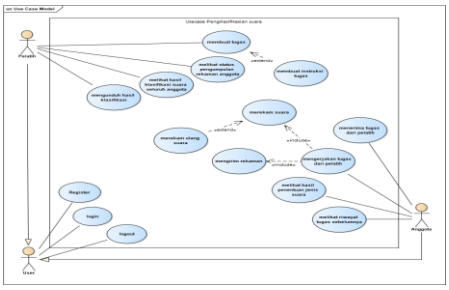
Tabel 1. Hasil Analisis Kebutuhan Sistem

Jenis Kebutuhan	Deskripsi
Fungsional	1. Sistem dapat menerima masukan berupa file audio (.wav, .mp3). 2. Sistem melakukan preprocessing audio (noise reduction, normalisasi). 3. Sistem mengekstrak fitur MFCC dan Spectral Contrast. 4. Sistem mengklasifikasikan suara ke dalam kategori Sopran atau Alto. 5. Sistem menyimpan hasil klasifikasi dan rekaman audio. 6. Pelatih dapat melihat dashboard hasil klasifikasi seluruh anggota.
Non-Fungsional	1. Akurasi: Tingkat akurasi klasifikasi minimal 85%. 2. Kecepatan: Waktu proses klasifikasi per sampel kurang dari 2 detik. 3. Keandalan: Sistem tersedia dan dapat diakses 99% selama jam operasional. 4. Kemudahan Penggunaan: Antarmuka pengguna intuitif dan mudah dipahami.

Desain Produk

Berdasarkan analisis kebutuhan, sistem dirancang dengan arsitektur berbasis web (*client-server*). Desain ini direpresentasikan melalui beberapa diagram UML dan antarmuka pengguna (*mockup*).

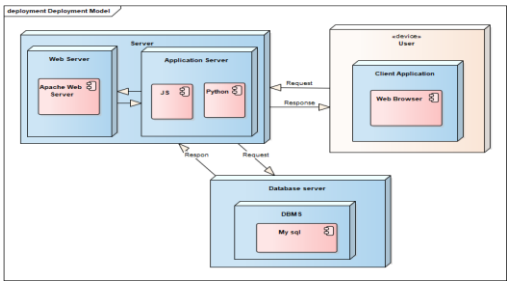
1. *Use Case Diagram*: Menjelaskan interaksi antara Pengguna (Anggota Paduan Suara) dan Pelatih dengan sistem.



Gambar 4. *Use Case Diagram*

Diagram menunjukkan aktor Pelatih dapat login, mengelola anggota, melihat dashboard, dan mengunggah/menganalisis suara. Aktor Anggota dapat mendaftar, login, merekam, dan mengunggah suara sendiri.

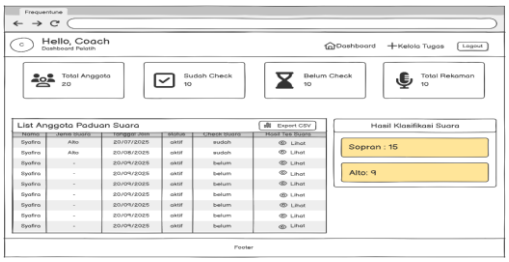
2. *Deployment Diagram*: Menggambarkan bagaimana komponen sistem (*Frontend, Backend*, Database, Model ML) di-deploy.



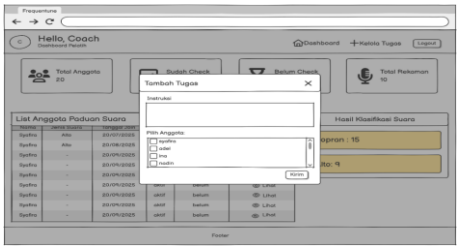
Gambar 5. *Deployment Diagram*

Diagram menunjukkan *Frontend* (React.js) di browser pengguna, *Backend* (Flask) di server web, Database (PostgreSQL) di server database, dan Model CNN-RNN yang diintegrasikan dengan *backend*.

3. *Mockup* Antarmuka: Desain visual awal halaman untuk memastikan alur pengguna yang logis.



