MAKALAH

DATA MINING INTELECTA 2025

# Ensemble Citra Berbasis Spectral Contrast untuk Klasifikasi Spesies Mamalia Laut


Oleh:

| | |
|---|---|
| Daniel Sandi Bratanata | 5026231216 |
| Muhammad Abid Baihaqi Al Faridzi | 5025241133 |
| Hartmann Kanisius Galla' Massang | 5025241160 |
| Danish Rafie Ekaputra | 5052231024 |


**ak mw final**


**Institut Teknologi Sepuluh Nopember**

# ABSTRAK

Mamalia laut memainkan peran penting dalam ekosistem laut dan menunjukkan keragaman spesies yang tinggi. Namun, identifikasi keberadaan mereka masih menantang karena setiap spesies memiliki karakteristik akustik berbeda yang seringkali berada pada rentang frekuensi yang sulit dideteksi oleh pendengaran manusia. Studi ini mengatasi tantangan tersebut dengan fokus pada klasifikasi vokalisasi 32 spesies mamalia laut menggunakan pendekatan berbasis citra. Kami mengusulkan penggunaan fitur *spectral contrast* untuk merepresentasikan perbedaan intensitas antara puncak dan lembah spektral, sehingga menonjolkan pola frekuensi yang khas bagi tiap spesies. Representasi *spectral contrast* kemudian dikonversi menjadi citra 2D dan dianalisis menggunakan tiga arsitektur model citra berbeda; WhaleNet, ResNet-50, dan VGG-16. Hasil eksperimen menunjukkan bahwa WhaleNet dengan fitur *spectral contrast* mencapai F1-Macro sebesar **0.9635**. Pengembangan ensemble dari ketiga model percobaan untuk menangkap variasi pola spektral dan meningkatkan performa menjadi **0.9857**. Kontribusi utama penelitian ini adalah demonstrasi bahwa ekstraksi fitur *spectral contrast* dan pemodelan berbasis citra dapat meningkatkan klasifikasi vokalisasi pada 32 spesies mamalia laut meskipun data yang tersedia terbatas.



**Kata Kunci:** klasifikasi akustik mamalia laut, spectral contrast, ResNet, vokalisasi laut.

# DAFTAR ISI

# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Mamalia laut memiliki peran penting dalam menjaga keseimbangan ekosistem perairan dan menjadi indikator kualitas lingkungan laut. Namun, pemantauan spesies mamalia laut masih terkendala karena identifikasi melalui vokalisasi alami mereka tidak mudah dilakukan. Setiap spesies memiliki karakteristik suara yang unik dan sering berada pada rentang frekuensi yang sulit diidentifikasi manusia. Selain itu, dinamika lingkungan laut, kebisingan, serta tumpang tindih frekuensi antarspesies semakin mempersulit proses klasifikasi, sehingga diperlukan metode otomatis yang lebih akurat untuk mendukung upaya konservasi.

Penelitian terkait klasifikasi akustik mamalia laut telah berkembang dengan berbagai pendekatan, seperti fitur mel-spectrogram, MFCC, spectral contrast, dan transformasi wavelet. Berbagai model seperti CNN, ResNet, LSTM, dan attention-based networks pun telah digunakan. Meskipun demikian, sebagian besar studi hanya memanfaatkan satu jenis representasi fitur sehingga belum mampu menangkap seluruh karakteristik sinyal. Keterbatasan dataset, baik dari sisi jumlah, ketidakseimbangan kelas, maupun variasi kualitas rekaman, juga membuat performa model sering kurang optimal.

Untuk mengatasi masalah tersebut, penelitian ini mengusulkan model Multi-Representative dengan Late-Fusion Ensemble yang menggabungkan empat representasi akustik: spectral contrast, mel-spectrogram,wavelet scattering transform, dan scattering transform. Masing-masing fitur diekstraksi menggunakan ResNet, kemudian digabungkan melalui MLP sebagai pengambil keputusan akhir. Pendekatan late fusion dipilih karena mampu memanfaatkan kekuatan setiap representasi secara independen, sekaligus meningkatkan robustness pada kondisi data yang terbatas.

## 1.2 Tujuan

Tujuan penelitian ini adalah sebagai berikut.

1. Mengembangkan arsitektur multi-representative yang lebih handal dibanding penelitian sebelumnya
2. Menganalisis performa representasi fitur baru pada model,dalam menangkap pola panggilan mamalia laut.
3. Mengukur performa model (Akurasi, F1-Score) dan membandingkannya dengan model baseline (model fitur tunggal) untuk membuktikan keunggulan pendekatan multi-representative.

## 1.3 Manfaat

Manfaat penelitian ini adalah sebagai berikut.

1. Memberikan bukti empiris mengenai performa arsitektur *Late-Fusion* yang menggabungkan tiga fitur (*spectral contrast, mel-spectrogram, scattering transform*) dalam menangani kompleksitas suara mamalia laut, terutama pada dataset yang berbeda dari studi aslinya.
2. Memberikan wawasan mendalam mengenai karakteristik masing-masing fitur akustik dan bagaimana masing-masing representasi fitur mengisi pola yang tidak tertangkap oleh model
3. Memberi kesempatan bagi penelitian bioakustik dalam meningkatkan performa sistem otomasi deteksi suara.

## 1.4 Batasan Masalah

1. Penelitian hanya menggunakan tiga jenis fitur akustik: spectral contrast, mel-spectrogram, dan scattering transform.
2. Dataset yang digunakan terbatas pada rekaman suara mamalia laut tertentu dan tidak mencakup semua spesies.

3. Dataset yang digunakan memiliki jumlah yang sangat kecil pada species tertentu.

# BAB II
# PENELITIAN TERKAIT

## 2.1 Implementasi Dataset Terkait

Penelitian terkait klasifikasi suara mamalia laut telah mendapat perhatian signifikan seiring berkembangnya metode Deep Learning. Hagiwara et al. mengembangkan benchmark dataset BEANS dan model AVES yang berbasis self-supervision, mencapai akurasi sekitar 87,9% hingga 92,8% pada subset data pilihan (best of subset). Pendekatan lain oleh Murphy et al. menggunakan arsitektur ResNet pada subset serupa dan menghasilkan akurasi sebesar 83%. Lebih lanjut, Carbone dan Licciardi memperkenalkan WhaleNet, sebuah arsitektur deep ensemble yang menggabungkan Wavelet Scattering Transform (WST) dan Mel Spectrogram. Model ini memanfaatkan sifat invariansi WST terhadap deformasi sinyal dan mencapai akurasi tertinggi sebesar 97,61% pada dataset penuh WMMD, membuktikan bahwa penggabungan representasi fitur yang berbeda dapat meningkatkan diskriminabilitas model secara signifikan .

