Klasifikasi Dialek Bahasa Jawa Berbasis Arsitektur *Hybrid* CNN-TCN dengan Strategi *Speaker-Level Voting*

Danika Najwa Ardelia
Informatika, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia

Diterima: 11 Januari, 2024 | Revisi: 11 Mei, 2024 | Diterbitkan: 11 Juni 2025
DOI:

ABSTRAK
Pelestarian variasi dialek Bahasa Jawa menghadapi tantangan komputasi akibat kompleksitas fitur spektral dan temporal, seperti intonasi dan tempo, yang sulit ditangkap model konvensional. Penelitian ini mengajukan arsitektur hibrida Convolutional Neural Network (CNN) dan Temporal Convolutional Network (TCN) untuk mengatasi kelemahan CNN murni yang sering mengabaikan ketergantungan urutan waktu jangka panjang dalam sinyal audio. Menggunakan representasi Log-Mel Spectrogram dari data dialek Banyumas, Solo, dan Malang, penelitian menerapkan skema evaluasi Stratified Group K-Fold serta strategi agregasi Speaker-Level Voting. Secara empiris, integrasi ini meningkatkan akurasi klasifikasi secara signifikan dari 83,98% pada tingkat segmen menjadi 90,55% pada tingkat penutur. Pendekatan hibrida terbukti efektif menangkap karakteristik dialek yang dinamis, menawarkan solusi robust untuk dokumentasi digital bahasa daerah meskipun terdapat tantangan bias pada kelas data minoritas.

Kata Kunci: Klasifikasi Dialek, Bahasa Jawa, *Hybrid* CNN-TCN, *Speaker-Level Voting*, *Deep Learning*

ABSTRACT
The preservation of Javanese dialect variations faces computational challenges due to complex spectral and temporal features, such as intonation and tempo, which are difficult for conventional models to capture. This study proposes a hybrid architecture combining Convolutional Neural Networks (CNN) and Temporal Convolutional Networks (TCN) to address the limitations of pure CNNs, which often overlook long-term sequential dependencies in audio signals. Utilizing Log-Mel Spectrogram representations from Banyumas, Solo, and Malang dialect data, the research implements a Stratified Group K-Fold evaluation scheme and a Speaker-Level Voting aggregation strategy. Empirically, this integration significantly improves classification accuracy from 83.98% at the segment level to 90.55% at the speaker level. The hybrid approach proves effective in capturing dynamic dialect characteristics, offering a robust solution for the digital documentation of regional languages despite existing challenges regarding minority class bias.

Keywords: Dialect Classification, Javanese Language, Hybrid CNN-TCN, Speaker-Level Voting, Deep Learning.

PENDAHULUAN

Bahasa daerah adalah elemen penting dalam identitas budaya di Indonesia. Namun, keberadaan bahasa daerah saat ini mulai terancam karena adanya perubahan kebiasaan komunikasi antar-generasi. Data Sensus Penduduk 2020 menunjukkan bahwa penggunaan bahasa daerah di rumah terus menurun, dari 79,64% pada tahun 2010 menjadi 74,77% pada tahun 2020, terutama di kalangan generasi muda (PROFIL SUKU DAN KERAGAMAN BAHASA DAERAH HASIL LONG FORM SENSUS PENDUDUK 2020, 2020). Masalah ini juga terjadi pada Bahasa Jawa, bahasa dengan jumlah penutur terbanyak di Indonesia. Walaupun jumlah penuturnya masih banyak, pelestarian Bahasa Jawa cukup sulit dilakukan karena variasi dialeknya sangat luas, mulai dari dialek Ngapak di barat, Tengahan, hingga dialek Timuran (Hasisah & Suryadi, 2022). Oleh karena itu, usaha pelestarian tidak cukup hanya fokus pada bahasa standar, tetapi juga perlu mendokumentasikan variasi dialek tersebut menggunakan teknologi digital (Ajani dkk., 2024).

Masalah utama dalam mengelompokkan dialek Bahasa Jawa menggunakan komputer terletak pada fitur suaranya yang rumit. Secara bahasa, perbedaan antar-dialek bukan hanya soal kosakata, tetapi juga menyangkut aspek spektral (tinggi rendahnya nada) dan temporal (durasi dan kecepatan bicara). Studi terbaru menunjukkan bahwa dialek Pekalongan memiliki nada yang tinggi dan melengking, berbeda dengan dialek pedalaman yang lebih berat (Ardini & Sunarya, 2024). Selain itu, aspek waktu juga sangat penting; dialek Jawa Timuran seperti Surabaya terbukti memiliki tempo bicara yang jauh lebih cepat dibandingkan dialek Jawa Tengahan yang cenderung lambat (Mawarni dkk., 2024). Ciri khas seperti "cengkok" atau alunan nada di akhir kalimat ini membutuhkan model komputasi yang bisa menangkap detail frekuensi dan urutan waktu secara bersamaan.

