Dampak Filter Digital Terhadap Kinerja *Convolutional Neural Network* pada Klasifikasi Suara Lingkungan

http://dx.doi.org/10.28932/jutisi.v12i1.14018

Riwayat Artikel
Received: 08 Desember 2025 | *Final Revision*: 08 April 2026 | *Accepted*: 09 April 2026

I Kadek Arya Sugianta✉

[#] *Program Studi Informatika, Fakultas Bisnis, Sosial, Teknologi dan Humaniora, Universitas Bali Internasional*
Jalan Seroja, Gang Jeruk No.9A, Kelurahan Tonja, Kota Denpasar, 80234, Indonesia
aryabisabikin@gmail.com

✉Corresponding author: aryabisabikin@gmail.com

Abstrak — **Sering kali, teknik pembatasan *bandwidth* standar telefoni diterapkan secara langsung pada sistem *Environmental Sound Classification* (ESC) atau klasifikasi suara lingkungan tanpa validasi empiris yang memadai. Penelitian ini bertujuan untuk mengevaluasi dampak berbagai konfigurasi filter digital *Finite Impulse Response* (FIR) dan *Infinite Impulse Response* (IIR) dengan variasi tipe (*Low-Pass, High-Pass, Band-Pass, Band-Stop*) dan orde (2–128) terhadap kinerja *Convolutional Neural Network* (CNN) pada *dataset Environmental Sound Classification (*ESC-50). Melalui perbandingan fitur, penelitian ini mengonfirmasi superioritas *Log-Mel Spectrogram* dibandingkan *Mel-Frequency Cepstral Coefficients* (MFCC) dalam merepresentasikan fitur spektral spasial. Hasil eksperimen menunjukkan bahwa filter *Band-Pass* standar (300–3400 Hz) dan *Low-Pass Filter* bersifat destruktif karena mengeliminasi komponen frekuensi tinggi esensial pada suara lingkungan yang bersifat *broadband*. Temuan krusial menunjukkan bahwa penggunaan *High-Pass Filter* (HPF) orde rendah (FIR-32) dengan *cut-off* 1000 Hz berhasil meningkatkan akurasi hingga 66,20% di atas *baseline*. Analisis spektral mengungkapkan bahwa konfigurasi ini efektif mereduksi *noise* frekuensi rendah tanpa memicu fenomena *transient smearing* atau distorsi waktu yang merusak fitur *onset*. Penelitian ini merekomendasikan penggunaan *High-Pass Filter* (HPF) orde rendah sebagai standar pra-pemrosesan baru, serta menyarankan transisi menuju *learnable filters* untuk mengatasi rigiditas filter statis di masa depan.**

Kata kunci— ***CNN; ESC-50;* Filter Digital*;* Klasifikasi Suara Lingkungan*; Log-Mel Spectrogram.***

Digital Filter Impact on Convolutional Neural Network Performance for Environmental Sound Classification

Abstract — ***Often, standard telephony bandwidth-limiting techniques are applied directly to Environmental Sound Classification (ESC) systems without adequate empirical validation. This study aims to evaluate the impact of various Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) digital filter configurations—with variations in type (Low-Pass, High-Pass, Band-Pass, Band-Stop) and order (2–128)—on the performance of Convolutional Neural Networks (CNNs) on the Environmental Sound Classification (ESC-50) dataset. Through feature comparison, this study confirms the superiority of the Log-Mel Spectrogram over Mel-Frequency Cepstral Coefficients (MFCC) in representing spatial spectral features. Experimental results show that the standard Band-Pass filter (300–3400 Hz) and Low-Pass Filter are destructive because they eliminate essential high-frequency components in broadband environmental sounds. A crucial finding indicates that the use of a low-order High-Pass Filter (HPF) (FIR-32) with a***

cutoff of 1000 Hz successfully improved accuracy by up to 66.20% over the baseline. Spectral analysis reveals that this configuration effectively reduces low-frequency noise without triggering transient smearing or time distortion that degrades onset features. This study recommends the use of a low-order High-Pass Filter (HPF) as a new preprocessing standard and suggests a transition toward learnable filters to address the rigidity of static filters in the future

Keywords— ***CNN; Digital Filter; Environmental Sound Classification; ESC-50; Log-Mel Spectrogram.***

I. PENDAHULUAN

Environmental Sound Classification (ESC) memegang peran krusial dalam komputasi audio; tujuannya tidak terbatas pada identifikasi sinyal non-wicara, melainkan mencakup kategorisasi spektrum luas mulai dari suara antropogenik (seperti lalu lintas) hingga fenomena geofisika. Secara spesifik, domain ini telah mengalami transisi masif dari pemrosesan sinyal klasik menuju arsitektur *Deep Learning* tingkat lanjut; sebuah pergeseran yang mencerminkan signifikansi ESC dalam ekosistem modern. Urgensi pengembangannya pun kian tak terelakkan akibat integrasi infrastruktur *smart city* dan AIoT, yang menuntut hadirnya sistem keselamatan publik serta pemantauan lingkungan yang tangguh [1], [2].

Interferensi derau latar belakang (*background noise*) menjadi hambatan utama dalam implementasi ESC; fenomena ini kerap mendegradasi kinerja model klasifikasi secara signifikan [3], [4]. Merespons masalah tersebut, paradigma konvensional biasanya menerapkan filter digital seperti *Low-Pass* atau *Band-Pass* sebagai solusi standar guna mereduksi *bandwidth* sinya [5], [6]. Namun, terdapat perbedaan fundamental: sinyal suara lingkungan tidaklah terstruktur layaknya sinyal wicara, melainkan bersifat *broadband* dan acak [7]. Kondisi ini memicu ambiguitas serius terkait efektivitas metode tersebut: apakah reduksi *bandwidth* benar-benar meningkatkan *Signal-to-Noise Ratio* (SNR) demi akurasi, atau sebaliknya, justru menghilangkan fitur spektral vital yang berujung pada degradasi model.

Meskipun teknik pra-pemrosesan audio untuk arsitektur *Deep Learning* telah berkembang pesat, literatur saat ini masih menunjukkan kesenjangan kritis dalam memahami dampak mekanis filter digital terhadap integritas fitur audio. Penelitian terbaru oleh Galic dkk. [8] mengeksplorasi penggunaan *inverse filtering* sebagai strategi augmentasi data untuk mengenali suara bisikan, namun studi tersebut lebih berfokus pada transformasi karakteristik vokal manusia dan mengabaikan parameterisasi orde filter. Di sisi lain, Bautista dkk. [9] mendemonstrasikan efektivitas jaringan *Paralel CNN-Attention* dengan augmentasi data yang masif, namun pendekatan tersebut memperlakukan input audio sebagai entitas statis tanpa mengevaluasi bagaimana distorsi pra-pemrosesan dapat mengaburkan fitur transien yang menjadi target mekanisme *attention.*

Pentingnya integritas fase dalam representasi audio sebenarnya telah disinggung oleh Rajan dan Sivan [10] melalui penggunaan *Modgdgram* (*Group Delay*) untuk pengenalan musik. Namun, terdapat kekosongan literatur (*research gap*) yang signifikan mengenai bagaimana desain filter konvensional (FIR dan IIR) yang sering digunakan untuk reduksi *noise* secara tidak sengaja mendistorsi respon fase dan *group delay* sinyal lingkungan yang bersifat *broadband.*

Kebaruan (*novelty*) penelitian ini terletak pada pengungkapan korelasi negatif antara orde filter tinggi dengan akurasi klasifikasi CNN. Berbeda dengan asumsi umum bahwa filter yang lebih tajam (orde tinggi) menghasilkan sinyal yang lebih bersih, penelitian ini membuktikan secara eksperimental bahwa peningkatan orde filter memicu fenomena *transient smearing*. Fenomena ini mengaburkan pola spasial pada spektrogram yang sangat krusial bagi CNN untuk mengenali kelas suara impulsif pada *dataset* ESC-50 [11].