Namun, sebagian besar penelitian tersebut cenderung memiliki membatasi dalam cakupan data, seperti penggunaan subset "Best of" atau jumlah kelas yang terbatas sedikit. Dan beberapa penelitian berfokus pada optimalisasi satu jenis arsitektur atau fitur standar tanpa mengeksplorasi potensi kombinasi model heterogen (heterogeneous ensemble) yang lebih dalam. Untuk itu, penelitian ini berkontribusi dalam pengembangan sistem klasifikasi berbasis ensemble pada dataset Data Mining dari Intelecta yang mengintegrasikan tiga arsitektur berbeda: WhaleNet, VGG16, dan ResNet50, dengan penambahan fitur Spectral Contrast untuk menangkap karakteristik tekstur spektral. Dan menggunakan strategi fusi dengan Weighted Ensemble, di mana prediksi probabilitas dari ketiga model digabungkan menggunakan bobot adaptif untuk menyeimbangkan kontribusi masing-masing model. Pendekatan ini bertujuan untuk meningkatkan generalisasi dan stabilitas prediksi, khususnya pada metrik F1-Macro untuk kelas spesies minoritas.

## 2.2 Implementasi Metode Terkait

### 2.2.1 Ekstraksi Fitur Wavelet Scattering Transform

Wavelet Scattering Transform (WST) orde 1 merupakan *layer* pertama dari *transformasi scattering* yang mengekstrak fitur stabil dan invarian dari sinyal waktu [1]. WST orde 1 menggunakan konvolusi wavelet diikuti modulus absolut dan averaging low-pass filter. Secara matematis koefisien scattering orde 1 didefinisikan berikut:

$$S_1 x(t, \lambda_1) = |x * \psi_{\lambda_1}| * \phi(t)$$

Dimana:

- $x$ atau $x(t)$: Sinyal asli (input).
- $\psi_{\lambda_1}$: Wavelet pada skala $\lambda_1$ sebagai band-pass filter.
- $|\cdot|$: Modulus absolut untuk untuk stabilitas terhadap noise dan deformasi.
- $*$: Konvolusi
- $\phi(t)$: Low-pass filter (biasanya fungsi Gaussian) yang memberikan sifat invarian terhadap translasi waktu lokal (deformasi kecil)

Penelitian oleh Carbone dan Licciardi tentang WST menunjukkan bahwa orde rendah dari WST (orde 1) sudah cukup untuk merepresentasikan sekitar 98% energi sinyal sehingga efisien secara komputasi [2]. Melalui implementasi pada arsitektur deep residual, pendekatan WST ini berhasil mencapai akurasi pengujian sebesar 94% dan weighted F1-score 0,94. Hasil ini terbukti mengungguli benchmark state-of-the-art sebelumnya sebesar 6% serta secara kualitatif menunjukkan stabilitas konvergensi loss pelatihan yang lebih konsisten dibandingkan metode *spectrogram* konvensional.

### 2.2.2 Ekstraksi Fitur Mel-Spectogram

Mel Spectrogram merupakan representasi waktu-frekuensi yang dirancang untuk meniru persepsi pendengaran manusia (non-linear) dengan memetakan spektrum energi ke dalam skala Mel. Metode ini menghitung magnitudo dari

Short-Time Fourier Transform (STFT) yang kemudian diproyeksikan menggunakan serangkaian filter segitiga (Mel filter bank) yang saling tumpang tindih. Secara matematis, Mel Spectrogram didefinisikan sebagai penjumlahan energi pada setiap pita filter berikut:

$$M(t,m) = \sum_{k=0}^{N-1} |X(t,\omega_k)|^2 H_m(\omega_k)$$

Dimana:

- $M(t,m)$ : Koefisien Mel Spectrogram pada waktu $t$ dan pita frekuensi Mel $m$
- $|X(t,\omega_k)|^2$ : Spektrogram daya (kuadrat magnitudo) dari STFT sinyal input
- $H_m(\omega_k)$ : Nilai dari *filter bank* segitiga Mel ke-$m$ pada frekuensi $\omega_k$
- $N$ : Jumlah total frequency bins dalam STFT

Studi yang sama oleh Carbone dan Licciardi (2024) pada Watkins Marine Mammal Sound Database (WMMD) tentang Mel Spectrogram bahwa penggunaan Mel Spectrogram dengan 64 bin frekuensi terbukti memberikan performa klasifikasi tertinggi dibandingkan metode lainnya [2]. Implementasi ini berhasil mencapai akurasi pengujian sebesar 96% dan weighted F1-score 0,96, yang secara signifikan mengungguli arsitektur klasifikasi benchmark yang ada sebesar 8% serta mengurangi separuh jumlah sampel yang salah prediksi.

### 2.2.3 Ekstraksi Fitur Short-Time Fourier Transform (STFT)

STFT (Short-Time Fourier Transform) merupakan metode dasar dalam pemrosesan sinyal yang memetakan sinyal satu dimensi ke dalam domain waktu-frekuensi untuk menganalisis distribusi energi lokal yang berubah seiring waktu [4]. STFT membagi sinyal menjadi segmen-segmen pendek menggunakan fungsi window (seperti Hann atau Gaussian) yang digeser sepanjang sumbu waktu, kemudian menghitung transformasi Fourier untuk setiap segmen tersebut. Secara matematis, transformasi STFT didefinisikan melalui persamaan integral berikut:

$$STFT\{x\}(t,\omega) = \int_{-\infty}^{\infty} x(\tau) h(\tau - t) e^{-i\omega\tau} d\tau$$

Dimana:

- $x(\tau)$: Sinyal asli (input)
- $h(\tau - t)$: Fungsi window (jendela) yang digeser sebesar $t$
- $\omega$: Variabel frekuensi angular

- $e^{-i\omega\tau}$ : Basis kompleks Transformasi Fourier

Studi terkait mengenai STFT yang dilakukan dalam penelitian Bach dkk. (2023) pada dataset bioakustik National Oceanic and Atmospheric Administration (NOAA). Dalam penelitian tersebut, STFT menggunakan fungsi window Hanning dengan ukuran FFT 256 dan overlap 75% untuk menghasilkan *spectrogram* RGB berukuran 224x224 sebagai input model klasifikasi [5]. Evaluasi eksperimental menunjukkan bahwa penggunaan fitur STFT dengan arsitektur CNN menghasilkan akurasi klasifikasi sebesar 80,5% untuk empat kelas mamalia laut.