Selain kompleksitas fitur linguistik, tantangan lain dalam pemrosesan sinyal audio adalah variabilitas kondisi data di lapangan. Rekaman suara dialek sering kali memiliki kualitas yang beragam, mulai dari adanya gangguan suara latar (*background noise*) hingga durasi rekaman yang tidak seragam. Kondisi ini menuntut adanya strategi pra-pemrosesan data yang tepat agar informasi penting tidak hilang. Oleh karena itu, penggunaan representasi visual audio yang kuat sangat diperlukan untuk memastikan model tetap dapat mengenali pola dialek meskipun kualitas rekamannya bervariasi atau kurang ideal.

Walaupun teknologi kecerdasan buatan (*Artificial Intelligence*) sudah sangat maju, penelitian sebelumnya tentang klasifikasi bahasa daerah di Indonesia hasilnya belum maksimal. Sebagai contoh, penelitian Fauzi dkk. (2022) yang menggunakan metode MFCC dan ANFIS untuk mengenali ragam bahasa di Pulau Jawa hanya menghasilkan akurasi 32,5%. Hasil yang rendah ini menunjukkan bahwa metode konvensional belum cukup kuat untuk membedakan fitur dialek yang perbedaannya sangat tipis.

Saat ini, Convolutional Neural Network (CNN) menjadi metode yang paling sering digunakan untuk mengolah data audio dalam bentuk representasi visual, khususnya Log-Mel Spectrogram. CNN sangat bagus dalam menangkap pola frekuensi dari representasi ini dan tahan terhadap gangguan suara (*noise*) (Setianingrum dkk., 2023). Akan tetapi, CNN murni memiliki kelemahan karena menganggap input audio seperti gambar diam, sehingga sering kali mengabaikan aspek urutan waktu yang panjang

(Fantaye dkk., 2020). Padahal, dalam dialek Jawa, informasi mengenai tempo dan cengkok tersebar dalam rentang waktu tertentu yang tidak bisa ditangkap jika hanya melihat potongan fitur sesaat.

Untuk mengatasi kekurangan tersebut, penelitian ini mengusulkan metode gabungan (*hybrid*) antara CNN dengan *Temporal Convolutional Network* (TCN). TCN dipilih karena kemampuannya yang spesifik dalam memodelkan data berurutan. Model ini dapat menangkap ketergantungan jangka panjang (long-term dependencies) dalam suara tanpa mengalami masalah teknis (vanishing gradient) yang sering terjadi pada model lama seperti RNN (Jo & Kwak, 2025).

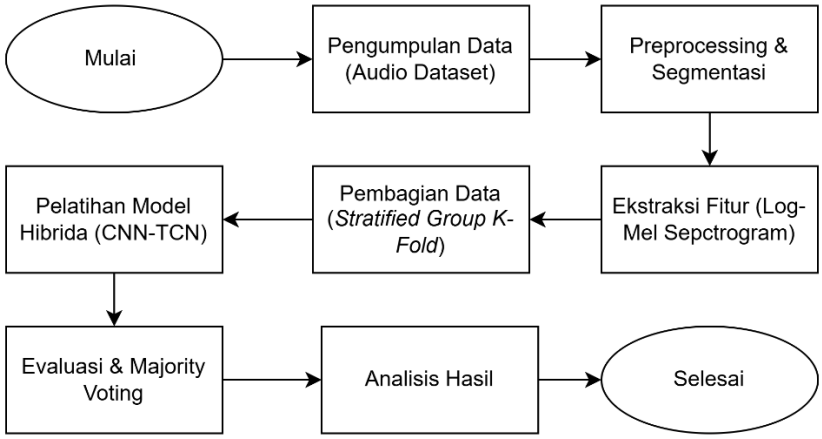
Melalui gabungan kedua metode ini, penelitian ini bertujuan membangun sistem klasifikasi dialek Bahasa Jawa yang lebih akurat. Dalam desain yang diajukan, CNN bertugas sebagai *front-end* untuk mengambil fitur spektral dari *Log-Mel Spectrogram*, kemudian hasilnya diproses lebih lanjut oleh TCN sebagai *back-end* untuk menangkap pola waktu dan ritme bicara. Kombinasi fitur spasial dan temporal ini diharapkan membuat model lebih peka terhadap perbedaan dialek yang halus dibandingkan jika hanya menggunakan satu model saja (Chen dkk., 2025), sehingga dapat menjadi dasar teknis yang berguna untuk pelestarian bahasa daerah secara digital.