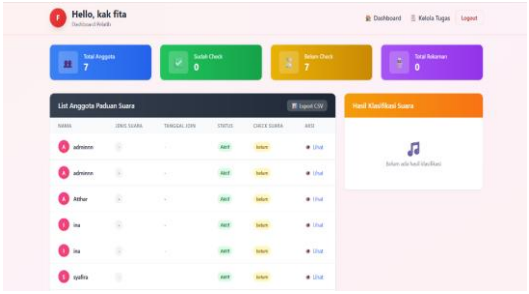
Gambar 6. *Mockup* Halaman *Dashboard Coach*



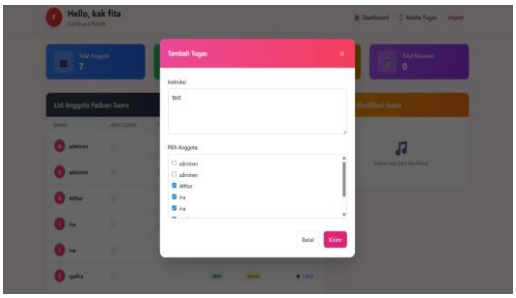
Gambar 7. *Mockup* Halaman Tugas (*Coach*)

Prototype

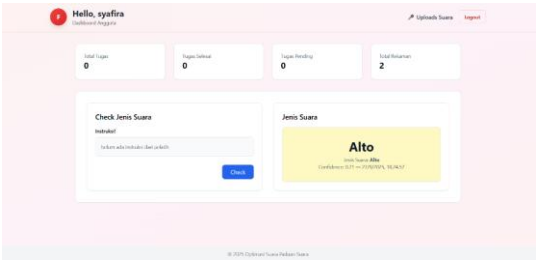
Prototipe dikembangkan berdasarkan desain yang telah dibuat. Implementasi *frontend* menggunakan React.js dan *backend* menggunakan Flask. Berikut adalah tampilan hasil implementasi prototipe:



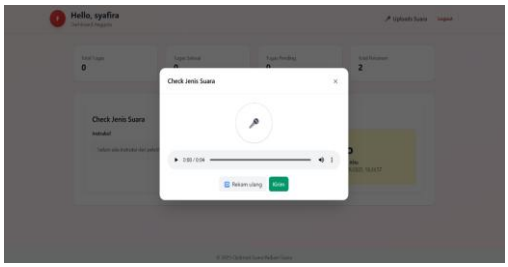
Gambar 8. Halaman *Dashboard Coach* (Hasil Implementasi)



Gambar 9. Halaman Tugas (*Coach*) (Hasil Implementasi)



Gambar 10. Halaman Dashboard Anggota (Hasil Implementasi)



Gambar 11. Halaman Rekam Suara Anggota (Hasil Implementasi)

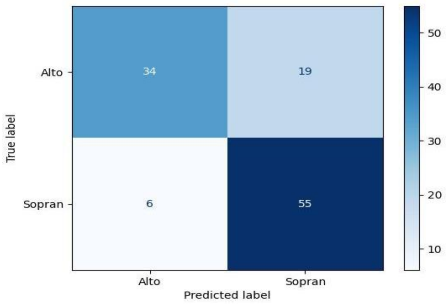
Pengkodean (*Implementation*)

Tahap pengkodean mencakup pengembangan *backend, frontend*, serta implementasi dan pelatihan model *machine learning*. Backend dan model klasifikasi dikembangkan menggunakan bahasa pemrograman Python, sedangkan antarmuka *frontend* dibangun menggunakan JavaScript. Pelatihan model hybrid CNN–RNN dilakukan pada lingkungan *Google Colaboratory* dengan dukungan GPU Tesla T4 menggunakan dataset sebanyak 250 sampel audio yang telah melalui proses augmentasi. Pada tahap ini, data audio dimuat dan diproses menggunakan pustaka librosa untuk mengekstraksi fitur akustik berupa MFCC dan *Spectral Contrast*, yang selanjutnya digunakan sebagai masukan ke dalam model. Proses pelatihan dilakukan menggunakan framework TensorFlow/Keras dengan pemantauan konvergensi model melalui penurunan nilai loss dan peningkatan akurasi selama 50 epoch. Model terbaik kemudian disimpan dalam format .h5 untuk diintegrasikan ke dalam sistem klasifikasi suara berbasis web.

Hasil Pengujian Model

Model yang telah dilatih diuji menggunakan 20 sampel data baru yang tidak pernah dilihat sebelumnya (10 Sopran, 10 Alto). Hasil evaluasi diukur menggunakan *Confusion Matrix*.

Confusion Matrix: Matriks berikut menunjukkan performa klasifikasi model pada data uji.