### 2.2.4 Ekstraksi Fitur Spectral Contrast

Spectral Contrast merupakan metode ekstraksi fitur yang dirancang untuk merepresentasikan karakteristik tekstur spektral dari sinyal audio. Berbeda dengan MFCC yang merupakan representasi spektral rata-rata untuk menangkap timbre secara umum, Spectral Contrast berfokus pada perbedaan level energi antara puncak (peaks) dan lembah (valleys) dalam spektrum frekuensi [6]. Metode ini membagi spektrum sinyal ke dalam beberapa rentang frekuensi (biasanya skala oktaf) dan menghitung kontras energi di setiap rentang tersebut. Secara matematis, nilai rata-rata Spectral Contrast ($SC_k$) untuk rentang frekuensi ke-$k$ didefinisikan sebagai persamaan berikut:

$$SC_k = \underbrace{\frac{1}{T}\sum_{t=1}^{T}\log(E_{peak,k}^{(t)})}_{\text{Mean Peak}} - \underbrace{\frac{1}{T}\sum_{t=1}^{T}\log(E_{valley,k}^{(t)})}_{\text{Mean Valley}}$$

Dimana:

- $SC_k$: Nilai rata-rata Spectral Contrast pada rentang frekuensi ke-$k$
- $T$: Jumlah total time frames dalam sinyal audio.
- $E_{peak,k}^{(t)}$: Magnitudo energi peaks pada rentang $k$ saat waktu $t$
- $E_{valley,k}^{(t)}$: Magnitudo energi valleys pada rentang $k$ saat waktu $t$

Efektivitas fitur ini dalam *case* pemantauan akustik bawah air ditunjukkan oleh penelitian Kong et al. (2024). Dalam studi tersebut, *Spectral Contrast* dengan *fusion* fitur MFCC dan pola ritme untuk mengklasifikasikan dataset gabungan yang terdiri dari 37 kelas (32 Mammal + 5 Vessel) dengan total 558 sampel audio [7]. Hasilnya menunjukkan bahwa Spectral Contrast berperan krusial dalam mendiskriminasi pola

distribusi energi yang dinamis pada sinyal Mammal dibandingkan pola statis pada sinyal Vessel, yang memungkinkan model SVM mencapai akurasi hingga 99%.

### 2.2.5 Ekstraksi Fitur Concatenation (Penggabungan Fitur)

Konkatenasi fitur (*Feature Concatenation*) merupakan metode *feature-level fusion* yang mengintegrasikan informasi dari berbagai domain representasi ke dalam satu vektor fitur gabungan yang kontinu. Pendekatan ini bertujuan untuk memperkaya ruang fitur dengan menggabungkan karakteristik intrinsik dari masing-masing ekstraktor tanpa mengubah nilai aslinya, sehingga memungkinkan model pembelajaran mesin untuk mengeksploitasi korelasi antar-fitur secara simultan. Secara matematis, jika diberikan dua vektor fitur input $\mathbf{x} \in \mathbb{R}^m$ dan $\mathbf{y} \in \mathbb{R}^n$, operasi konkatenasi didefinisikan sebagai pembentukan vektor baru $\mathbf{z}$ sebagai berikut:

$$\mathbf{z} = \text{Concat}(\mathbf{x}, \mathbf{y}) = [x_1, \ldots, x_m, y_1, \ldots, y_n] \in \mathbb{R}^{m+n}$$

Dimana:

- $\mathbf{z}$ : Vektor fitur gabungan hasil fusi dengan dimensi total $m + n$.
- $\mathbf{x}$ : Vektor fitur pertama
- $\mathbf{y}$ : Vektor fitur kedua
- $m, n$ : Dimensi dari masing-masing vektor fitur input

Studi yang dilakukan oleh Meng et al. (2025) pada dataset Watkins Marine Mammal Sound Database (WMMD). Penelitian tersebut menggunakan arsitektur Dual-Branch ResNet yang memproses fitur CQT dan STFT secara paralel sebelum digabungkan melalui konkatenasi. Model ini terbukti efektif dengan mencapai akurasi 94,76% dan F1-score 94,77%. Hasil ini menunjukkan *outperform* dibandingkan penggunaan fitur tunggal, di mana model berbasis STFT murni mencapai akurasi 94,37% dan CQT murni hanya 61,43% [8].

### 2.2.6 Ensemble Model

Ensemble Model dengan pendekatan Soft Voting adalah metode fusi yang menentukan prediksi akhir berdasarkan rata-rata probabilitas yang dihasilkan oleh setiap model dasar. Berbeda dengan Hard Voting yang hanya menghitung jumlah label terbanyak, Soft Voting mempertimbangkan tingkat keyakinan (confidence) dari setiap model, sehingga menghasilkan keputusan klasifikasi yang lebih akurat dan

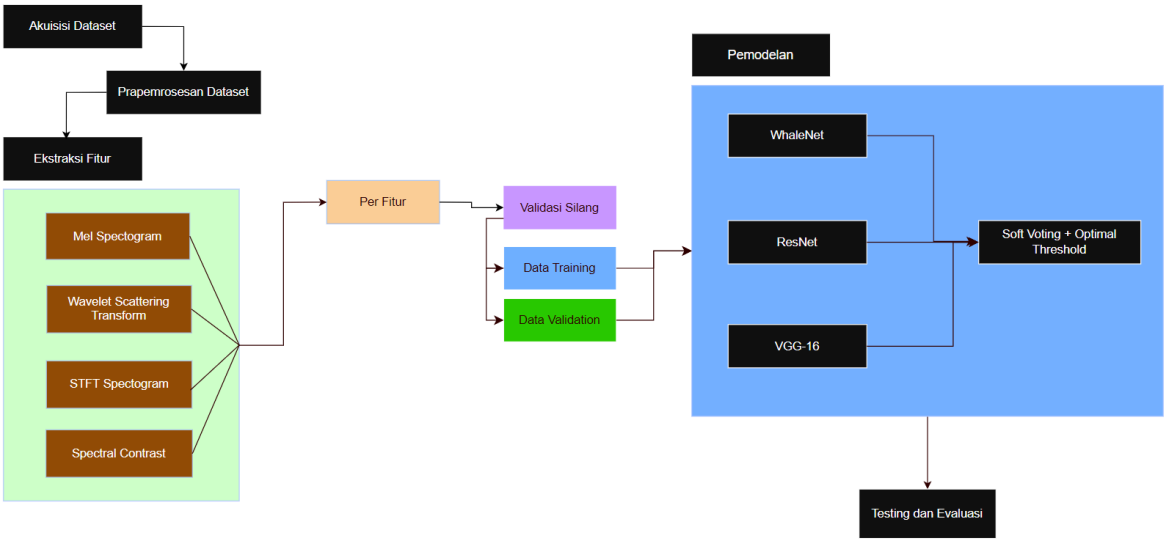
stabil. Secara matematis, kelas akhir ditentukan dengan memilih kelas yang memiliki nilai rata-rata probabilitas tertinggi dari seluruh model yang digabungkan.