Urgensi evaluasi tersebut berakar kuat pada mekanisme fundamental arsitektur *Deep Learning* modern; khususnya CNN yang memang didesain spesifik untuk memproses data bertopologi grid. Kendati awalnya dikembangkan untuk tugas visi komputer, CNN memiliki kapasitas unik untuk mengekstraksi fitur hierarkis secara otomatis dari data input [12], [13]. Dalam konteks klasifikasi audio, kapabilitas tersebut diadopsi melalui transformasi sinyal suara menjadi representasi waktu-frekuensi, seperti *Log-Mel Spectrogram*. Akibatnya, representasi ini tidak lagi dipandang sebagai gelombang audio biasa, melainkan diperlakukan layaknya citra dua dimensi; di sini, sumbu waktu dan frekuensi dipetakan sebagai koordinat spasial yang memuat pola akustik vital [14].

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada analisis kritis terhadap dampak pra-pemrosesan filter digital, baik *Finite Impulse Response* (FIR) maupun *Infinite Impulse Response* (IIR), terhadap kinerja CNN pada *dataset* ESC-50. Secara eksplisit, penelitian ini merumuskan masalah utama pada ketidakpastian batasan efektivitas filter konvensional dalam mempertahankan integritas fitur temporal dan spektral yang dibutuhkan oleh model *Deep Learning.* Masalah ini diturunkan ke dalam dua pertanyaan penelitian: (1) sejauh mana variasi orde filter memicu distorsi *ringing* yang mendegradasi fitur transien suara lingkungan, dan (2) bagaimana karakteristik respons frekuensi dari berbagai filter (LPF, HPF, BFS, BSF) secara selektif mengeliminasi komponen akustik esensial pada kelas suara yang berbeda.

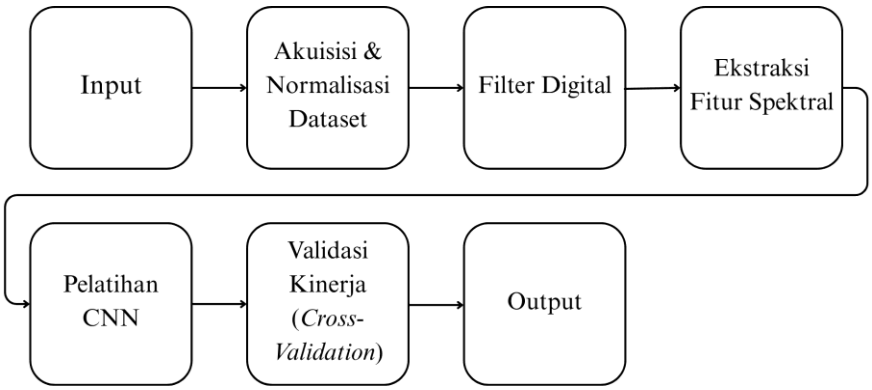
Tujuan utama penelitian ini adalah untuk mengevaluasi secara sistematis dan granular pengaruh parameterisasi filter terhadap akurasi klasifikasi CNN. Secara spesifik, eksperimen ini bertujuan untuk mengidentifikasi *trade-off* optimal antara reduksi *noise* dan preservasi fidelitas sinyal melalui variasi tipe filter dan orde yang ketat (FIR:32-128; IIR:2-8). Skenario

eksperimental ini diterapkan untuk membedah kontribusi elemen spektral, mulai dari dominasi frekuensi rendah pada suara alam hingga isolasi komponen menengah pada suara perkotaan. Pada akhirnya, hasil investigasi ini diharapkan mampu memberikan wawasan empiris baru mengenai batasan metode konvensional, serta merumuskan rekomendasi berbasis data terkait urgensi pelestarian integritas spektral dalam ekosistem *Deep Learning*

II. METODE PENELITIAN

A. Alur Penelitian

Penelitian ini berfokus pada evaluasi kinerja klasifikasi akibat dampak reduksi *bandwidth* sinyal; sebuah pendekatan eksperimental kuantitatif pun diterapkan untuk menjawab masalah tersebut. Secara spesifik, alur sistematis riset tidak dilakukan sekaligus, melainkan terbagi ke dalam lima fase krusial. Tahap awal dimulai dari akuisisi dan normalisasi *dataset*, yang segera diikuti oleh penerapan bank filter digital dengan parameter variatif. Setelah fitur spektral berhasil diekstraksi, proses berlanjut ke pelatihan model *Convolutional Neural Network* (CNN). Terakhir, validasi kinerja dilakukan menggunakan skema *Cross-Validation* seperti yang tervisualisasi pada diagram blok di Gambar 1.



Gambar 1.Diagram Blok Fase Penelitian

B. Dataset dan Persiapan Data

Dataset utama yang didayagunakan dalam penelitian ini adalah ESC-50; koleksi ini mencakup 2.000 rekaman audio lingkungan berdurasi 5 detik dengan frekuensi sampel asli 44,1 kHz. Struktur data tidak disusun secara acak, melainkan terkelompok ke dalam 50 kelas semantik di bawah lima kategori mayor: mulai dari suara hewan, fenomena alam, hingga eksterior perkotaan [11]. Guna menjamin konsistensi *input* pada model *Deep Learning*, diterapkan rangkaian pra-pemrosesan yang ketat. Tahap pertama dimulai dengan *resampling* seluruh audio menjadi 16 kHz; keputusan ini didasarkan pada prinsip Nyquist yang dinilai cukup untuk menangkap mayoritas informasi suara lingkungan sekaligus mereduksi beban komputasi [15], [16]. Proses kemudian berlanjut ke tahap segmentasi. Durasi asli 5 detik tidak dipertahankan, melainkan diseragamkan menjadi 4 detik (*trimming*) untuk seluruh sampel. Langkah strategis ini diambil demi menjaga efisiensi komputasi tanpa mengorbankan fitur semantik utama, serta memastikan kompatibilitas dimensi untuk potensi generalisasi silang (*cross-generalization*) dengan dataset standar lain seperti UrbanSound8K

C. Desain Eksperimen Filter Digital

Tahap ini merupakan inti dari investigasi; serangkaian filter digital pun diterapkan pada domain waktu sebelum ekstraksi fitur guna menguji hipotesis dampak reduksi *bandwidth*. Desain filter tidak terbatas pada satu pendekatan, melainkan dibagi ke dalam dua arsitektur utama. Pertama, arsitektur *Finite Impulse Response* (FIR) diterapkan, di mana hubungan input $x[n]$ dan output $y[n]$ diatur oleh operasi konvolusi diskrit

$$y[n] = \sum_{k=0}^{N} b_k \cdot x[n-k] \qquad (1)$$

Dalam implementasinya, metode *Hamming Window* dipilih secara khusus untuk meminimalkan kebocoran spektral (*spectral leakage*) pada penentuan koefisien b_k [17]. Pemilihan ini didasarkan pada kemampuan *Hamming Window* dalam menyeimbangkan lebar *main-lobe* dan penekanan *side-lobe*. Secara spesifik, penelitian terbaru menunjukkan bahwa *Hamming Window* mampu menghasilkan redaman *side-lobe* relatif hingga -42,1 dB dengan faktor kebocoran yang sangat

rendah, yaitu sebesar 0,04%. Karakteristik ini jauh lebih unggul dibandingkan *Rectangular window* (-13 dB) dalam konteks penekanan *side-lobe* [18]. Hal ini menjadi sangat krusial dalam klasifikasi suara lingkungan untuk mencegah energi dari frekuensi dominan bocor ke bin frekuensi tetangga. Melalui pemeliharaan integritas fitur spektral pada *Log-Mel Spectrogram* dari gangguan artefak komputasi, CNN dapat lebih efektif dalam menangkap fitur frekuensi yang kritis dan relevan, terutama pada sinyal lingkungan yang memiliki struktur waktu-frekuensi yang kompleks [19]. Variasi orde yang diuji mencakup 32, 64, dan 128 *taps*. Berbeda halnya dengan FIR, arsitektur kedua yang diuji adalah *Infinite Impulse Response* (IIR) yang memanfaatkan umpan balik (*feedback*) dari output sebelumnya.

$$y[n] = \frac{1}{a_0}\left(\sum_{k=0}^{P} b_k \cdot x[n-k] - \sum_{k=1}^{Q} a_k \cdot y[n-k]\right) \quad (2)$$

Pada tahap ini, desain Butterworth diadopsi sebagai standar. Adopsi ini bukan tanpa dasar; karakteristik respons frekuensinya yang datar (*maximally flat*) pada *passband*. Karakteristik ini dinilai krusial untuk memastikan bahwa filter tidak memperkenalkan fluktuasi amplitudo pada frekuensi yang diloloskan, sehingga integritas energi pada *Log-Mel Spectrogram* tetap terjaga sesuai dengan sinyal aslinya [20]. Hal ini sangat penting bagi arsitektur CNN, karena penyimpangan amplitudo sekecil apapun pada domain frekuensi dapat menyebabkan pergeseran distribusi fitur yang berpotensi menurunkan akurasi klasifikasi. Eksperimen dilakukan dengan variasi orde 2,4,6 dan 8 untuk mengevaluasi *trade-off* antara kecuraman *roll-off* dan potensi munculnya distorsi fase atau *ringing effect* yang dapat mengaburkan komponen *transient* pada suara lingkungan.