Lebih jauh lagi, keberhasilan pengembangan model ini diharapkan dapat membuka peluang untuk aplikasi praktis yang lebih luas. Jika sistem ini terbukti andal, teknologi ini dapat diterapkan dalam pembuatan kamus digital otomatis, aplikasi pembelajaran bahasa daerah, atau sistem pengenalan penutur yang lebih canggih. Dengan demikian, upaya pelestarian budaya tidak hanya berhenti pada pengarsipan data, tetapi juga berkembang menjadi alat bantu yang relevan dan bermanfaat bagi masyarakat modern.

METODE PENELITIAN

Penelitian ini menerapkan kerangka kerja sistematis untuk menyelesaikan permasalahan klasifikasi dialek yang kompleks. Secara garis besar, alur penelitian dirancang untuk mengubah data audio mentah menjadi prediksi kelas dialek yang akurat melalui serangkaian tahapan yang saling berkesinambungan. Proses diawali dengan akuisisi data suara dari tiga variasi dialek target, yang kemudian melewati tahap pra-pemrosesan sinyal yang ketat untuk memastikan kualitas input. Data yang telah bersih selanjutnya diubah menjadi representasi visual *Log-Mel Spectrogram* agar dapat diproses oleh mesin.

Tahapan inti dari penelitian ini terletak pada pemodelan arsitektur hibrida, saat fitur spektral diekstraksi oleh komponen CNN dan dinamika waktunya dipelajari oleh komponen TCN. Rangkaian proses ini diakhiri dengan evaluasi performa menggunakan skema validasi silang (*cross-validation*) untuk menguji konsistensi model, baik pada level potongan segmen maupun pada level penutur secara keseluruhan. Gambaran menyeluruh mengenai tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan *subset* terpilih dari korpus data Bahasa Jawa yang didokumentasikan oleh *The Language Archive* (TLA), Max Planck Institute for Psycholinguistics (https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0022_8651_8). Dataset ini mencakup rekaman ujaran penutur asli yang difokuskan pada tiga kelas dialek utama, yaitu: Banyumas (merepresentasikan dialek wilayah barat/Ngapak), Solo (merepresentasikan dialek standar), dan Malang (merepresentasikan dialek Jawa Timuran). Pemilihan ketiga variasi ini didasarkan pada perbedaan karakteristik fonetik dan intonasi yang cukup distingtif, namun tetap berada dalam satu rumpun bahasa yang sama. Seluruh sampel audio disimpan dalam format .wav dengan frekuensi pencuplikan (*sampling rate*) yang diseragamkan menjadi 16 kHz untuk menjaga konsistensi kualitas sinyal input selama proses komputasi.

B. *Preprocessing* dan Segmentasi

Mengingat data rekaman asli memiliki durasi yang bervariasi dan mengandung gangguan suara latar, tahap pra-pemrosesan menjadi langkah krusial untuk meningkatkan kualitas sinyal sebelum diolah lebih lanjut. Proses ini dimulai dengan *Silence Removal* (penghapusan keheningan) menggunakan metode *energy-based thresholding* untuk membuang bagian audio yang tidak memuat informasi suara aktif. Langkah ini selaras dengan pendekatan yang umum dilakukan dalam pemrosesan sinyal audio digital untuk memisahkan sinyal wicara dari noise lingkungan, sehingga model dapat fokus pada fitur yang relevan (Najah Ulfah dkk., 2025). Setelah dibersihkan, sinyal audio dipotong-potong melalui proses segmentasi menjadi unit-unit yang lebih kecil dengan durasi tetap selama 3 detik tanpa tumpang tindih (*non-overlapping*).

C. Ekstraksi Fitur

Potongan sinyal audio dalam domain waktu kemudian diubah menjadi representasi dua dimensi dalam domain frekuensi-waktu menggunakan transformasi *Log-Mel Spectrogram*. Pemilihan fitur ini didasarkan pada karakteristik persepsi pendengaran

manusia yang merespons frekuensi suara secara logaritmik, bukan linier, sehingga lebih akurat dalam merepresentasikan karakteristik suara (Seo dkk., 2022).

D. Pembagian Data

Penelitian ini menggunakan skema validasi silang Stratified Group K-Fold dengan $k = 5$. Berbeda dengan pembagian acak biasa, skema ini menggunakan informasi sumber rekaman (*Group* ID) sebagai basis pemisahan. Pendekatan ini diadopsi untuk mencegah terjadinya kebocoran data (data leakage) antar-segmen, di mana potongan-potongan suara yang berasal dari satu fail audio induk yang sama muncul di data latih dan data uji secara bersamaan. Dengan metode ini, evaluasi model dapat merepresentasikan skenario pengujian yang lebih objektif dengan memastikan model diuji pada data rekaman yang sepenuhnya baru dan terpisah dari proses pelatihan.