Penerapan metode ini pada dataset Watkins Marine Mammal Sound Database (WMMD) telah dilakukan oleh Carbone dan Licciardi (2024) dalam arsitektur WhaleNet. Penelitian tersebut menerapkan prinsip fusi probabilitas dengan menggabungkan output prediksi dari tiga cabang jaringan berbeda (WST dan Mel Spectrogram). Hasil eksperimen membuktikan bahwa strategi penggabungan probabilitas ini mampu meningkatkan akurasi klasifikasi secara signifikan hingga mencapai 97,60%, mengungguli kinerja model tunggal.

# BAB III
# METODE PENELITIAN

Dalam penelitian ini dilakukan dalam beberapa tahap, meliputi prapemrosesan dataset, ekstraksi fitur spektral, strategi konkatenasi, serta pemodelan berbasis tabular dan citra. Alur penelitian dilanjutkan dengan penerapan teknik ensemble, dan diakhiri dengan testing serta evaluasi model. Diagram di bawah menunjukkan metode dan alur analisis secara garis besar.

**Gambar 1**. Alur Penelitian




## 3.1 Eksplorasi Data Analisis (EDA)

### 3.1.1 Dataset

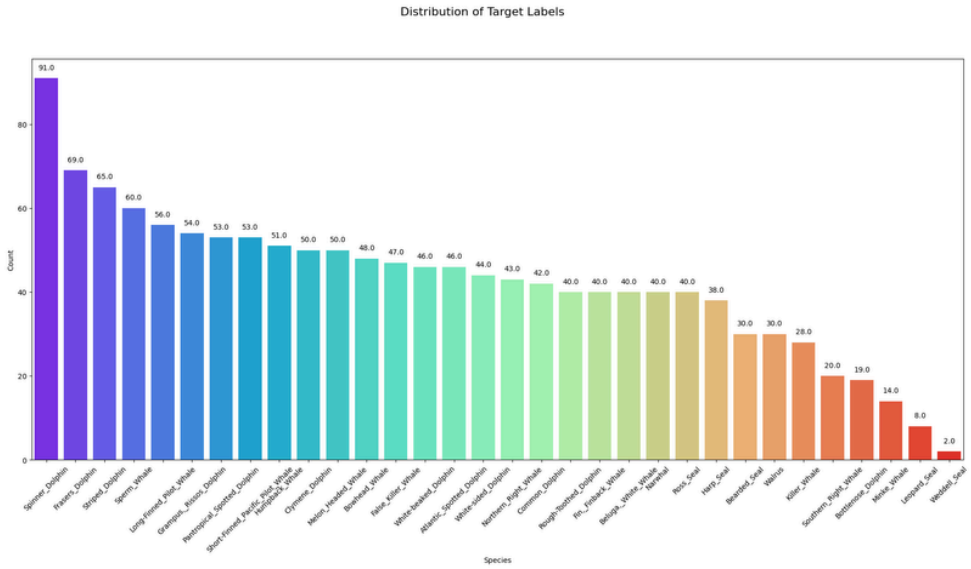
Dataset ini terdiri dari 1.357 data train dan 340 data test dalam format audio (.wav) yang mencakup 32 kelas spesies mamalia laut sebagai target klasifikasi [9]. Untuk proses pelatihan dan validasi, diterapkan metode Stratified K-Fold Cross Validation dengan nilai $k = 5$. Pendekatan stratified ini dipilih untuk memastikan proporsi sebaran kelas spesies tetap seimbang di setiap lipatan (*fold*), sehingga

meningkatkan reliabilitas evaluasi dan mencegah *overfitting* yang rentan terjadi pada pembagian data statis (train-test split).

### 3.1.2 Distribusi Target

Mengetahui distribusi data target merupakan langkah penting karena memberikan insight tentang class balance. Karena dengan mengetahui class balance akan menentukan strategi pemodelan bagaimana melakukan strategi splitting untuk validasi internal.

**Gambar 2**. Distribusi Target Label

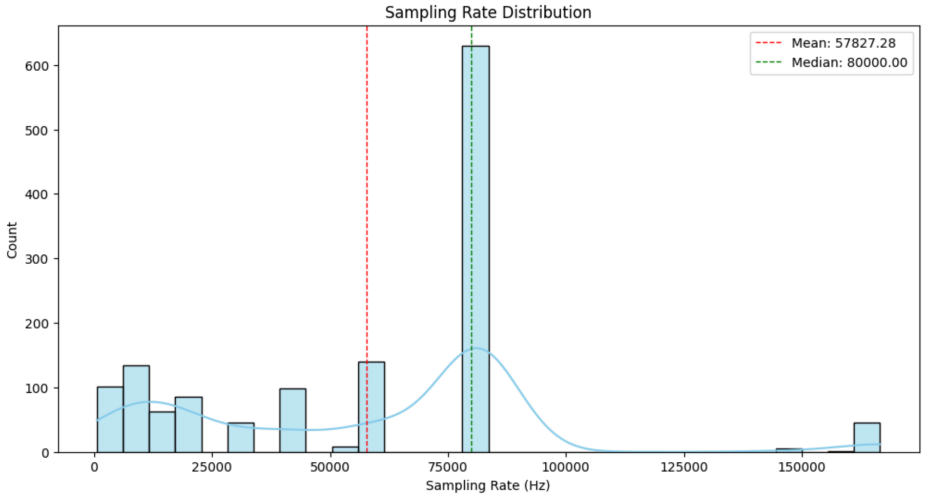



Dapat dilihat bahwa dalam **Gambar 2.** Menunjukkan bahwa distribusi dari data target memiliki persebaran yang *imbalance*. Untuk mengatasi hal tersebut, tim peneliti menggunakan Stratified K-Fold untuk menjaga proporsi representatif di setiap partisi data, serta mencegah bias model yang cenderung memprioritaskan kelas mayoritas dan mengabaikan kelas minoritas.

### 3.1.2 Distribusi Sampling Rate

Visualisasi terhadap distribusi sampling rate penting untuk melihat konsistensi representasi fitur dalam domain time-frequency. Variasi frekuensi yang beragam dapat menyebabkan ketidaksesuaian dimensi pada input model serta distorsi informasi akibat perbedaan frekuensi Nyquist antar sampel audio.