Eksperimen ini tidak dirancang secara acak, melainkan mencakup empat strategi manipulasi spektral yang berbeda; masing-masing konfigurasi ditargetkan untuk menguji hipotesis spesifik terkait distribusi informasi pada sinyal lingkungan. Mulai dari evaluasi hilangnya detail spasial harmonik hingga simulasi batasan infrastruktur telekomunikasi, seluruh parameter teknis beserta justifikasi analitisnya dirangkum secara sistematis dalam Tabel 1.

TABEL 1
MATRIKS KONFIGURASI FILTER DAN TUJUAN EVALUASI SPEKTRAL

Tipe Filter	Parameter Cut-off / Rentang (Hz)	Tujuan Investigasi & Justifikasi Spektral
Low-Pass Filter (LPF)	1k, 2k, 4k	Evaluasi Informasi Frekuensi Tinggi: Menguji dampak penghilangan detail spasial dan harmonik halus yang krusial bagi suara biakustik (seperti serangga dan burung).
High-Pass Filter (HPF)	500, 1k, 2k	Analisis Kontribusi Frekuensi Rendah: Memvalidasi apakah pembuangan energi rendah (gemuruh angin/mesin) meningkatkan SNR, atau justru menghilangkan fitur utama kelas mekanis.
Band-Pass Filter (BPF)	500-2000, 1000-4000	Simulasi Batasan Narrowband: Mereplikasi karakteristik kanal telekomunikasi wicara tradisional [21] guna menguji validitas penggunaannya pada sinyal lingkungan yang bersifat broadband.
Stop-Band Filter (SBF)	1000-2000, 500-4000	Uji Integritas Pita Tengah: Menganalisis ketahanan model CNN terhadap fenomena *spectral notch* (penghilangan pita frekuensi tengah).

D. Ekstraksi Fitur

Pasca tahap pemfilteran, sinyal audio tidak diproses dalam bentuk gelombang mentah, melainkan ditransformasikan ke domain waktu-frekuensi menggunakan representasi *Log-Mel Spectrogram*. Pemilihan transformasi ini bukanlah langkah arbitrer, melainkan didasarkan pada kapasitasnya dalam memodelkan karakteristik persepsi pendengaran manusia yang bersifat non-linier; sebuah aspek di mana resolusi frekuensi tinggi menjadi kurang relevan dibandingkan frekuensi rendah [22]. Mekanisme konversi dari frekuensi linier (f) ke skala Mel (m) tersebut diformulasikan melalui persamaan berikut:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad (3)$$

Namun, pemetaan frekuensi saja dinilai belum cukup. Besaran amplitudo spektrum (S) selanjutnya dikonversi ke skala logaritmik (Desibel). Langkah ini berfungsi krusial untuk memperluas rentang dinamis fitur sekaligus menekan variasi amplitudo yang ekstrem, sebagaimana dideskripsikan pada persamaan:

$$S_{dB} = 10 \cdot \log_{10}(S + \epsilon) \quad (4)$$

Di mana ϵ merupakan konstanta kecil 1×10^{-6} yang disisipkan guna menjamin stabilitas numerik. Penambahan konstanta ini bertujuan untuk mencegah terjadinya kesalahan matematis berupa nilai tak hingga $(-\infty)$ saat operasi logaritma dilakukan pada komponen spektral yang memiliki nilai magnitudo nol atau sangat mendekati nol. Langkah ini dilakukan untuk menghindari kegagalan komputasi pada proses ekstraksi fitur yang dapat menghentikan fase pelatihan model CNN. Secara empiris, fitur *Log-Mel Spectrogram* telah terbukti superior dalam menangkap dinamika temporal sinyal audio; hal ini berkorelasi langsung dengan peningkatan akurasi, ketahanan (*robustness*) di lingkungan bising, serta kompatibilitas tinggi dengan arsitektur *Deep Learning* [23], [24]. Terkait konfigurasi teknis, parameter ekstraksi tidak ditentukan secara acak, melainkan mengadopsi standar arsitektur *Pre-trained Audio Neural Networks* (PANNs) [25] guna menyeimbangkan resolusi spektral dan efisiensi komputasi. Rincian parameter tersebut disajikan secara lengkap pada Tabel 2.

TABEL 2
PARAMETER EKSTRAKSI FITUR LOG-MEL SPECTROGRAM

Parameter	Nilai	Keterangan
Window Size (n_fft)	1024 sampel	Ukuran jendela FFT untuk analisis frekuensi
Hop Length	128 sampel	Jarak antar frame, menghasilkan ~500 frame untuk durasi 4 detik
Mel Bands	128 bin	Jumlah filter bank Mel untuk representasi spektral
Normalisasi	Z-score normalization	mean = 0, std = 1 untuk mempercepat konvergensi model

Kulminasi dari rangkaian ekstraksi fitur ini mewujud dalam bentuk tensor input berdimensi (128 x 500); sebuah struktur data yang secara spesifik merepresentasikan 128 pita frekuensi Mel dan 500 *frame* waktu. Secara operasional, tensor tersebut tidak lagi diperlakukan sebagai sinyal audio temporal konvensional, melainkan difungsikan sebagai input visual dua dimensi yang menjadi basis pembelajaran bagi model CNN.

E. Arsitektur Model Convolutional Neural Network (CNN)

Model klasifikasi dalam penelitian ini mengadopsi arsitektur CNN-14, yang merupakan varian dari PANNs [25]. Arsitektur ini dirancang khusus untuk memproses input spektrogram dengan mempertahankan invariansi translasi (*translation invariance*). Proses ekstraksi fitur dilakukan melalui operasi konvolusi 2D. Untuk input spektrogram I dan kernel K berukuran $(u\ x\ v)$, nilai peta fitur (*feature map*) pada posisi (i, j) dihitung berdasarkan persamaan berikut:

$$(I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \qquad (5)$$

Secara keseluruhan, arsitektur model ini dibangun dari empat blok konvolusi utama yang disusun secara sekuensial. Setiap blok dirancang dengan struktur lapisan yang seragam untuk memaksimalkan efisiensi ekstraksi fitur dari input spektrogram. Rincian konfigurasi lapisan, parameter, serta fungsi dari setiap komponen dalam blok tersebut disajikan pada Tabel 3.