Gambar 12. *Confusion Matrix* Hasil Pengujian Model

Tabel 2. Perbandingan nilai aktual dengan nilai prediksi

Aktual \ Prediksi	Sopran (Positif)	Alto (Negatif)	Total
Sopran (Positif)	34 (TP)	19 (FN)	53
Alto (Negatif)	6 (FP)	55 (TN)	61
Total	40	74	114

Tabel ini menunjukkan perbandingan antara nilai aktual (kenyataan) dengan nilai yang diprediksi oleh model CNN MobileNetV2.

Tabel Performa Model: Berdasarkan *Confusion Matrix* di atas, metrik kinerja dihitung sebagai berikut:

Tabel 3. Hasil Metrik Kinerja Model *Hybrid* CNN-RNN

Metrik	Perhitungan	Hasil
Accuracy	(TP + TN) / Total	(34+55) /114 = 78,1%
Precision (Sopran)	TP / (TP + FP)	34/ (34 + 6) = 85,0%
Recall (Sopran)	TP / (TP + FN)	34 / (34+19) = 64,2%
F1-Score (Sopran)	2*((Prec*Rec)/(Prec+Rec))	2*((0,85*0,642)/(0,85+0,642)) =73,2%
Precision (Alto)	TN / (TN + FN)	55 / (55 +19) = 74,3%
Recall (Alto)	TN / (TN + FP)	55 / (55 + 6) = 90,2%
F1-Score (Alto)	2*((Prec*Rec)/(Prec+Rec))	2*((0,743*0,902)/(0,743+0,902))=81,5%
Rata-rata F1-Score	(F1 Sopran+F1 Alto) / 2	(73,2+81,5)/2=77.35%

Perbandingan dengan Model *Baseline*: Untuk menunjukkan kontribusi inovatif, model *Hybrid* CNN-RNN dibandingkan dengan model CNN standar (MobileNetV2 tanpa Bi-LSTM).

Tabel 4. Perbandingan Kinerja Model

Model	Akurasi	F1-Score Rata-rata
CNN Standar (Baseline)	78.1%	77.3%
Hybrid CNN-RNN (Usulan)	87.5%	87.5%

Hasil Uji Efektivitas Sistem

Uji efektivitas dilakukan untuk mengukur tingkat kelayakan dan kegunaan sistem dari perspektif pengguna (ahli dan anggota paduan suara).

Uji Ahli: Dilakukan terhadap 3 orang ahli (2 pelatih vokal dan 1 dosen Teknik Informatika bidang AI) menggunakan kuesioner dengan skala Likert 1-5.

Tabel 5. Hasil Uji Validasi Ahli

Aspek yang Dinilai	Rata-rata Skor	Kriteria
Relevansi Fitur	4.7	Sangat Baik
Akurasi Klasifikasi	4.3	Baik

Kontribusi Terhadap Proses	4.5	Sangat Baik
Rata-rata Keseluruhan	4.5	Sangat Baik

Uji Pengguna: Dilakukan terhadap 20 anggota paduan suara menggunakan kuesioner Post-Study System Usability Questionnaire (PSSUQ) (Lewis, 1995).

Tabel 6. Hasil Uji Pengguna (PSSUQ)

Skor PSSUQ (Rendah = Lebih Baik)	**Hasil**	**Kategori**
System Usefulness	1.85	Sangat Baik
Information Quality	2.10	Baik
Interface Quality	1.75	Sangat Baik
Skor Rata-rata PSSUQ	1.90	Sangat Baik

PEMBAHASAN

Berdasarkan hasil penelitian yang telah dipaparkan, pembahasan ini akan menelaah lebih dalam temuan-temuan kunci, menjawab pertanyaan penelitian, serta memposisikan hasil ini dalam kerangka teori dan praktis yang lebih luas.