**Gambar 3**. Distribusi Sampling Rate




## 3.2 Preprocessing

Pemrosesan data suara dilakukan dengan menganalisis karakteristik noise yang terdapat pada rekaman, khususnya noise riak air yang mendominasi sebagian besar sampel. Noise tersebut memiliki pola spektral yang terus berubah (non-stationary) sehingga berpotensi menutupi fitur akustik utama. Denoising dapat membuat model lebih mudah menangkap fitur utama pada suara mamalia pau

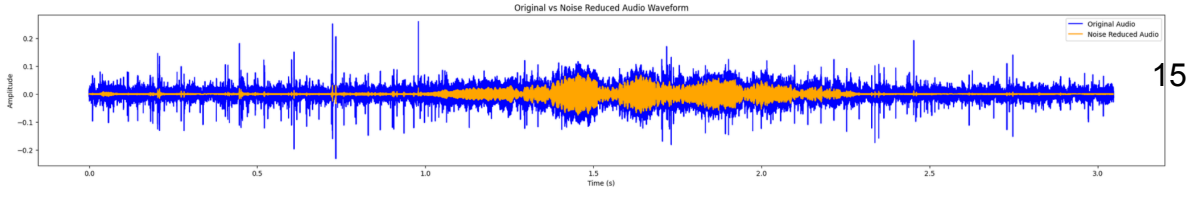
### 3.2.2 Standardisasi Sampling Rate

Dapat dilihat pada **Gambar 3**, distribusi sampling rate menunjukkan heterogenitas yang signifikan dengan sebaran nilai yang sangat bervariasi, mulai dari nilai rendah hingga dominasi di area 80.000 Hz. Untuk mengatasi ini tim peneliti menerapkan teknik resampling untuk menyamakan seluruh data menjadi satu nilai tetap, yaitu 46.700 Hz. Pendekatan ini mengadopsi strategi yang diterapkan dalam penelitian WhaleNet [2].

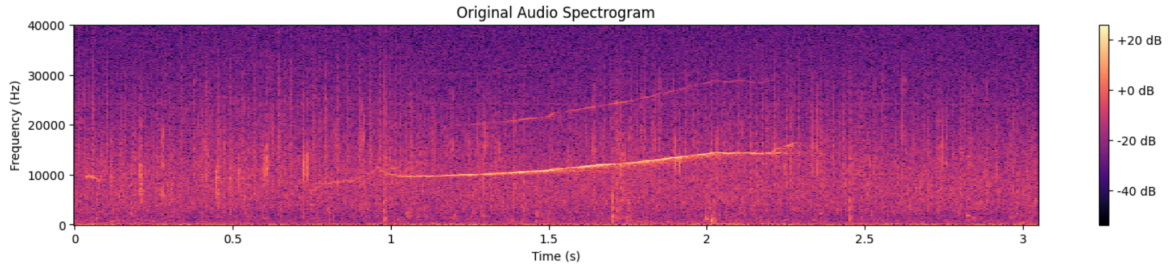
### 3.2.2 Signal Enhancement

Lingkungan bawah air memiliki tingkat noise yang tinggi dan dinamis (non-stationary) dapat dilihat pada **Gambar 4**. Untuk mengatasi ini, tim peneliti menerapkan teknik Non-Stationary Spectral Gating dimana dengan mengestimasi profil noise dari statistik latar belakang lingkungan laut, kemudian menetapkan ambang batas (threshold) adaptif untuk menekan frekuensi yang dikategorikan sebagai gangguan.

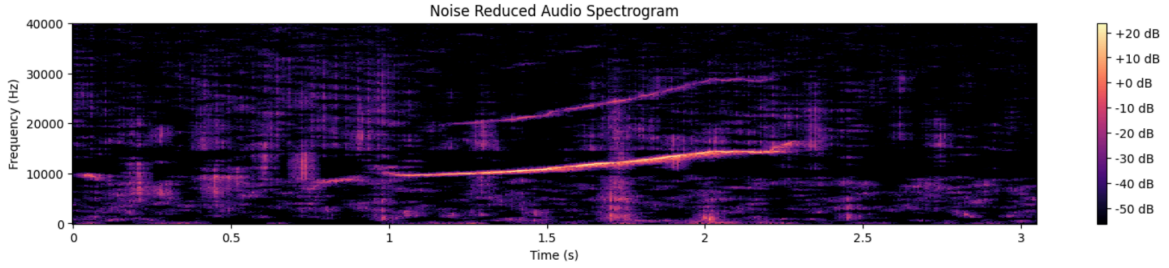
**Gambar 4**. Original vs Noise Reduced Audio



**Gambar 5**. Original Audio Spectogram



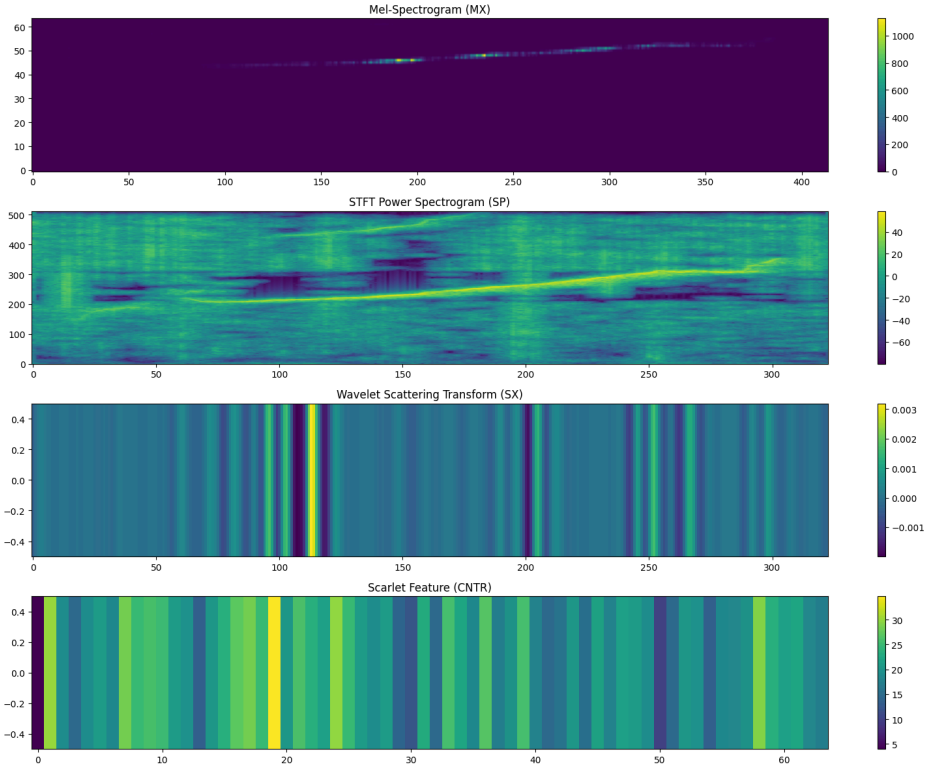

**Gambar 6**. Noise Reduced Audio Spectogram




Pada **Gambar 5** sebelum dilakukannya signal enhancement, terlihat bahwa terdapat broadband noise (gangguan spektrum lebar) yang ditandai dengan latar belakang ungu terang yang mengaburkan detail sinyal. Setelah dilakukannya signal enhancement pada **Gambar 6** spektrum menjadi jauh lebih gelap (bersih). Hal ini mengindikasikan bahwa gangguan frekuensi rendah maupun tinggi berhasil dihilangkan.