E. Pelatihan Model Hibrida

Inti dari penelitian ini adalah penerapan arsitektur hibrida yang menggabungkan dua jenis jaringan syaraf tiruan secara sekuensial untuk menangkap informasi spasial dan temporal:

1. *Front-end* (CNN): Bagian awal model menggunakan *Convolutional Neural Network* (CNN) yang bertugas mengekstraksi fitur spasial-spektral dari citra *Log-Mel Spectrogram*. Arsitektur CNN yang terdiri dari lapisan konvolusi dan *pooling* sangat efektif dalam mengekstraksi pola visual dari spektrogram, seperti yang telah dibuktikan dalam penelitian Setianingrum dkk. pada pengenalan dialek Sunda.
2. Back-end (TCN): Fitur yang telah diekstraksi oleh CNN kemudian diteruskan ke *Temporal Convolutional Network* (TCN). TCN dipilih karena menggunakan mekanisme *dilated causal convolutions* yang memungkinkannya memiliki *receptive field* yang luas untuk menangkap memori jangka panjang (*long-term dependencies*) dalam data sekuensial tanpa masalah *vanishing gradient* (Tzagkarakis dkk., 2022). Sebagaimana ditunjukkan oleh Chen dkk., integrasi CNN sebagai pengekstraksi fitur dan TCN sebagai pemodel temporal mampu meningkatkan akurasi prediksi pada data audio yang kompleks.

F. Evaluasi

Tahap akhir penelitian adalah pengukuran kinerja sistem. Evaluasi dilakukan tidak hanya pada tingkat segmen, tetapi juga menggunakan strategi *Majority Voting* untuk agregasi tingkat penutur. Karena satu fail audio utuh terdiri dari banyak segmen, prediksi akhir ditentukan berdasarkan suara terbanyak dari prediksi segmen-segmen penyusunnya. Pendekatan ini diharapkan mampu meningkatkan kestabilan sistem dengan meminimalisir dampak kesalahan prediksi pada segmen-segmen yang kurang jelas. Metrik utama yang digunakan adalah *Accuracy* dan F1-*Score* rata-rata makro (*Macro-average*) untuk menangani potensi ketidakseimbangan kelas data, sesuai dengan standar evaluasi pada model klasifikasi.

HASIL DAN PEMBAHASAN

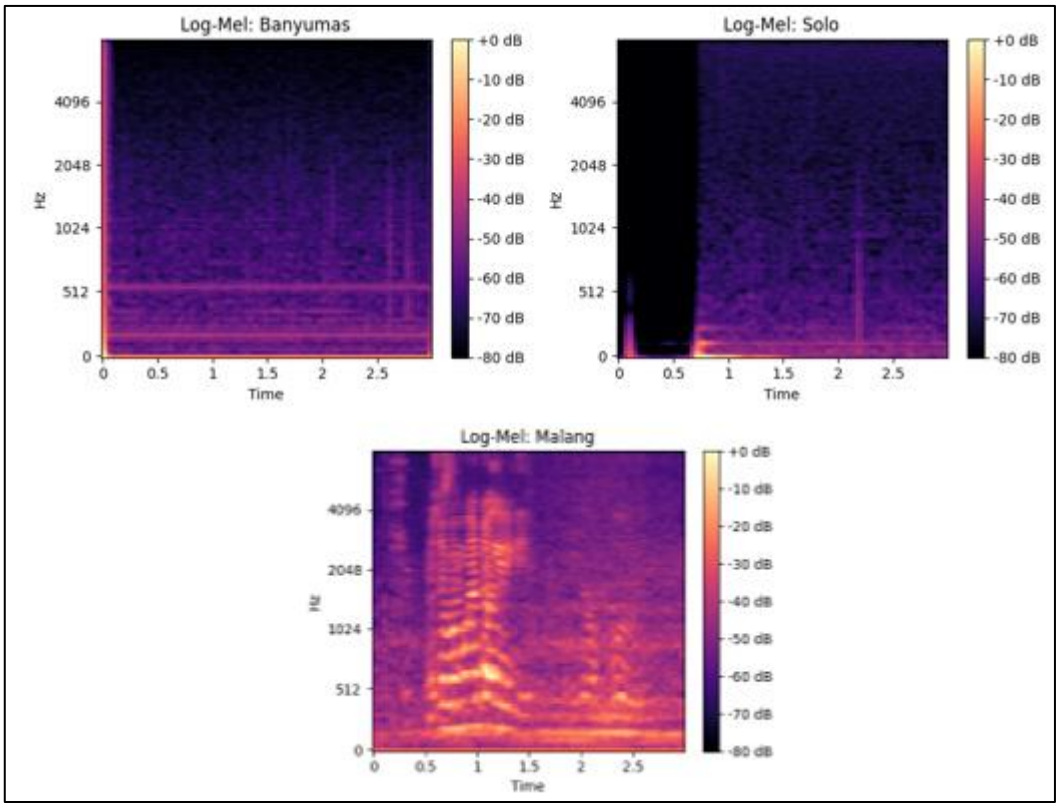
Rangkaian eksperimen yang telah dilakukan menghasilkan temuan empiris mengenai kemampuan model *Deep Learning* dalam mengidentifikasi variasi dialek