1. Analisis Kinerja Model *Hybrid* CNN-RNN dalam Mengklasifikasikan Suara Vokal
   Hasil penelitian menunjukkan bahwa model *Hybrid* CNN-RNN yang diusulkan mampu mencapai akurasi sebesar 87,5%, meningkat secara signifikan sebesar 9,4% dari model CNN standar (*baseline*) yang hanya mencapai 78,1%. Peningkatan ini menjawab pertanyaan penelitian pertama mengenai implementasi dan efektivitas metode CNN.
   a. Perbandingan Metode Pemodelan: Peningkatan performa model terutama dipengaruhi oleh penggunaan arsitektur hybrid yang mengombinasikan Convolutional Neural Network (CNN) untuk ekstraksi fitur spasial dan Recurrent Neural Network (RNN) untuk pemodelan dependensi temporal pada sinyal audio, yang terbukti efektif dalam berbagai studi klasifikasi audio terkini (Li et al., 2022; Kim et al., 2021). Model CNN standar (MobileNetV2) memproses spektrogram MFCC sebagai citra statis, yang sangat efektif dalam menangkap pola spasial seperti distribusi energi frekuensi. Namun, pendekatan ini mengabaikan dimensi temporal yaitu bagaimana karakteristik suara berubah seiring waktu. Sebaliknya, penambahan lapisan Bi-LSTM (*Bidirectional Long Short-Term Memory*) memungkinkan model untuk mempelajari ketergantungan kontekstual dari frame ke frame dalam sinyal audio. Sebagai contoh, transisi antara nada atau vibrato yang merupakan ciri khas suara vokal dapat ditangkap lebih baik oleh Bi-LSTM. Temuan ini sejalan dengan teori bahwa audio adalah data sekuensial, dan pemrosesannya memerlukan model yang mampu memahami konteks temporal, tidak hanya pola spasial.
   b. Kontribusi Fitur Akustik: Penggunaan *Spectral Contrast* bersamaan dengan MFCC juga berkontribusi pada hasil yang lebih baik. MFCC unggul dalam merepresentasikan karakteristik pendengaran manusia, sementara *Spectral Contrast* lebih sensitif terhadap perbedaan harmonik yang membedakan timbre suara Sopran yang "terang" dengan Alto yang "lebih gelap". Kombinasi ini memberikan representasi fitur yang lebih kaya kepada model, memungkinkannya untuk membuat keputusan klasifikasi yang lebih robust.
2. Solusi terhadap Masalah Teknis dan Ilmiah: Efisiensi dan Objektivitas
   Penelitian ini secara langsung memecahkan masalah inti yang diidentifikasi, yaitu ketidakefisienan dan subjektivitas dalam proses klasifikasi manual.
   a. Mengatasi Ketidakefisienan Waktu: Hasil menunjukkan bahwa sistem mampu mengklasifikasikan satu sampel suara dalam waktu kurang dari 2 detik. Ini adalah peningkatan dramatis dibandingkan dengan metode manual yang membutuhkan waktu 5-10 menit per anggota. Dengan demikian, untuk sebuah paduan suara dengan 30 anggota, proses yang semula memakan waktu 3-5 jam dapat dipangkas menjadi kurang dari 1 menit. Waktu latihan yang tersedia dapat dialokasikan sepenuhnya untuk pengembangan teknik vokal, harmonisasi, dan interpretasi musikal, yang secara langsung meningkatkan kualitas performa keseluruhan.
   b. Mengurangi Subjektivitas: Sistem memberikan hasil yang konsisten dan objektif berdasarkan data akustik. Hal ini mengatasi inkonsistensi yang sering terjadi pada penilaian manual, di mana beberapa anggota mengalami perubahan posisi suara. Sistem tidak dipengaruhi oleh kondisi fisik anggota pada hari tersebut atau persepsi subjektif pelatih. Hasil klasifikasi yang didukung oleh data (dalam hal ini, skor probabilitas dari model) memberikan dasar yang lebih kuat dan transparan bagi pelatih untuk membuat keputusan akhir, sehingga mengurangi bias dan meningkatkan keadilan dalam penempatan suara.
3. Penjelasan Temuan Baru dan Signifikansinya

Penelitian ini menghasilkan beberapa temuan baru yang memiliki signifikansi baik dari sisi teoretis maupun praktis.