TABEL 3
KONFIGURASI LAPISAN PADA BLOK KONVOLUSI UTAMA

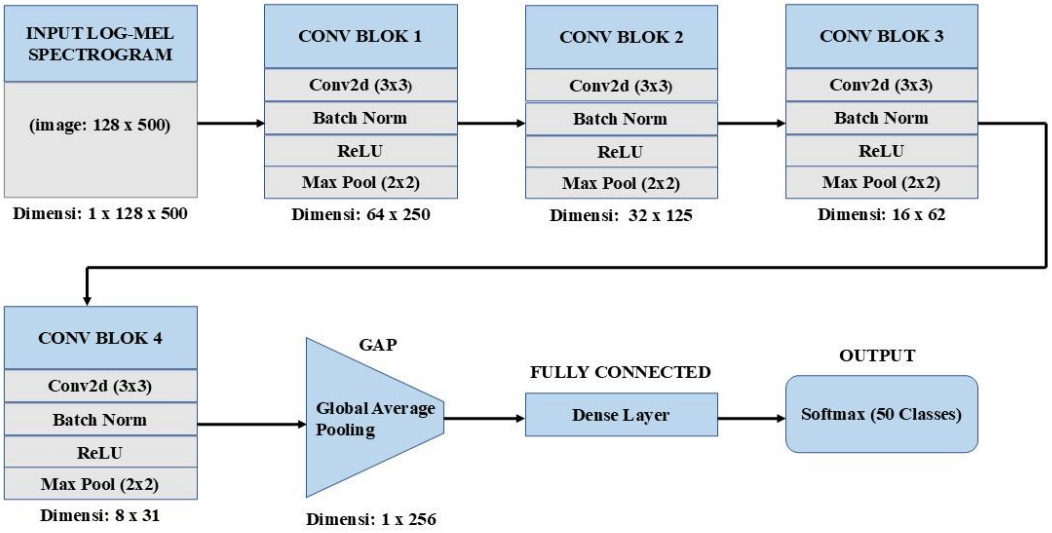
Urutan	Nama Lapisan	Parameter / Konfigurasi	Fungsi Utama
1	Conv2d	Kernel: 3×3Fitur: Padding aktif	Mengekstraksi fitur spasial lokal tanpa mereduksi dimensi input secara signifikan.
2	Batch Normalization	Normalisasi distribusi aktivasi	Menstabilkan proses pelatihan dan mempercepat konvergensi model.
3	ReLU Activation	f(x)=max(0,x)	Menambahkan non-linieritas agar model dapat mempelajari pola fitur yang kompleks.
4	Max Pooling	Kernel: 2×2	Melakukan downsampling (reduksi dimensi) dan mempertahankan fitur dominan.

Setelah blok konvolusi terakhir, mekanisme *Global Average Pooling* (GAP) diterapkan untuk mengagregasi informasi spasial. GAP dipilih karena efektif menangani variasi posisi temporal suara [26]. Dari peta fitur terakhir yang memiliki 256 kanal dengan dimensi spasial (8 x 31), GAP mereduksi setiap kanal menjadi nilai skalar, menghasilkan vektor fitur 1 x 256 yang memuat karakteristik spektral global. Vektor ini kemudian diteruskan ke lapisan *Fully Connected* dan fungsi aktivasi *Softmax* untuk mengklasifikasikan input ke salah satu dari 50 kelas ESC-50. Visualisasi arsitektur utama model *Convolutional Neural Network* (CNN) ini dapat dilihat pada Gambar 2.

F. Skenario Evaluasi

Sesuai standar *dataset* ESC-50, penelitian ini menerapkan skema *5-Fold Cross-Validation*. Metode ini krusial untuk memitigasi risiko bias dan kebocoran data (*data leakage*) antar-sampel yang berasal dari fail audio yang sama [27], [28] . Secara teknis, *dataset* dipartisi menjadi lima *folds*. Di setiap iterasi, model dilatih menggunakan 1.600 sampel (4 *folds*) dan diuji pada 400 sampel (1 *fold*). Rotasi dilakukan lima kali hingga semua bagian teruji. Metrik utama yang digunakan adalah rata-rata Akurasi dari kelima *fold* tersebut, yang didefinisikan sebagai:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \qquad (6)$$



Gambar 2. Diagram Blok Arsitektur Utama CNN

Dengan *TP* (*True Positive*), *TN* (*True Negative*), *FP* (*False Negative*), dan *FN* (*False Negative*) merepresentasikan paramater hasil prediksi model. Selain akurasi global, analisis Confusion Matrix juga dilakukan untk membedah distribusi kesalahan klasifikasi pada tingkat kelas, guna mengidentifikasi dampak spesifik dari masing-masing filter terhadap kategori suara tententu.

III. HASIL DAN PEMBAHASAN

A. Analisis Komparatif Fitur Baseline

Sebagai langkah awal sebelum penerapan filter digital, dilakukan studi komparatif antara dua fitur akustik standar: MFCC dan *Log-Mel Spectrogram*. Kedua fitur ini diekstraksi langsung dari sinyal audio mentah (*raw waveform*) tanpa adanya pemrosesan awal. Evaluasi kinerja didasarkan pada akurasi puncak (*peak accuracy*) yang diperoleh melalui serangkaian eksperimen *5-fold cross-validation*, dengan dukungan optimasi *learning rate scheduler* serta regulasi *weight decay*. Tahapan ini krusial guna menetapkan fondasi input yang paling optimal bagi arsitektur CNN. Hasil perbandingan akurasi rata-rata dari kedua representasi fitur tersebut dirangkum dalam Tabel 4.

TABEL 4
PERBANDINGAN AKURASI FITUR BASELINE PADA *DATASET* ESC-50

Jenis Fitur	Dimensi Input	Akurasi Rata-rata (%)
MFCC	(40 x 500)	62.55%
Log-Mel Spectrogram	(128 x 500)	63.30%

Hasil eksperimen memperlihatkan bahwa representasi *Log-Mel Spectrogram* mengungguli MFCC secara signifikan dengan margin keunggulan akurasi sebesar +0.75%. Meskipun MFCC menawarkan stabilitas konvergensi berkat dimensinya yang ringkas, penelitian ini tetap memprioritaskan *Log-Mel Spectrogram* karena pertimbangan kekayaan

informasi sinyal. Hal ini didasarkan pada karakteristik *Log-Mel Spectrogram* yang mempertahankan korelasi spasial antara bin frekuensi, yang sangat krusial bagi arsitektur CNN dalam mengenali pola tekstur waktu-frekuensi yang kompleks. Sebaliknya, tahap *Discrete Cosine Transform* (DCT) pada MFCC cenderung mereduksi informasi spektral yang bersifat diskriminatif pada suara lingkungan demi mencapai dekorrelasi sinyal. Karakteristik suara lingkungan dalam *dataset* ESC-50 yang bersifat *non-stasioner* dan memiliki spektrum yang luas menyebabkan pelestarian intergritas data mentah pada domain Mel menjadi lebih efektif bagi model. Pendekatan ini memungkinkan ekstraksi fitur-fitur unik dengan lebih optimal dibandingkan representasi MFCC yang terlalu terkompresi.

Temuan ini sekaligus memvalidasi hipotesis bahwa arsitektur CNN, yang mengandalkan ekstraksi fitur spasial lokal, mendapatkan manfaat lebih besar dari resolusi spektral tinggi pada *Log-Mel* (128 *bins*) dibandingkan representasi terkompresi MFCC (40 *bins*). Secara teoritis, penerapan *Discrete Cosine Transform* (DCT) pada MFCC memang bertujuan mendekorelasikan fitur, namun proses ini cenderung mengaburkan detail tekstur halus (*fine-grained texture*) di domain frekuensi. Padahal, dalam klasifikasi suara lingkungan yang kompleks, nuansa tekstur spektral tersebut merupakan fitur diskriminatif yang vital. Atas dasar keunggulan performa dan ketersediaan detail fitur inilah, *Log-Mel Spectrogram* ditetapkan sebagai input standar untuk seluruh eksperimen filter selanjutnya.

B. Evaluasi Kuantitatif Dampak Filter Digital

Temuan eksperimental menyoroti fenomena menarik: penerapan *High-Pass Filter* (HPF) orde rendah terbukti mampu mendongkrak akurasi model melampaui batas *baseline*. Konfigurasi HPF-1000-FIR32 mencatatkan performa terbaik dengan akurasi 66.20%. Hal ini mengindikasikan bahwa atenuasi frekuensi di bawah 1000 Hz efektif dalam mereduksi derau latar belakang tanpa menghilangkan fitur diskriminatif utama.

Pola sebaliknya terjadi pada penerapan *Low-Pass Filter* (LPF), yang secara konsisten mendegradasi kinerja model hingga titik terendah 48.70% (pada konfigurasi LPF-1000-IIR). Disparitas hasil ini membuktikan bahwa komponen frekuensi tinggi (>1000 Hz) mengandung informasi vital bagi klasifikasi suara lingkungan. Karakteristik ini sangat kontras dengan pemrosesan sinyal wicara, yang umumnya masih toleran terhadap pemangkasan frekuensi tinggi

Lebih jauh lagi, teridentifikasi kecenderungan konsisten di mana filter orde rendah (seperti FIR-32 atau IIR-2) secara umum menghasilkan akurasi yang lebih tinggi dibandingkan filter orde tinggi (misalnya FIR-128 atau IIR-8). Fenomena ini mengindikasikan bahwa transisi *roll-off* yang terlalu tajam pada filter orde tinggi berisiko memicu distorsi fase atau munculnya *ringing artifacts* pada domain waktu. Distorsi tersebut pada akhirnya mengaburkan informasi sinyal, sehingga menyulitkan CNN dalam mengenali pola transien cepat. Sebagai rujukan komparatif, Tabel 5 merangkum performa filter terbaik dan terburuk dari setiap kategori terhadap kondisi *Baseline* (Tanpa Filter).