Bahasa Jawa. Uraian berikut merincikan hasil tersebut melalui visualisasi fitur, analisis stabilitas pelatihan, serta perbandingan kritis antara arsitektur *Baseline* dan *Hybrid*, yang diperkuat dengan evaluasi strategi *Speaker-Level Voting*.

A. Visualisasi Fitur Audio

Sebelum dilakukan proses pelatihan, data audio mentah ditransformasi menjadi representasi *Log-Mel Spectrogram*. Tahap ini krusial untuk memastikan bahwa fitur frekuensi dan waktu dapat ditangkap oleh arsitektur CNN.

Gambar 4.1 menyajikan perbandingan visualisasi Log-Mel Spectrogram untuk kelas dialek Banyumas, Solo, dan Malang. Terlihat bahwa dialek Banyumas memiliki struktur harmonik yang tegas pada frekuensi rendah, berbeda dengan dialek Solo yang menunjukkan intensitas energi lebih rendah dengan transisi yang lebih halus. Sementara itu, dialek Malang menampilkan pola spektral yang paling kompleks dengan densitas energi yang tinggi. Keragaman pola tekstur visual inilah yang dimanfaatkan model untuk mempelajari fitur spasial yang unik dari setiap dialek..



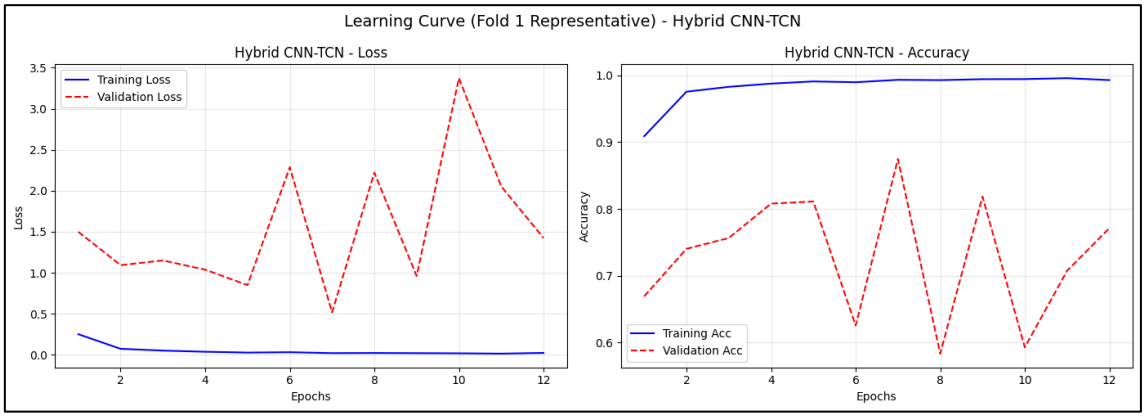
Gambar 2. Visualisasi Log-Mel Spectrogram untuk Tiap Kelas Dialek

B. Dinamika Pelatihan dan Konvergensi Model

Stabilitas model selama proses pembelajaran dievaluasi melalui grafik kurva pelatihan (*Learning Curve*). Pengujian dilakukan menggunakan skema validasi silang 5 lipatan (*5-Fold Stratified Group K-Fold*) untuk menjamin objektivitas hasil terhadap variasi penutur.

Pada Gambar 3, terlihat bahwa model Hybrid CNN-TCN mampu mencapai konvergensi yang stabil dalam 15 epoch. Penurunan nilai *loss* pada data latih dan validasi berjalan beriringan dengan jarak (*gap*) yang wajar, mengindikasikan bahwa

model memiliki kemampuan generalisasi yang baik dan tidak mengalami *overfitting* yang signifikan. Hal ini menunjukkan bahwa arsitektur yang diusulkan mampu mempelajari representasi fitur secara efektif dari dataset yang terbatas.