## 3.3 Ekstraksi Fitur Citra

**Gambar 7**. Ekstraksi Fitur *Each Method*




### 3.3.1 STFT Spectogram

Pada ekstraksi fitur menggunakan STFT, sinyal terstandarisasi $\hat{x}$ kemudian ditransformasikan ke dalam domain frekuensi. Parameter transformasi dikonfigurasi sebagai berikut:

- Panjang FFT ($N_{FFT}$): 1024 sampel, untuk memberikan resolusi frekuensi
- Hop Length: 256 sampel, yang menentukan resolusi temporal spektrogram
- Power: 2.0, untuk menghitung spectrogram

Hasil transformasi STFT berupa besaran daya $|S|^2$ dikonversi ke skala log (desibel) untuk menjadikan rentang dinamis. Output akhir dari proses ini adalah tensor spectrogram 2D yang merepresentasikan distribusi frekuensi terhadap waktu sebagai input model.

### 3.3.2 Mel-Spectogram

Sama seperti STFT, pada ekstraksi fitur Mel-spectogram sinyal ($\hat{x}$) diproses untuk menangkap karakteristik spektral. Kemudian parameter transformasi dikonfigurasi sebagai berikut:

- Jumlah Mel Bins (n_mels): 64 bands, untuk mereduksi dimensi frekuensi tinggi
- Normalized: True, untuk menstandarisasi luas area filter

Hasilnya menghasilkan representasi time-frequent dalam skala Mel (Mel-frequency scale). Output akhir dari tahapan ini adalah tensor 2D kemudian diperluas menjadi 3D dengan penambahan channel dimension sebagai input model.

### 3.3.3 Wavelet Scattering Transform (WST) Orde 1

Ekstraksi fitur menggunakan WST diterapkan untuk menghasilkan representasi sinyal yang invarian terhadap translasi dan stabil. Sinyal $\hat{x}$ diproses melalui Scattering1D dengan konvolusi wavelet dan modulus non-linear dengan parameter transformasi dikonfigurasi sebagai berikut:

- Skala Invariansi $J$: 8, menentukan ukuran maksimum window
- $Q$: 14, menentukan resolusi frekuensi
- Panjang Sinyal $T$: dimensi temporal dari sinyal input $X_{std}$

Hasil transformasi dari setiap *batch* kemudian *concatenated* menjadi satu tensor. Output ini adalah koefisien scattering orde pertama.

### 3.3.4 Spectral Contrast

Dengan menghitung Short-Time Fourier Transform (STFT) dari sinyal input $\hat{x}$. Parameter ekstraksi sebagai berikut:

- Panjang FFT ($N_{FFT}$): 1024 sampel
- Hop Length: 256 sampel
- $n_{bands}$: 4, membagi spektrum menjadi beberapa sub

Output dari Spectral Contrast (dimensi frekuensi rendah sesuai jumlah band) diinterpolasi menggunakan teknik Zooming Orde-1 (Linear Interpolation) agar memiliki dimensi frekuensi yang sama dengan Mel-Spectrogram, yaitu 64 bin frekuensi $n_{mels}$, serta panjang waktu yang sinkron $T_{frames}$. Hal ini untuk memastikan dimensi spasial saat fitur digabungkan model. Lalu menghasilkan output akhir

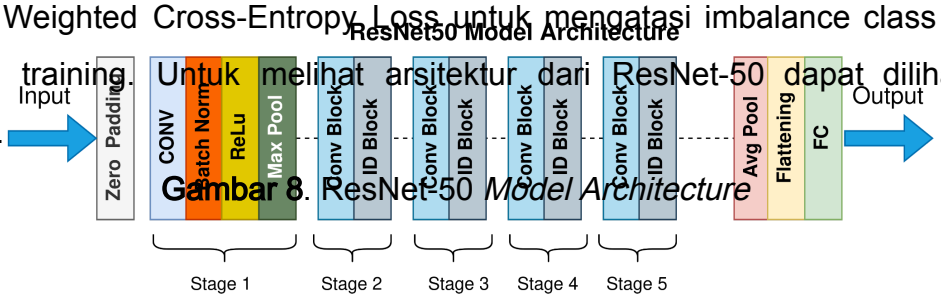
dikonversi menjadi tensor 4D dengan penambahan dimensi channel dan batch untuk model.

### 3.4 Modeling

Proses modeling dilakukan menggunakan fungsi *loss Weighted Cross-Entropy Loss* untuk mengatasi *imbalance class*. Kemudian untuk melakukan optimalisasi dilakukan two-phase training untuk memastikan stabilitas konvergensi parameter. Fase pertama adalah head warm up selama 5 epoch, di mana backbone ekstraktor fitur dilakukan freeze dan hanya *classification head* yang dilatih dengan learning rate $1 \times 10^{-3}$ menggunakan optimizer AdamW, bertujuan untuk menginisialisasi bobot klasifikasi tanpa merusak fitur *pre-trained*. Fase kedua adalah fine-tuning penuh selama 25 epoch, di mana seluruh layer dilakukan *unfreeze* dan *training* kembali dengan skema learning rate diferensial $1 \times 10^{-5}$ untuk backbone dan $1 \times 10^{-4}$ untuk head dengan OneCycleLR untuk adaptasi learning rate yang dinamis sepanjang iterasi.

#### 3.4.1 ResNet-50

Arsitektur ResNet-50 [10] berbasis Transfer Learning yang memanfaatkan residual connection untuk ekstraksi fitur dengan modifikasi pada *fully connected layer* untuk fitur ke 32 target. Seperti yang telah dijelaskan pada bagian **3.4** dilakukan Weighted Cross-Entropy Loss untuk mengatasi imbalance class dengan two-phase training. Untuk melihat arsitektur dari ResNet-50 dapat dilihat pada **Gambar 8**.