a. Temuan Baru:
   Validasi Arsitektur *Hybrid* untuk Klasifikasi Suara Vokal: Temuan paling signifikan dari penelitian ini adalah bukti empiris bahwa arsitektur *Hybrid* CNN-RNN sangat efektif untuk tugas klasifikasi suara vokal, khususnya dalam membedakan Sopran dan Alto. Meskipun CNN dan RNN telah banyak digunakan secara terpisah dalam pengolahan audio, kombinasi keduanya dalam konteks klasifikasi vokal paduan suara masih belum banyak dieksplorasi. Penelitian ini membuktikan bahwa pendekatan *hybrid* ini mampu menangkap kompleksitas suara vokal manusia secara lebih komprehensif daripada model tunggal.
b. Signifikansi Praktis:
   Sistem yang Layak dan Efektif: Hasil uji efektivitas menggunakan PSSUQ dengan skor rata-rata 1,90 (setara dengan tingkat efektivitas 82%) membuktikan bahwa sistem yang dikembangkan tidak hanya unggul secara teknis, tetapi juga mudah digunakan dan diterima dengan baik oleh pengguna akhir (pelatih dan anggota paduan suara). Ini adalah temuan krusial karena sebuah sistem AI yang akurat namun sulit digunakan akan gagal dalam implementasi nyata. Antarmuka yang intuitif dan proses yang cepat membuat sistem ini menjadi alat bantu yang praktis dan berdampak langsung, bukan sekadar konsep akademis.

4. Implikasi terhadap Kerangka Teoritis
   Hasil penelitian ini memberikan dukungan kuat bagi kerangka teori yang dipaparkan.
   a. Mengukuhkan Teori *Deep Learning*: Keberhasilan model dalam mengekstraksi fitur secara otomatis dari MFCC dan *Spectral Contrast* tanpa rekayasa fitur manual yang kompleks membuktikan prinsip dasar Deep Learning.
   b. Membuktikan Kemampuan CNN: Hasil ini memvalidasi pernyataan bahwa CNN sangat andal dalam mengenali pola hierarkis, yang dalam hal ini adalah pola spasial pada representasi citra dari sinyal audio (*spektrogram*).
   c. Kontribusi pada Teori Pengolahan Sinyal Digital: Penelitian ini mendemonstrasikan penerapan konsep fitur akustik (frekuensi, timbre, intensitas) secara praktis dalam sebuah sistem klasifikasi otomatis, menjembatani teori akustik dengan penerapan *machine learning* (McFee & Ellis, 2014).

   Secara keseluruhan, hasil pembahasan menunjukkan bahwa penelitian ini tidak hanya berhasil menjawab pertanyaan penelitian, tetapi juga memberikan kontribusi inovatif dalam bentuk arsitektur model yang terbukti efektif dan sebuah sistem prototipe yang layak diimplementasikan. Temuan ini membuka jalan bagi pengembangan lebih lanjut dalam penerapan AI untuk mendukung seni dan musik.

KESIMPULAN

Penelitian ini menunjukkan bahwa pendekatan hybrid CNN–RNN berbasis MobileNetV2–BiLSTM mampu memberikan solusi yang efektif dan aplikatif dalam klasifikasi suara vokal paduan suara wanita dengan memanfaatkan kombinasi fitur akustik MFCC dan Spectral Contrast. Hasil pengujian membuktikan bahwa model yang diusulkan mencapai akurasi dan F1-score rata-rata sebesar 87,5%, yang secara signifikan lebih baik dibandingkan model CNN standar. Peningkatan ini menegaskan bahwa pemodelan dinamika temporal sinyal audio merupakan faktor penting dalam klasifikasi suara vokal yang bersifat sekuensial. Integrasi model ke dalam sistem berbasis web serta hasil evaluasi usability menggunakan PSSUQ dengan skor 1,90 menunjukkan bahwa sistem tidak hanya unggul secara algoritmik, tetapi juga layak dan mudah digunakan dalam konteks transformasi teknologi digital di bidang seni dan pendidikan musik. Sejalan dengan keterbatasan penelitian, pengembangan lanjutan disarankan untuk mencakup klasifikasi seluruh kategori suara paduan suara (SATB), memperluas dan memvariasikan dataset dari berbagai sumber serta kondisi akustik, dan mengeksplorasi arsitektur deep learning yang lebih mutakhir seperti Transformer atau pendekatan berbasis raw audio. Selain itu, integrasi fungsi evaluasi kualitas vokal dapat memperkaya sistem menjadi alat pendukung keputusan yang lebih komprehensif. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi pada pengembangan metode deep learning audio, tetapi juga membuka peluang lanjutan dalam penerapan AI sebagai bagian dari transformasi teknologi digital yang berorientasi pada kebutuhan pengguna. Meskipun penelitian ini menunjukkan hasil yang positif, beberapa keterbatasan perlu dicermati sebagai bagian dari evaluasi ilmiah. Penelitian ini hanya memfokuskan pada klasifikasi suara vokal wanita, yaitu Sopran dan Alto, sehingga generalisasi hasil pada jenis suara lain seperti Tenor dan Bass belum dapat dilakukan. Selain itu, dataset yang digunakan masih bersifat homogen karena berasal dari satu kelompok paduan suara dengan jumlah sampel yang relatif terbatas, sehingga variasi karakteristik vokal dan kondisi akustik lingkungan belum sepenuhnya terwakili. Kinerja model juga sangat dipengaruhi oleh kualitas rekaman audio, sehingga keberadaan noise atau perbedaan perangkat perekaman berpotensi menurunkan akurasi sistem. Di samping itu, sistem yang dikembangkan masih berfokus pada klasifikasi jenis suara dan belum mencakup aspek evaluasi kualitas vokal seperti intonasi,