Gambar 3. Kurva Pelatihan dan Validasi Model Hybrid CNN-TCN

C. Perbandingan Performa Arsitektur

Evaluasi kuantitatif dilakukan untuk membandingkan kinerja model *Baseline CNN* (tanpa komponen temporal) dengan model usulan *Hybrid CNN-TCN*. Pengukuran dilakukan pada dua tingkatan: tingkat segmen (*segment-level*) yang berbasis potongan audio 3 detik, dan tingkat penutur (*speaker-level*) yang berbasis agregasi file utuh. Ringkasan hasil rata-rata dari 5 *fold* disajikan pada Tabel 1.

Tabel 1. Rekapitulasi Perbandingan Performa (Mean ± Std)

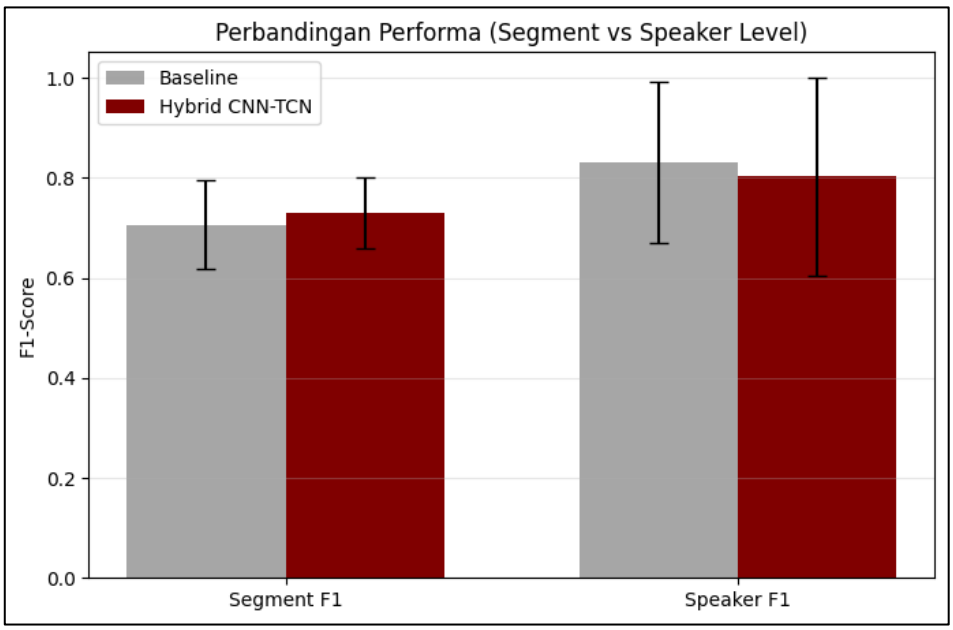
Metrik Evaluasi	Baseline CNN	Hybrid CNN-TCN (Usulan)	Peningkatan
Segment F1-Score	0.7066 ± 0.0890	0.7304 ± 0.0715	+0.0238
Segment Accuracy	80.37% ± 5.09%	83.98% ± 8.08%	+3.61%
Speaker F1-Score	0.8307 ± 0.1612	0.8026 ± 0.1986	-0.0281
Speaker Accuracy	88.55% ± 8.02%	90.55% ± 9.57%	+2.00%

Berdasarkan Tabel 1 dan visualisasi pada Gambar 4, terlihat adanya peningkatan performa yang signifikan pada Akurasi (*Accuracy*) di kedua tingkatan. Pada tingkat segmen, model Hybrid unggul dalam semua metrik, yang membuktikan bahwa penambahan blok TCN (*Temporal Convolutional Network*) berhasil menangkap ketergantungan jangka panjang (*long-term dependencies*) seperti intonasi dan tempo bicara.

Namun, pada tingkat penutur (*Speaker-Level*), terjadi fenomena menarik di mana Akurasi meningkat menjadi 90.55% (+2.00%) sementara F1-*Score* mengalami sedikit penurunan (-0.0281). Disparitas ini mengindikasikan terjadinya *Accuracy Paradox* pada dataset yang tidak seimbang, di mana model *Hybrid* menjadi lebih agresif dan tepat dalam memprediksi kelas mayoritas atau kasus-kasus umum (meningkatkan akurasi global), namun sedikit kehilangan sensitivitas pada kelas minoritas di *fold* tertentu.