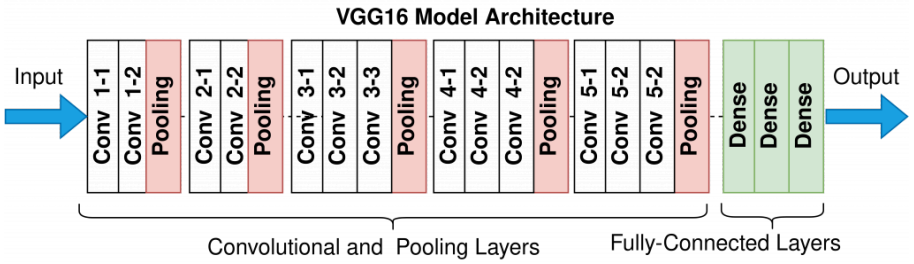

**Gambar 8**. ResNet-50 *Model Architecture*

#### 3.4.2 VGG16

Selain ResNet, penelitian ini juga menggunakan arsitektur VGG16 [11] dengan penggunaan filter konvolusi berukuran kecil $3 \times 3$ dengan depth yang signifikan untuk mengekstraksi fitur spasial dari citra spektrogram. Model diinisialisasi menggunakan bobot pre-trained ImageNet dengan modifikasi pada

lapisan fully connected layer untuk memproyeksikan fitur 32 class spesies mamalia laut. VGG16 menggunakan fungsi loss yang sama dengan ResNet50 untuk menerapkan optimasi yang identik yaitu dengan Weighted Cross-Entropy. Serta menggunakan two-phase training untuk menjaga stabilitas bobot dan memaksimalkan konvergensi model. Untuk lebih detail tentang arsitektur VGG16 dapat dilihat pada **Gambar 9** berikut.

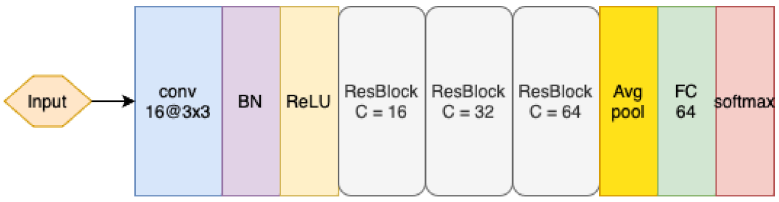
**Gambar 9**. VGG16 *Model Architecture*




### 3.4.3 WhaleNet

Penelitian ini mengembangkan arsitektur WhaleNet [2] oleh Carbone dan Licciardi (2024). Sebuah Convolutional Neural Network (CNN) ringan berbasis blok residual (BasicBlock) yang dirancang untuk mengklasifikasikan spectogram mamalia laut dengan efisien. Struktur WhaleNet terdiri dari tiga lapisan residual sekuensial dengan jumlah channel yang meningkat (16, 32, 64) untuk mengekstraksi fitur hierarkis dari input satu channel grayscale, diakhiri dengan Adaptive Average Pooling dan fully connected layer. Tahap model WhaleNet juga menerapkan fungsi loss Weighted Cross-Entropy serta menggunakan optimizer AdamW dengan OneCycleLR scheduler selama 100 epoch untuk memastikan konvergensi dari berbagai jenis input fitur (Mel, STFT, WST, dan Spectral Contrast). Arsitektur detail dari WhaleNet dapat dilihat pada **Gambar 10**.

**Gambar 10**. WhaleNet *Model Architecture*



### 3.4.4 Ensemble

Bagian pemodelan diakhiri dengan Ensemble Learning menggunakan pendekatan Soft Voting sebagai strategi fusi keputusan akhir. Metode ini bekerja dengan mengintegrasikan output probabilitas terprediksi dari setiap model dasar yang telah dilatih secara independen (WhaleNet, VGG16, dan ResNet-50). Secara matematis, prediksi kelas akhir $\hat{y}$ ditentukan oleh kelas dengan nilai rata-rata probabilitas tertinggi dari seluruh model yang digabungkan:

$$\hat{y} = \underset{j \in \{1,\ldots,C\}}{\operatorname{argmax}} \left( \frac{1}{M} \sum_{i=1}^{M} p_{i,j} \right)$$

Dimana $M$ adalah jumlah model dalam ensemble dan $p_{i,j}$ adalah probabilitas yang diprediksi oleh model ke-$i$ untuk kelas ke-$j$. Pendekatan ini memastikan bahwa performa klasifikasi dimaksimalkan dengan memanfaatkan kekuatan komplementer dari berbagai arsitektur.

# BAB IV
# HASIL DAN PEMBAHASAN

## 4.1 Hasil Eksperimen

Untuk menganalisis ekstraksi fitur mana yang memiliki metrics terbaik pada setiap model, dilakukan modeling pada setiap ekstraksi fitur dengan model-model yang diajukan. Dengan difokuskan pada metrics F1-Macro untuk mengidentifikasi kombinasi representasi fitur dan model yang paling efektif.

| Metode | Model | F1-Macro |
|---|---|---|
| STFT Spectogram | ResNet-50 | 0.7319 |
| | VGG16 | 0.7729 |
| | WhaleNet | 0.6952 |
| Mel Spectogram | ResNet-50 | 0.7177 |
| | VGG16 | 0.7484 |
| | WhaleNet | 0.7580 |
| WST Orde 1 | ResNet-50 | 0.7668 |
| | VGG16 | 0.7143 |
| | WhaleNet | 0.8009 |
| **Spectral Contrast** | ResNet-50 | 0.9542 |
| | VGG16 | 0.9419 |
| | **WhaleNet** | **0.9635** |

Berdasarkan hasil eksperimen, terlihat bahwa model mengalami improvement yang signifikan jika menggunakan metode ekstraksi fitur yang tepat. Pada

penggunaan fitur berbasis representasi time-frequency seperti STFT dan Mel Spectrogram, model cenderung memiliki rentang F1-Macro rata-rata terendah jika dibandingkan ekstraksi fitur lainnya yakni antara 0.69 hingga 0.77. Di mana VGG16 menunjukkan dominasi pada fitur STFT (0.7729) sedangkan WhaleNet lebih unggul pada Mel Spectrogram (0.7580). Metode Wavelet Scattering Transform (WST) Orde 1 terbukti mampu meningkatkan prediksi, khususnya pada arsitektur WhaleNet yang mengalami kenaikan performa signifikan menjadi 0.8009. Hal ini mengindikasikan bahwa sifat invariansi WST terhadap deformasi sinyal lebih efektif dalam menangkap karakteristik bioakustik dibandingkan spectogram konvensional.