TABEL 5
PERBANDINGAN KINERJA FILTER DIGITAL TERBAIK PER KATEGORI

Kategori	Filter Terbaik (Konfigurasi)	Akurasi (%)	Delta vs Baseline
High-Pass (HPF)	**FIR Orde 32, Cutoff 1000 Hz**	**66.20%**	**+2.90%**
Band-Stop (BSF)	FIR Orde 32, Rentang 1000-2000 Hz	64.90%	+1.60%
BASELINE	**Raw Audio (*Log-Mel Spectrogram*)**	**63.30%**	**0.00%**
Band-Pass (BPF)	FIR Orde 32, Rentang 1000-4000 Hz	62.10%	-1.20%
Low-Pass (LPF)	IIR Orde 2, Cutoff 1000 Hz	55.10%	-8.20%

C. Analisis Spektral dan Pergeseran Distribusi Frekuensi

Guna memvalidasi temuan kuantitatif pada Tabel 3, dilakukan analisis visual komparatif terhadap representasi *Log-Mel Spectrogram* sebelum dan sesudah pemfilteran. Analisis ini bertujuan untuk mengobservasi dampak filter terhadap struktur harmonik dan tekstur sinyal yang menjadi input bagi CNN. Gambar 3 memvisualisasikan perbandingan respons spektral dari sampel suara *insects* (Kelas 41) dalam tiga kondisi: (A) *Baseline* tanpa filter, (B) Filter Terbaik (*High-Pass* 1000Hz), dan (C) Filter Terburuk (*Low-Pass* 1000Hz).

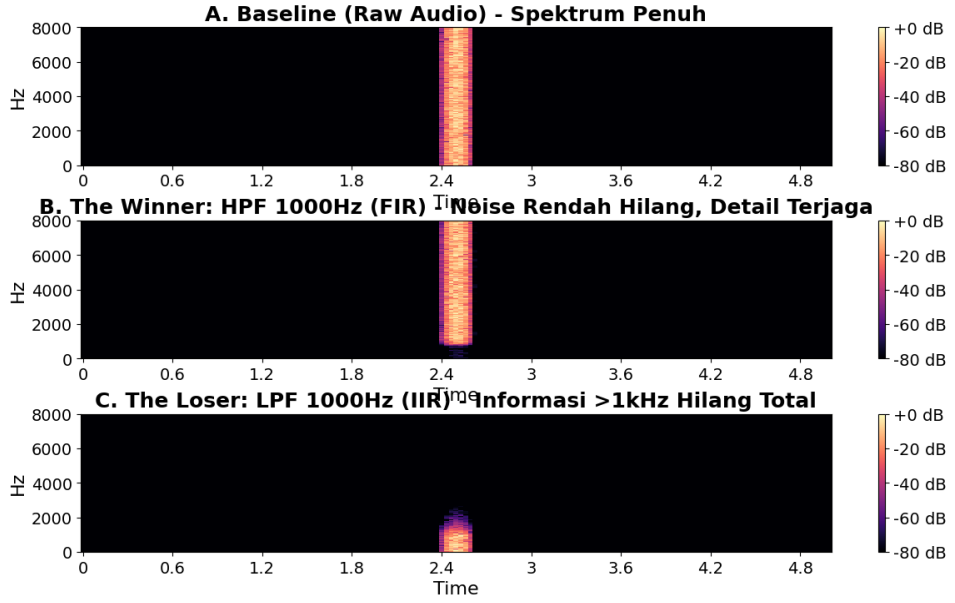
Visualisasi dampak filter terhadap distribusi energi spektral memperlihatkan perbedaan yang mencolok: (A) *Baseline* memuat spektrum frekuensi penuh namun terkontaminasi derau latar, (B) HPF efektif meredam derau frekuensi rendah (<1 kHz) sekaligus menonjolkan fitur harmonik atas, sedangkan (C) LPF menghilangkan mayoritas informasi sinyal (>1 kHz), menyisakan area spektral yang kosong. Berdasarkan observasi visual pada Gambar 3, dapat diuraikan dua fenomena spektral utama yang memengaruhi kinerja model:

1. Peningkatan Kontras Fitur pada HPF (Gambar 3B): Pada konfigurasi *High-Pass Filter* (HPF) dengan *cut-off* 1000 Hz, area spektrum bagian bawah (frekuensi rendah) tampak menjadi gelap, mengindikasikan hilangnya energi pada pita tersebut. Hal ini membuktikan bahwa filter berhasil mengeliminasi derau latar (*background rumble*) dan

interferensi angin yang kerap bersifat stasioner. Pembersihan ini secara efektif meningkatkan *Signal-to-Noise Ratio* (SNR) pada pita frekuensi menengah-tinggi. Akibatnya, tekstur suara jangkrik (pola garis putus-putus pada frekuensi (>4 kHz) menjadi jauh lebih kontras dan mudah diekstraksi oleh lapisan konvolusi CNN. Faktor fisik inilah yang mendorong peningkatan akurasi sebesar +2.9%.

2. Hilangnya Informasi Vital pada LPF (Gambar 3C): Sebaliknya, penerapan *Low-Pass Filter* (LPF) dengan *cut-off* 1000 Hz memicu fenomena *spectral emptiness* (kekosongan spektral) yang masif. Lebih dari 80% area visual spektrogram khususnya frekuensi tinggi yang memuat detail harmonik dan transien cepat terhapus total. Mengingat CNN memproses input layaknya data visual (*computer vision*), spektrogram ini terbaca sebagai gambar kosong dengan minim informasi. Kondisi ini menyebabkan model kehilangan fitur diskriminatif utama untuk membedakan kelas biakustik (seperti serangga atau burung), yang menjelaskan mengapa akurasi anjlok drastis hingga di bawah 50%.

Secara keseluruhan, analisis spektral ini menyimpulkan bahwa dalam domain ESC-50, integritas informasi pada frekuensi tinggi jauh lebih bernilai dibandingkan frekuensi rendah.



Gambar 3. Visualisasi dampak filter terhadap distribusi energi spektral

D. Pengaruh Orde Filter dan Distorsi Transien

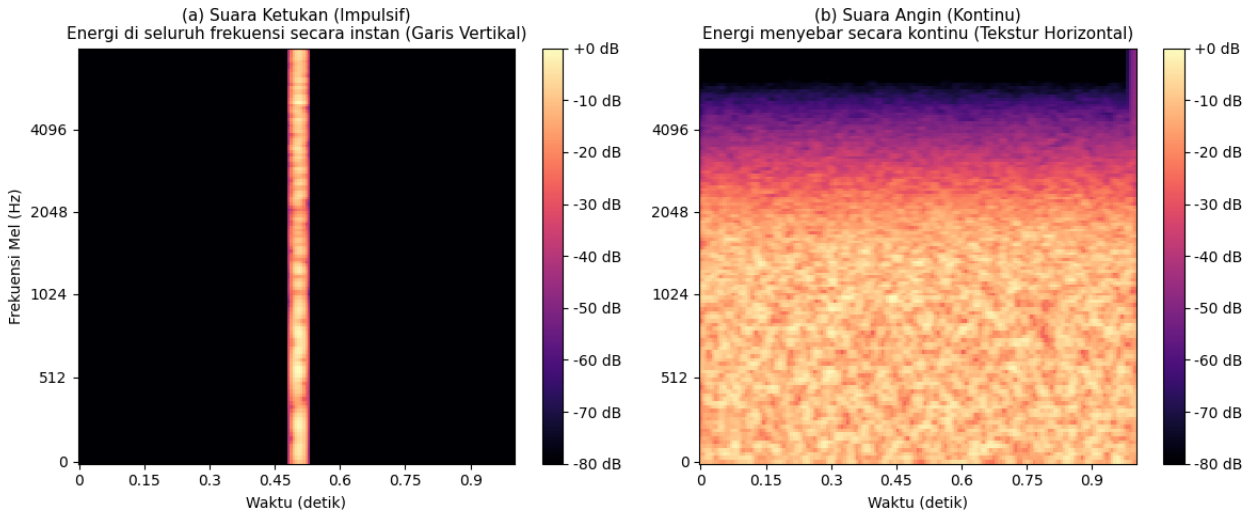
Temuan krusial lainnya dari eksperimen ini adalah adanya korelasi negatif antara orde filter (*filter order*) dan akurasi klasifikasi pada arsitektur CNN. Merujuk pada data evaluasi di Tabel 5, penggunaan filter FIR dengan orde rendah (32 *taps*) secara konsisten mengungguli varian orde tinggi (128 *taps*) dengan margin selisih rata-rata yang signifikan, mencapai ~6,0%. Tren serupa juga teridentifikasi pada filter IIR, di mana Orde 2 menunjukkan stabilitas kinerja yang jauh lebih baik dibandingkan Orde 8.