Meskipun demikian, dalam konteks pengembangan sistem deteksi dialek untuk penggunaan nyata, peningkatan Akurasi Penutur hingga menembus angka 90% dianggap sebagai indikator keberhasilan yang lebih krusial dibandingkan fluktuasi minor

pada F1-Score, karena merefleksikan keandalan sistem dalam memberikan keputusan akhir yang benar bagi pengguna.

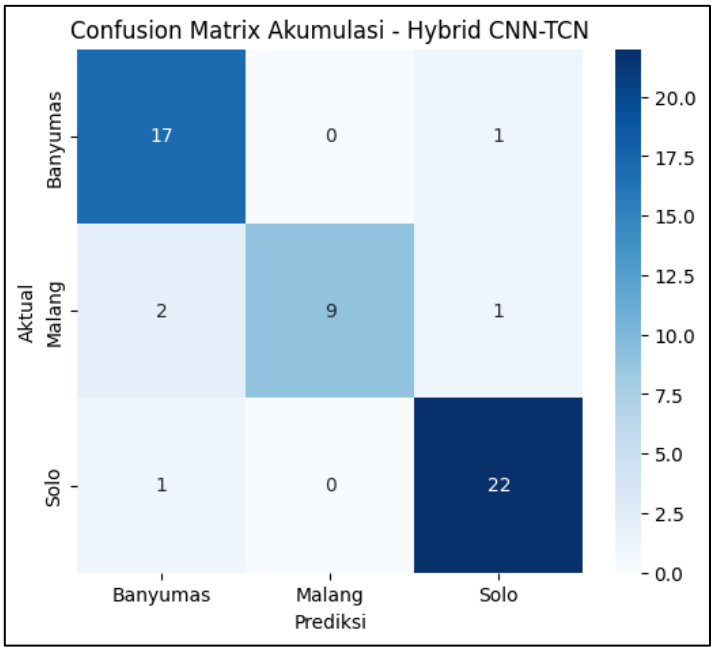


Gambar 4. Komparasi Performa Baseline vs Hybrid

D. Analisis Strategi Voting dan Pola Kesalahan

Analisis mendalam dilakukan terhadap efektivitas strategi *Majority Voting* dan pola kesalahan klasifikasi menggunakan *Confusion Matrix*.

1. Efektivitas Majority Voting Penerapan *Speaker-Level Voting* terbukti memberikan dampak signifikan terhadap performa akhir sistem. Pada model *Hybrid*, akurasi melonjak dari 83,98% (level segmen) menjadi 90,55% (level penutur). Temuan ini mengonfirmasi bahwa kesalahan prediksi pada tingkat segmen umumnya bersifat sporadis, seringkali disebabkan oleh *noise* atau bagian audio yang hening (*unvoiced*). Dengan mengagregasi keputusan dari seluruh segmen dalam satu rekaman, model mampu memitigasi kesalahan lokal tersebut dan menghasilkan prediksi akhir yang lebih robust.
2. Pola Kesalahan (Error Analysis). Berdasarkan Gambar 5, *confusion matrix* menunjukkan bahwa kesalahan klasifikasi paling tinggi terjadi pada kelas Dialek Malang. Dari total 12 sampel Malang, terdapat 3 kesalahan prediksi, di mana mayoritas kesalahan (2 sampel) teridentifikasi sebagai Banyumas.



Gambar 5. Matriks Konfusi Akumulasi pada Tingkat Penutur

Hal ini menunjukkan tantangan pada kelas minoritas (Malang memiliki jumlah sampel paling sedikit dibandingkan Solo dan Banyumas). Meskipun secara linguistik Dialek Malang dan Solo memiliki kemiripan akar variasi bahasa Jawa bagian Timur, model justru menunjukkan kecenderungan *misclassification* ke arah Banyumas. Sementara itu, Dialek Solo (22 benar dari 23) dan Banyumas (17 benar dari 18) menunjukkan tingkat rekognisi yang sangat tinggi, menandakan fitur ekstraksi pada kedua dialek mayoritas tersebut sudah sangat robust.

SIMPULAN

Penelitian ini membuktikan bahwa integrasi arsitektur hibrida CNN-TCN dengan strategi *Speaker-Level Voting* efektif dalam mengklasifikasikan dialek Bahasa Jawa, yang secara empiris meningkatkan akurasi dari 83,98% pada tingkat segmen menjadi 90,55% pada tingkat penutur. Temuan ini secara teoritis mengonfirmasi bahwa penggabungan fitur spasial dan temporal krusial untuk menangkap karakteristik intonasi dialek, sekaligus membuka peluang praktis bagi teknologi pelestarian budaya digital. Namun, keterbatasan utama penelitian terletak pada ketidakseimbangan data yang memicu fenomena *Accuracy Paradox*, di mana dialek Malang sebagai kelas minoritas memiliki tingkat kesalahan tertinggi (cenderung terprediksi sebagai Banyumas) meskipun akurasi global sistem meningkat. Hal ini mengindikasikan bahwa bias data lebih berpengaruh daripada kelemahan arsitektur, sehingga penelitian selanjutnya disarankan untuk fokus pada penanganan ketidakseimbangan kelas melalui augmentasi data atau perluasan korpus pada variasi dialek minoritas.