Temuan utama adalah ekstraksi fitur Spectral Contrast sangat signifikan, yang mana secara konsisten *outperform* seluruh arsitektur model (ResNet-50, VGG16, dan WhaleNet) hingga 0.94. Peningkatan performa ini menunjukkan bahwa informasi dari tekstur spektral merepresentasikan perbedaan peaks dan valleys frekuensi. Lonjakan performa ini menegaskan bahwa informasi tekstur spektral yang merepresentasikan perbedaan energi antara peaks dan valleys frekuensi memiliki pengaruh jauh lebih kuat untuk membedakan spesies mamalia laut dibandingkan fitur lainnya. Dengan kombinasi terbaik yaitu WhaleNet dengan fitur Spectral Contrast mencapai nilai F1-Macro sebesar 0.9635. Hal ini membuktikan bahwa WhaleNet efektif sebagai arsitektur yang efisien dalam menangkap fitur kontras spektral dengan optimal.

## 4.1 Ensemble Model

Temuan utama adalah ekstraksi fitur Spectral Contrast sangat signifikan, yang mana secara konsisten *outperform* seluruh arsitektur model (ResNet-50, VGG16, dan WhaleNet) hingga 0.94. Peningkatan performa ini menunjukkan bahwa informasi dari tekstur spektral merepresentasikan perbedaan peaks dan valleys frekuensi. Lonjakan performa ini menegaskan bahwa informasi tekstur spektral yang merepresentasikan perbedaan energi antara peaks dan valleys frekuensi memiliki pengaruh jauh lebih kuat untuk membedakan spesies mamalia laut dibandingkan fitur lainnya. Dengan kombinasi terbaik yaitu WhaleNet dengan fitur Spectral Contrast mencapai nilai F1-Macro sebesar 0.9635. Hal ini membuktikan bahwa

WhaleNet efektif sebagai arsitektur yang efisien dalam menangkap fitur kontras spektral dengan optimal.

# BAB V
# PENUTUP

## 5.1 Kesimpulan

Penelitian ini memiliki motivasi karena urgensi pemantauan mamalia laut sebagai indikator kesehatan ekosistem perairan yang sering terkendala oleh kompleksitas sinyal bioakustik dan gangguan noise lingkungan. Tujuan utama dari studi ini adalah mengembangkan arsitektur klasifikasi yang robust melalui eksplorasi representasi fitur akustik multi-dimensi dan pendekatan pemodelan Deep Learning. Penelitian ini memiliki kontribusi pada penerapan strategi prapemrosesan adaptif melalui Non-Stationary Spectral Gating dan standardisasi sampling rate, serta pembuktian bahwa fitur Spectral Contrast memiliki ekstraksi fitur yang superior dibandingkan ekstraksi fitur lainnya dalam menangkap tekstur spektral vokalisasi spesies.

Berdasarkan hasil eksperimen dan evaluasi model, performa metrik terbaik dalam penelitian ini didapat pada model Soft Voting Ensemble. Model ini

mengintegrasikan kekuatan arsitektur WhaleNet, VGG16, dan ResNet-50 yang dilatih menggunakan fitur Spectral Contrast, di mana prediksi dari ketiga model digabungkan menggunakan pembobotan. Pendekatan ini terbukti mampu memaksimalkan generalisasi prediksi dengan memanfaatkan performa masing-masing model dengan mendapatkan metrics pada F1-Macro sebesar 0.9857 Sehingga menghasilkan model klasifikasi yang robust dan reliabel untuk mendukung upaya konservasi satwa laut secara otomatis.

## 5.2 Saran

Untuk pengembangan penelitian selanjutnya, model ini dapat membantu peneliti dalam memantau dan menjaga satwa mamalia laut langka, sehingga identifikasi spesies dapat dilakukan secara lebih cepat dan akurat. Model ini juga berpotensi dikembangkan lebih lanjut jika tersedia dataset yang cukup besar dan beragam, misalnya dataset audio seperti XLSR, sehingga kemampuan adaptasi terhadap variasi lingkungan dapat meningkat. Selain itu, strategi augmentasi data atau pengumpulan data terarah dapat diterapkan untuk mengatasi imbalance class, khususnya untuk spesies minoritas, agar distribusi data lebih seimbang. Terakhir, integrasi metode explainable AI (XAI) disarankan untuk meningkatkan interpretabilitas model, sehingga hasil prediksi dapat dipahami secara transparan dan digunakan sebagai dasar pengambilan keputusan konservasi.

## DAFTAR PUSTAKA


[1] J. Andén and S. Mallat, “Deep Scattering Spectrum,” *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014. doi: 10.1109/TSP.2014.232699.

[2] A. Licciardi and D. Carbone, “WhaleNet: A novel deep learning architecture for marine mammals vocalizations on Watkins marine mammal sound database,” arXiv:2402.17775 [eess.AS], Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.17775

[3] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980. doi: 10.1109/TASSP.1980.1163420.

[4] J. B. Allen, “Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 3, pp. 235–238, Jun. 1977. doi: 10.1109/TASSP.1977.1162950

[5] N. H. Bach, L. H. Vu, V. D. Nguyen, and D. P. Pham, “Classifying marine mammals signal using cubic splines interpolation combining with triple loss variational auto-encoder,” *Scientific Reports*, vol. 13, Art. no. 19984, 15 Nov. 2023. doi: 10.1038/s41598-023-47320-4.

[6] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, “Music type classification by spectral contrast feature,” in *IEEE Int. Conf. Multimedia and Expo (ICME)*, 2002, pp. 113–116. doi: 10.1109/ICME.2002.1024785.

[7] Y. Kong, X. Sun, Z. Sun, and J. Zhao, “Marine mammals and vessel noise classification based on multifeature fusion and rhythm pattern,” *Sensors*, vol. 24, no. 2, Art. no. 458, Jan. 2024. doi: 10.3390/s24020458.

[8] X. Meng, X. Liu, Y. Xu, Y. Wu, H. Li, K.-W. Kim, S. Liu, and Y. Xu, “A Multi-Time-Frequency Feature Fusion Approach for Marine Mammal Sound Recognition,” *J. Mar. Sci. Eng.*, vol. 13, no. 6, Art. no. 1101, May 2025. doi: 10.3390/jmse13061101.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, May 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[11] J. Andén and S. Mallat, “Deep scattering spectrum,” IEEE Trans. Signal Process., vol. 62, no. 16, pp. 4114–4128, Aug. 2014, doi: 10.1109/TSP.2014.2326991.

[12] Watkins Marine Mammal Sound Database, Woods Hole Oceanographic Institution and New College of Florida. [Online]. Available: https://whoicf2.whoi.edu/science/B/whalesounds/