Secara teoritis, filter orde tinggi memang menawarkan transisi *cut-off* yang lebih curam (*steeper roll-off*) pada domain frekuensi, yang secara konvensional dianggap ideal untuk memisahkan sinyal pada pita sempit Namun, dalam konteks klasifikasi suara lingkungan menggunakan *Log-Mel Spectrogram*, penggunaan orde tinggi justru memicu efek samping detrimental pada domain waktu. Hal ini disebabkan oleh peningkatan latensi fase yang tidak linier (pada IIR) atau *delay* yang panjang (pada FIR), sehingga komponen frekuensi yang berbeda mengalami pergeseran posisi temporal yang tidak seragam. Fenomena ini merusak struktur spasial pada *Log-Mel Spectrogram* yang sangat krusial bagi operasi konvolusi CNN. Dua efek samping utama yang teridentifikasi akibat penggunaan filter orde tinggi adalah sebagai berikut:

1. Fenomena *Ringing* dan *Transient Smearing*: Secara matematis, peningkatan orde filter berbanding lurus dengan panjang respons impuls (*impulse response length*) pada domain waktu. Hubungan ini dapat ditinjau melalui fungsi transfer sistem LTI, di mana untuk filter IIR, durasi peluruhan respons impuls $(h[n])$ ditentukan oleh posisi *pole* yang semakin kompleks seiring bertambahnya orde N mengikuti pola peluruhan:

$$h[n] \approx A.r^n cos(\omega_0 n + \phi) \quad (6)$$

2. Degradasi Fitur Visual pada Spektrogram: Mengingat *dataset* ESC-50 didominasi oleh suara transien tajam seperti gonggongan anjing, pecahan kaca, atau ketukan pintu, dampak artefak ini menjadi krusial. Kondisi tersebut didorong oleh perbedaan fundamental pada struktur waktu-frekuensi antara suara impulsif dan kontinu seperti yang divisualisasikan pada Gambar 4. Suara ketukan (impulsif) secara spektral tidak hanya memiliki beberapa *peak*, melainkan dicirikan oleh garis vertikal tajam karena energinya muncul serentak di hampir semua bin frekuensi dalam durasi yang sangat singkat. Sebaliknya, suara angin bersifat kontinu dengan konsentrasi energi yang menyebar luas sepanjang sumbu waktu namun cenderung dominan pada pita frekuensi rendah. Efek *ringing* menyebabkan garis vertikal pada suara impulsif melebar atau kabur sepanjang sumbu waktu, sehingga fitur visualnya menjadi tumpang tindih dan menyerupai tekstur suara angin yang kontinu. Bagi model CNN yang mengandalkan prinsip deteksi tepi (*edge detection*) untuk mengenali *onset* suara, pengaburan ini merusak ketajaman pola visual dan menurunkan kemampuan model dalam mendiferensiasi jenis suara.



Gambar 4. Perbandingan *Log-Mel Spectrogram* antara suara ketukan yang bersifat impulsif (kiri) dan suara angin yang bersifat kontinu (kanan).

Berdasarkan analisis tersebut, penelitian ini menyimpulkan bahwa dalam aplikasi pengenalan audio berbasis CNN, preservasi integritas transien (resolusi waktu) jauh lebih krusial dibandingkan ketajaman pemisahan frekuensi. Filter orde rendah (FIR-32 atau IIR-2) terbukti memberikan titik keseimbangan (*trade-off*) paling optimal antara reduksi *noise* dan pemeliharaan fidelitas bentuk gelombang.

E. Analisis Kesalahan Spesifik Kelas

Untuk Guna menganalisis dampak filter secara lebih granular, dilakukan evaluasi kinerja per kelas untuk membandingkan model terbaik (*High-Pass Filter* 1000 Hz) terhadap *Baseline*. Kendati filter HPF terbukti meningkatkan akurasi rata-rata secara global, evaluasi mendalam tetap diperlukan untuk memahami pengaruhnya terhadap karakteristik akustik unik pada setiap kelas. Gambar 5 memvisualisasikan hasil analisis *per-class performance* tersebut menggunakan grafik batang divergen (*diverging bar chart*). Grafik ini merepresentasikan selisih (deviasi) akurasi antara model HPF-1000-FIR32 dan model *Baseline.*

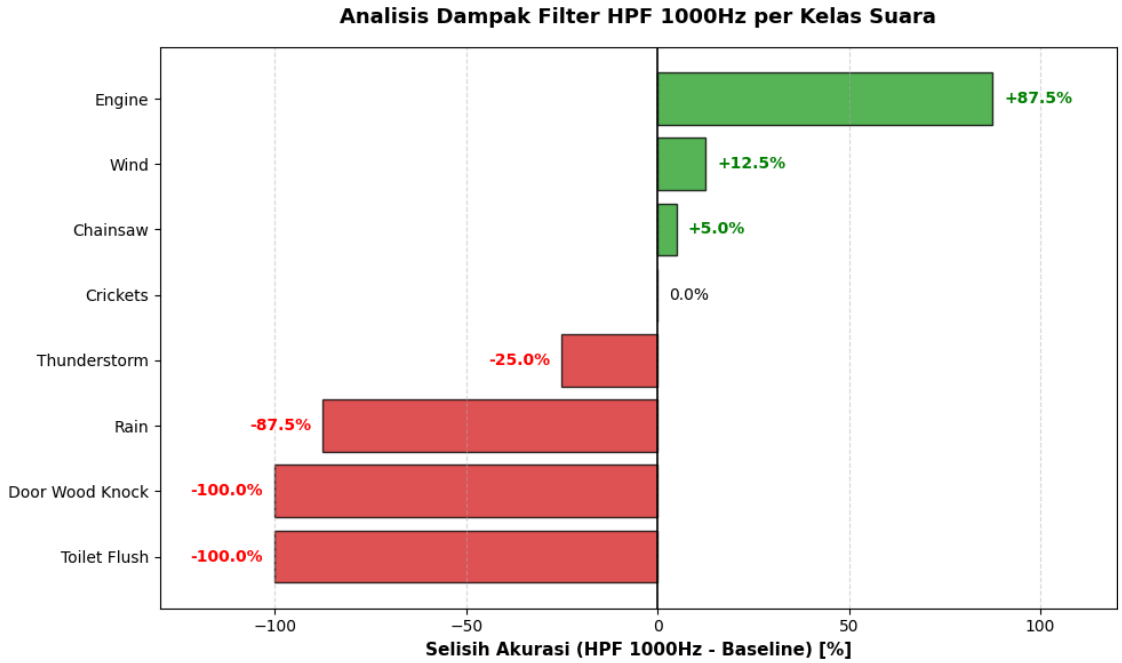
Dalam visualisasi ini, batang hijau merepresentasikan kelas dengan peningkatan kinerja (*gain*), sedangkan batang merah mengindikasikan penurunan (*loss*). Analisis mendalam melalui *Confusion Matrix* (Gambar 4) menyingkap adanya *trade-off* ekstrem pada kelas-kelas tertentu akibat penerapan filter. Penetapan nilai *cut-off* HPF pada 1000 Hz dipilih sebagai titik uji kritis karena secara empiris mayoritas gangguan kebisingan lingkungan, seperti gemuruh mesin dan dengung perangkat elektronik, memiliki akumulasi energi dominan pada rentang frekuensi rendah di bawah ambang tersebut [29].