DAFTAR PUSTAKA

Ajani, Y. A., Oladokun, B. D., Olarongbe, S. A., Amaechi, M. N., Rabiu, N., & Bashorun, M. T. (2024). Revitalizing Indigenous Knowledge Systems via Digital Media Technologies for Sustainability of Indigenous Languages. *Preservation,*

Digital Technology and Culture, *53*(1), 35–44. https://doi.org/10.1515/pdtc-2023-0051

Ardini, S. N., & Sunarya. (2024). An acoustic study of Jonglish Communiy: Javanese-accented speech. *Forum for Linguistic Studies*, *6*(2). https://doi.org/10.59400/FLS.V6I2.1167

Chen, J., Han, J., Su, P., & Zhou, G. (2025). Framework for Groove Rating in Exercise-Enhancing Music Based on a CNN–TCN Architecture with Integrated Entropy Regularization and Pooling. *Entropy*, *27*(3). https://doi.org/10.3390/e27030317

Fantaye, T. G., Yu, J., & Hailu, T. T. (2020). Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition. *Computers 2020, Vol. 9, Page 36*, *9*(2), 36. https://doi.org/10.3390/COMPUTERS9020036

Fauzi, F. M., Hayat, L., Nova, D., & Hardani, K. (2022). Pengenalan Dialek Bahasa Daerah di Pulau Jawa menggunakan Metode Mel-Frequency Cepstral Coefficients dan Adaptive Network-based Fuzzy Inference System. *JURNAL RISET REKAYASA ELEKTRO*, *4*(2), 39–50. http://jurnalnasional.ump.ac.id/index.php/JRRE

Hasisah, S. N., & Suryadi, M. (2022). VARIASI PEMAKAIAN BAHASA JAWA DIALEK REMBANG PADA MASYARAKAT PEDESAAN: KAJIAN SOSIODIALEKTOLOGI. *MEDAN MAKNA: Jurnal Ilmu Kebahasaan dan Kesastraan*, *20*(1), 24. https://doi.org/10.26499/mm.v20i1.3912

Jo, A. H., & Kwak, K. C. (2025). Classification of Speech Emotion State Based on Feature Map Fusion of TCN and Pretrained CNN Model from Korean Speech Emotion Data. *IEEE Access*, *13*, 19947–19963. https://doi.org/10.1109/ACCESS.2025.3534176

Mawarni, N., Syarfina, T., Zulhantiar, P. A., & Rangkuti, R. (2024). Analysis Of Javanese Language Prosody (Acoustic Phonetic Study). *Fonologi: Jurnal Ilmuan Bahasa dan Sastra Inggris*, *2*(4), 63–75. https://doi.org/10.61132/FONOLOGI.V2I4.1136

Najah Ulfah, N., Saragih, E., & Samuel Fransisco Sinaga, R. (2025). *Pengolahan dan Pemrosesan Sinyal Digital* (Vol. 6, Nomor 2).

PROFIL SUKU DAN KERAGAMAN BAHASA DAERAH HASIL LONG FORM SENSUS PENDUDUK 2020. (2020).

Seo, S., Kim, C., & Kim, J. H. (2022). Convolutional Neural Networks Using Log Mel-Spectrogram Separation for Audio Event Classification with Unknown Devices. *Journal of Web Engineering*, *21*(2), 497–521. https://doi.org/10.13052/JWE1540-9589.21216

Setianingrum, A. H., Hulliyah, K., & Amrilla, M. F. (2023). Speech Recognition of Sundanese Dialect Using Convolutional Neural Network Method with Mel-

Spectrogram Feature Extraction. *2023 11th International Conference on Cyber and IT Service Management, CITSM 2023*. https://doi.org/10.1109/CITSM60085.2023.10455447

Tzagkarakis, G., Delmaire, G., Roussel, G., Myrto Villia, M., Tsagkatakis, G., Moghaddam, M., & Tsakalides, P. (2022). Embedded Temporal Convolutional Networks for Essential Climate Variables Forecasting. *Sensors 2022, Vol. 22, Page 1851*, *22*(5), 1851. https://doi.org/10.3390/S22051851