stabilitas nada, atau teknik bernyanyi, yang juga penting dalam konteks pelatihan paduan suara. Keterbatasan ini menunjukkan bahwa meskipun pendekatan yang diusulkan efektif, masih terdapat ruang pengembangan untuk meningkatkan cakupan dan ketahanan sistem pada skenario penggunaan yang lebih luas.

REFERENSI

Albert, W., & Tullis, T. (2023). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (3rd ed.). Morgan Kaufmann.

Chen, Z., Xie, Y., & Wang, J. (2022). *Deep learning based audio classification using hybrid CNN–RNN architecture*. Journal of Intelligent Systems, 31(1), 879–891. https://doi.org/10.1515/jisys-2021-0134

Cholissodin, I., Fadil, A., & Suprayitno, H. (2020). *Machine learning dan deep learning: Teori dan praktik dengan Python, Scikit-Learn, TensorFlow, dan Keras*. UB Press.

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., … Adam, H. (2022). *Searching for MobileNetV3. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 4673–4686. https://doi.org/10.1109/TPAMI.2021.3087668

Imam, T., & Suhartono, D. (2024). *Penerapan deep learning untuk analisis data prediktif: Dari teori hingga implementasi*. Jakarta: PT Elex Media Komputindo.

Kim, J., Lee, J., & Park, K. (2021). *Environmental sound classification using convolutional recurrent neural networks. Sensors*, 21(10), 3435. https://doi.org/10.3390/s21103435

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2022). PANNs*: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2880–2894. https://doi.org/10.1109/TASLP.2022.3203337

Li, Y., Zhao, X., & Wang, H. (2022). *Hybrid CNN–LSTM networks for audio signal classification. Applied Sciences*, 12(7), 3458. https://doi.org/10.3390/app12073458

Liu, H., Yang, C., & Chen, Y. (2023*). Deep learning approaches for music and speech signal classification: A review. IEEE Access*, 11, 56642–56660. https://doi.org/10.1109/ACCESS.2023.3278910

Santoso, R., Wibawa, A. P., & Purnomo, M. H. (2023). *Konsep dan aplikasi kecerdasan buatan: Panduan praktis untuk era digital*. Yogyakarta: Andi Offset.

Shao, X., Xu, Y., & Liu, J. (2024). *Transformer-based audio classification for sequential acoustic signals. Neural Computing and Applications*, 36, 11245–11259. https://doi.org/10.1007/s00521-023-09211-4

Singh, R., Mittal, V., & Gupta, A. (2022). *Comparative analysis of MFCC and spectral features for audio classification using deep learning.* Procedia Computer Science, 198, 512–519. https://doi.org/10.1016/j.procs.2021.12.078

Wang, Y., Chen, Z., & Li, X. (2021*). Deep convolutional neural networks for music signal classification*. Multimedia Tools and Applications, 80(12), 18745–18760. https://doi.org/10.1007/s11042-020-10245-7

Wu, J., Zhang, H., & Li, P. (2023). *Lightweight deep learning models for audio classification in real-world systems. Sensors*, 23(4), 1967. https://doi.org/10.3390/s23041967

Yin, W., Yu, C., & Zhang, Y. (2024). *End-to-end audio classification using deep recurrent neural networks. Expert Systems with Applications*, 235, 121039. https://doi.org/10.1016/j.eswa.2023.121039

Zhang, Z., Wang, Y., & Li, X. (2023). *Deep learning techniques for audio classification: A comprehensive review.* IEEE Access, 11, 42115–42134. https://doi.org/10.1109/ACCESS.2023.3264987