Peningkatan akurasi paling drastis tercatat pada kelas *Engine*, yang melonjak signifikan sebesar +87.5% (dari 12.5% menjadi 100%). Secara spektral, rekaman suara mesin kerap terkontaminasi oleh gemuruh frekuensi rendah (*low-frequency rumble*) yang dominan namun tidak distingtif. Penerapan HPF 1000 Hz terbukti efektif mengeliminasi komponen gemuruh tersebut, sehingga CNN dapat berfokus sepenuhnya pada fitur harmonik mekanis frekuensi tinggi yang menjadi ciri khas

mesin. Sebaliknya, filter HPF menyebabkan degradasi fatal pada kelas-kelas yang memiliki energi dominan di frekuensi rendah atau bersifat *broadband*:

1. Suara Impulsif Frekuensi Rendah: Kelas *Door Wood Knock* (Ketukan Pintu) mengalami penurunan akurasi hingga -100%. Secara fisik, suara ketukan kayu memiliki frekuensi fundamental di bawah 500 Hz. Dengan *cut-off* filter di 1000 Hz, sinyal utama suara ini terhapus total, menyebabkan model gagal mengenali input tersebut.
2. Suara Air dan Transien: Kelas *Toilet Flush* dan *Rain* juga mengalami penurunan drastis (-87.5% hingga -100%). Suara air mengalir memiliki komponen *white noise* yang tersebar di seluruh spektrum. Pemangkasan frekuensi bawah menyebabkan suara hujan terdengar 'tipis' (menyerupai desis serangga), sehingga memicu kesalahan klasifikasi (*misclassification*).

Temuan ini mengonfirmasi bahwa meskipun HPF mampu meningkatkan akurasi rata-rata global dengan membersihkan *noise*, filter ini bersifat destruktif bagi kelas yang mengandalkan fitur akustik frekuensi rendah. Hal ini mengindikasikan perlunya pendekatan *class-aware preprocessing* atau penggunaan *learnable filterbank* agar adaptasi filter dapat disesuaikan dengan karakteristik masing-masing kelas.



Gambar 5. Visualisasi dampak penerapan filter HPF-1000Hz terhadap akurasi per kelas

F. Diskusi dan rekomendasi

Berdasarkan sintesis dari seluruh rangkaian eksperimen, penelitian ini berhasil mengungkap karakteristik spektral krusial yang mendiferensiasikan domain *Environmental Sound Classification* (ESC) dengan pengenalan wicara konvensional. Temuan ini membawa implikasi signifikan terhadap standar pra-pemrosesan sinyal audio, yang dirangkum sebagai berikut:

1. Inefektivitas Standar *Band-Pass* Telefoni: Hasil evaluasi secara tegas menolak efektivitas penerapan filter *Band-Pass* standar telefoni (300-3400 Hz) pada *dataset* suara lingkungan. Penurunan akurasi yang masif membuktikan bahwa sinyal suara lingkungan bersifat *broadband* (pita lebar). Informasi semantik tersebar luas, mulai dari frekuensi tinggi (>4 kHz) untuk kelas biakustik hingga frekuensi rendah untuk kelas mekanis. Oleh karena itu, teknik reduksi *bandwidth* yang agresif sangat tidak direkomendasikan untuk sistem ESC.
2. Rekomendasi Standar Baru: *Gentle High-Pass Filtering*: Penelitian ini mengidentifikasi 'Zona Optimal' pra-pemrosesan menggunakan *High-Pass Filter* dengan *cut-off* moderat (500-1000 Hz) dan orde rendah (FIR-32). Konfigurasi ini terbukti efektif sebagai *selective denoiser* yang membuang *recording rumble* tanpa mengorbankan integritas transien sinyal. Metode ini direkomendasikan sebagai standar baru untuk meningkatkan SNR pada *dataset* ESC-50.
3. Limitasi Filter Statis dan Urgensi Pendekatan Adaptif: Kegagalan filter digital klasik (LPF/BPF) dengan parameter statis menegaskan bahwa kompleksitas spektral suara lingkungan menuntut fleksibilitas yang tidak dapat dicapai oleh filter dengan *cut-off* kaku. Temuan ini sejalan dengan tren riset global yang mulai meninggalkan pemrosesan sinyal deterministik menuju pendekatan adaptif berbasis data.

Studi mutakhir seperti SincNet [30] dan LEAF [31] telah membuktikan bahwa filter yang paling efektif untuk *Deep Learning* adalah filter yang parameternya (*cut-off* dan *bandwidth*) dipelajari secara otomatis selama proses pelatihan (*learnable filters*). Tidak seperti seperti filter tradisional (seperti Mel-Filterbank atau Butterwoth) yang memiliki interval frekuensi tetap, *learnable filters* memungkinkan lapisan *front-end* untuk melakukan adaptasi bentuk gelombang secara dinamis. Secara otomatis, filter ini mengizinkan optimasi pada fungsi *sinc* atau parameter Gabor melalui algoritma *backpropagation*, sehingga model dapat secara mendiri menggeser fokus spektralnya pada area frekuensi yang mengandung fitur paling diskriminatif untuk setiap kelas suara.

Keunggulan utama pendekatan ini terletak pada kemampuannya untuk mengatasi masalah regiditas filter statis yang ditemukan dalam eksperimen ini. Sebagai contoh, alih-alih menggunakan satu nilai *cut-off* tunggal yang bersifat destruktif bagi kelas frekuensi rendah untuk suara ketukan pintu, namun tetap mampu melakukan pembersihan *noise* yang agresif untuk suara mesin. Merujuk pada fakta tersebut, penelitian selanjutnya sangat disarankan untuk beralih mengadopsi arsitektur *end-to-end* dengan *adaptive front-end* guna menciptakan sistem klasifikasi yang lebih tangguh terhadap variabilitas karaktersitik spektral lingkungan.

IV. Simpulan

Penelitian ini telah mengevaluasi secara sistematis dampak penerapan filter digital terhadap kinerja CNN dalam klasifikasi suara lingkungan. Berdasarkan rumusan masalah yang diajukan, penelitian ini menyimpulkan dua temuan utama. Pertama variasi orde filter terbukti berpengaruh signifikan terhadap integritas fitur transien; penggunaan orde tinggi (FIR-128) atau IIR-8) memicu distorsi *ringing* dan *transient smearing* yang mengaburkan *onset* suara, sehingga orde rendah (FIR-32 atau IIR-2) ditemukan sebagai konfigurasi paling optimal untuk mempertahankan resolusi waktu. Kedua, karakteristik respons frekuensi dari tipe filter yang berbeda menunjukkan dampak selektif terhadap kelas suara; penerapan filter standar telefoni (*Band-Pass* 300-3400 Hz) maupun *Low-Pass Filter* terbukti destruktif karena mengeliminasi komponen frekuensi tinggi yang vital bagi suara lingkungan bersifat *broadband*.

Temuan positif dari riset ini menunjukkan bahwa *High-Pass Filter* (HPF) dengan *cut-offf* moderat (500-1000 Hz) mampu meningkatkan performa model melampaui *baseline* dengan cara mengeliminasi derau frekuensi rendah tanpa mendistorsi informasi transien. Selain itu hasil eksperimen juga mengukuhkan superioritas *Log-Mel Spectrogram* dibandingkan MFCC dalam menyediakan representasi spektral spasial yang optimal bagi CNN. Sebagai implikasi praktis, disarankan penggunaan HPF orde rendah sebagai standar pra-pemrosesan baru. Untuk penelitian masa depan, direkomendasikan transisi dari filter statis menuju pendekatan adaptif berbasis data (*learnable filters*) guna mengatasi rigiditas filter konvensional pada kelas suara impulsif.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada institusi Universitas Bali Internasional atas dukungan yang diberikan selama proses penelitian ini. Dukungan berupa fasilitas lab komputasi, akses perangkat lunak pendukung, serta kesempatan untuk melakukan pengembanagn riset telah berperan penting dalam dalam kelancaran penyusunan penelitian ini.

Daftar Pustaka

[1] B. S. Reddy, D. M. Chowdary, R. Srinivas, and M. O. Rahmani, "Classification of Environmental and Urban Sounds Using Deep Learning Techniques," *Proceedings of the International Conference on Data, Electronics and Computing Engineering (ICDCECE)*, 2025.

[2] A. Singh, T. Deacon, And M. D. Plumbley, "Environmental Sound Classification Using Raw-Audio Based Ensemble Framework," In *53rd International Congress And Exposition On Noise Control Engineering, Internoise 2024*, 2024, Vol. 9, Pp. 6417–6425. [Online]. Available: Https://Www.Scopus.Com/Inward/Record.Uri?Eid=2-S2.0-105016121860&Partnerid=40&Md5=9719a7f39ec7d3e16b236c08ef19567e

[3] A. Ashurov, Y. Zhou, L. Shi, Y. Zhao, And H. Liu, "Environmental Sound Classification Based On Transfer-Learning Techniques With Multiple Optimizers," *Electronics*, Vol. 11, Pp. 1–20, 2022.

[4] A. Sivasankar, R. K. Mohith Niranjen, And S. S. Shah, "Sound Recognition System For People With Impaired Hearing," In *Proceedings Of The International Conference On Multi-Agent Systems For Collaborative Intelligence, Icmsci 2025*, 2025, Pp. 846–851.

[5] R. S. Shivanandham, M. Ishwarya Niranjana, D. Madhumitha, M. Sharmila Parveen, J. S. Roshaan, and K. Srisanjana, "A 45 nm CMOS OTA-Based Low-Pass Filter Designed in Cadence Virtuoso," *Proceedings of the International Conference on Communication and Computing Technologies (ICOCT)*, 2025.

[6] M. Mahdavi, "A Low-Latency And Programmable Band-Pass Filter," In *2024 7th International Balkan Conference On Communications And Networking, Balkancom 2024*, 2024, Pp. 107–112.

[7] A. Bansal And N. K. Garg, "Environmental Sound Classification: A Descriptive Review Of The Literature," *Intell. Syst. With Appl.*, Vol. 16, 2022.

[8] J. Galić, B. Marković, Đ. Grozdić, B. Popović, And S. Šajić, "Whispered Speech Recognition Based On Audio Data Augmentation And Inverse Filtering," *Appl. Sci.*, Vol. 14, No. 18, P. 8223, 2024.

[9] J. L. Bautista, Y. K. Lee, And H. S. Shin, "Speech Emotion Recognition Based On Parallel Cnn-Attention Networks With Multi-Fold Data Augmentation," *Electron.*, Vol. 11, No. 23, P. 3935, 2022.

[10] R. Rajan And S. Sivan, "Multi-Channel Cnn-Based Rāga Recognition In Carnatic Music Using Sequential Aggregation Strategy," *Circuits, Syst. Signal Process.*, Vol. 42, No. 7, Pp. 4072–4095, 2023.

[11] K. J. Piczak, "Esc: Dataset For Environmental Sound Classification," In *Proceedings Of The 23rd Acm International Conference On Multimedia*, 2015, Pp. 1015–1018.

[12] U. Dubey And R. Barskar, "Convolutional Neural Network In Deep Learning For Object Tracking: A Review," In *Lecture Notes In Networks And Systems*, 2024, Vol. 832, Pp. 343–353.
[13] J. Wang, X. Zeng, S. Duan, Q. Zhou, And H. Peng, "Image Target Recognition Based On Improved Convolutional Neural Network," *Math. Probl. Eng.*, Vol. 2022, 2022.
[14] C. Bin, "A Sound Classification Method Of Traditional Chinese Musical Instruments Based On Mel Spectrogram And Convolutional Neural Networks," In *Ieee Joint International Information Technology And Artificial Intelligence Conference (Itaic)*, 2025, Pp. 369–373.
[15] Y. Chen, F. Xiang, M. L. Hernandez, D. Carpenter, A. Bozkurt, And E. Lobaton, "Robust Multimodal Cough Detection With Optimized Out-Of-Distribution Detection For Wearables," *IEEE J. Biomed. Heal. Informatics*, 2025.
[16] A. M. Lorenzo, R. Barien, N. D. Favila, D. Basa, J. M. Ventura, and S. Catolos, "Trees Have Ears: An Acoustic Surveillance and TinyML-Based System for Detecting Illegal Logging," *Proceedings of the International Conference on Automation, Artificial Intelligence and Electrical Engineering Innovation (ICAAEEI)*, 2024.
[17] A. Pant And A. Kumar, "Hanning Fir Window Filtering Analysis For Eeg Signals," *Biomed. Anal.*, Vol. 1, No. 2, Pp. 111–123, 2024.
[18] E. O. Oyebode And A. Olatunji, "Towards Investigating The Properties Of Some Finite Impulse Response Filter In Signal Processing," *Ajayi Crowther J. Pure Appl. Sci.*, Vol. 2, Pp. 51–61, 2023,.
[19] W. Mu, B. Yin, X. Huang, J. Xu, And Z. Du, "Environmental Sound Classification Using Temporal-Frequency Attention Based Convolutional Neural Network," *Sci. Rep.*, Vol. 11, No. 1, P. 21552, 2021.
[20] Z. Pautzke, D. Kubanek, And T. J. Freeborn, "(3 + A)-Order Transfer Functions For Approximating Butterworth-Type Flat Passband Characteristics," In *Conference Proceedings - Ieee Southeastcon*, 2023, Vol. 2023, Pp. 815–820.
[21] L. R. Rabiner And R. W. Schafer, *Theory And Applications Of Digital Speech Processing*. Upper Saddle River, Nj: Pearson, 2010.
[22] L. C. Paulick, H. Relaño-Iborra, And T. Dau, "The Computational Auditory Signal Processing And Perception Model: A Revised Version," *J. Acoust. Soc. Am.*, Vol. 157, No. 5, Pp. 3232–3244, 2025.
[23] D. P. Goel, K. Mahajan, N. D. Nguyen, N. Srinivasan, And C. P. Lim, "Towards An Efficient Backbone For Preserving Features In Speech Emotion Recognition: Deep-Shallow Convolution With Recurrent Neural Network," *Neural Comput. Appl.*, Vol. 35, No. 3, Pp. 2457–2469, 2023.
[24] R. Huang And Y. Xie, "A CNN-Based Multi-Scale Pooling Strategy For Acoustic Scene Classification," *IEICE Trans. Inf. Syst.*, Vol. E107.D, No. 1, Pp. 153–156, 2024.
[25] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, And M. D. Plumbley, "Panns: Large-Scale Pretrained Audio Neural Networks For Audio Pattern Recognition," *IEEE/ACM Trans. Audio, Speech And Lang. Proc.*, Vol. 28, Pp. 2880–2894, 2020.
[26] Y. Dogan, "A New Global Pooling Method For Deep Neural Networks: Global Average Of Top-K Max- Pooling," *Trait. Du Signal*, Vol. 40, No. 2, Pp. 577–587, 2023.
[27] F. Kjeldsberg, Z. H. Munim, M. Bustgaard, S. Bhagat, E. Lindroos, And P. Haavardtun, "Sensitivity Of Predictive Performance Assessment Accuracy In Varying K-Fold Cross Validation," In *Lecture Notes In Networks And Systems*, 2025, Vol. 1274 Lnns, Pp. 71–82.
[28] H. L. Vu, K. T. W. Ng, A. Richter, And C. An, "Analysis Of Input Set Characteristics And Variances On K-Fold Cross Validation For A Recurrent Neural Network Model On Waste Disposal Rate Estimation," *J. Environ. Manage.*, Vol. 311, 2022.
[29] A. Pascale, C. Guarnaccia, And M. Coelho, "Analysis Of Single Vehicle Noise Emissions In The Frequency Domain For Two Different Motorizations.," *J. Environ. Manage.*, Vol. 370, P. 122905, 2024.
[30] B. Saritha, N. Shome, R. H. Laskar, and M. Choudhury,"Enhancement in Speaker Recognition Using SincNet Through Optimal Window and Frame Shift," *Proceedings of the International Conference on Intelligent Technologies (CONIT)*, 2022.
[31] N. Zeghidour, O. Teboul, F. De Chaumont Quitry, And M. Tagliasacchi, "Leaf: A Learnable Frontend For Audio Classification," 2021. [Online]. Available: Https://Www.Scopus.Com/Inward/Record.Uri?Eid=2-S2.0-85131246197&Partnerid=40&Md5=43f95ab49a38e4320d59bf529a2fb459