Underwater Sounds Classification - Report

Biagio Andreucci
Mat: 0522501814
b.andreucci@studenti.unisa.it

Irene Gaita
Mat: 0522501839
i.gaita@studenti.unisa.it

Massimo Domenico Giuggiola
Mat: 0522501872
m.giuggiola@studenti.unisa.it

ABSTRACT

L'impatto delle attività umane sugli ecosistemi marini è in costante aumento, con la pesca illegale non dichiarata e non regolamentata che emerge come una delle principali minacce alla sostenibilità delle risorse ittiche e alla salute degli oceani. Lo smaltimento illegale di rifiuti tossici e la navigazione non autorizzata in aree protette aggravano ulteriormente la situazione contribuendo all'inquinamento marino e alla perdita di biodiversità. Questo studio mira a sviluppare strumenti e metodologie efficaci per rilevare e classificare i segnali acustici associati a tali attività illegali, utilizzando tecniche avanzate di analisi del segnale come lo scalogramma. Sono state affrontate due principali sfide di classificazione: una classificazione binaria per distinguere tra suoni Target e Non Target, e una classificazione multiclasse per identificare specifiche sottocategorie di attività. I risultati mostrano che, nella classificazione binaria, EfficientNet-B3 ha mantenuto un buon equilibrio delle metriche ed è risultata meno onerosa dal punto di vista computazionale rispetto ad AlexNet. Nella classificazione multiclasse, ResNet ha offerto migliori prestazioni. L'early stopping ha aiutato a mitigare l'overfitting e il bilanciamento del dataset si è rivelato cruciale per la generalizzazione del modello. Per il futuro, sarà importante ottimizzare ulteriormente i parametri di preprocessing, esplorare diverse tecniche di segmentazione e ampliare il dataset per migliorare le prestazioni delle classificazioni. Questo studio rappresenta un passo significativo verso l'implementazione di sistemi di monitoraggio acustico avanzati per la protezione degli ecosistemi marini e la conservazione della biodiversità.

Per ulteriori dettagli e per consultare il codice sorgente, è possibile visitare il repository su GitHub al seguente link:

https://github.com/IreneGaita/UnderwaterSoundsClassification

1 INTRODUZIONE

L'impronta umana sull'ambiente marino è in costante crescita, alimentata da una serie di attività antropiche che esercitano un impatto significativo sugli ecosistemi oceanici. Tra queste, le attività di pesca illegale, non dichiarata e non regolamentata emergono come una delle principali minacce per la salute degli oceani. La pesca compromette la sostenibilità delle risorse ittiche, danneggia gli ecosistemi marini e minaccia la sicurezza alimentare delle comunità costiere dipendenti dalla pesca. Inoltre, altre attività umane non autorizzate, come lo smaltimento illegale di rifiuti tossici e la navigazione in aree protette, contribuiscono ulteriormente all'inquinamento marino e alla perdita di biodiversità. Queste attività illegali e non autorizzate rappresentano una sfida significativa per la gestione e la conservazione degli oceani, richiedendo un'azione coordinata a livello globale per contrastarle efficacemente e proteggere gli ecosistemi marini per le generazioni future [1, 2].

La biodiversità è mantenuta da un equilibrio delicato e vulnerabile che può essere facilmente interrotto: la scomparsa di una singola specie potrebbe portare all'estinzione di un intero ecosistema. Questo concetto si applica anche agli esseri umani. Ci sono numerose comunità il cui sostentamento dipende interamente dalle risorse marine. La diminuzione di una o più specie di pesci potrebbe rapidamente mettere in pericolo milioni di persone [3].

In particolare, l'obiettivo di questo studio è quello di sviluppare strumenti e metodologie efficaci per rilevare e classificare i segnali acustici associati a tali attività. Questo approccio innovativo mira a fornire una soluzione tecnologica avanzata per il monitoraggio e la protezione degli oceani.

Lo studio in questione mira ad affrontare due principali sfide di classificazione: una classificazione binaria e una classificazione multiclasse. La classificazione binaria si concentra sulla distinzione tra suoni target, associati ad attività illegali o non autorizzate, e suoni non target, che non presentano tali caratteristiche. La classificazione multiclasse, invece, si prefigge di identificare a quale sottocategoria specifica appartiene l'audio target, permettendo una comprensione più dettagliata delle diverse tipologie di attività rilevate.

Grazie all'utilizzo di tecniche avanzate di analisi del segnale, come lo scalogramma, è possibile ottenere una rappresentazione dettagliata dei segnali acustici, facilitando così il processo di classificazione. Questo metodo permette di individuare pattern e caratteristiche distintive nei segnali, migliorando l'accuratezza e l'affidabilità delle classificazioni effettuate.

In conclusione, il presente studio rappresenta un passo significativo verso l'implementazione di sistemi di monitoraggio acustico avanzati per la protezione degli ecosistemi marini. L'applicazione di tecniche di analisi del segnale e di algoritmi di classificazione può offrire un valido supporto nella lotta contro le attività illegali e non autorizzate, contribuendo alla conservazione della biodiversità marina e alla salvaguardia delle risorse ittiche per le future generazioni.

2 STRUMENTI DI ANALISI

In questo capitolo, verranno esaminati i progressi della letteratura attuale sull'argomento in questione e verranno descritti gli strumenti teorici e pratici utilizzati per l'analisi del segnale. La sezione 2.1 fornirà una panoramica dei lavori correlati, delineando lo stato attuale della ricerca e identificando le principali scoperte e lacune. Successivamente, nella sezione 2.2, verrà esplorato lo scalogramma come strumento per l'analisi del segnale, illustrandone i principi teorici.

2.1 Related Works

Negli ultimi anni, i modelli di deep learning hanno dimostrato una crescente superiorità rispetto ai classificatori convenzionali nell'affrontare complessi problemi di classificazione, mostrando un notevole miglioramento nelle prestazioni e nella capacità di adattamento a una vasta gamma di contesti e tipologie di dati.

Il monitoraggio acustico passivo, ad esempio, genera una grande quantità di dati che, fino a poco tempo fa, venivano principalmente analizzati manualmente attraverso l'ascolto dell'audio. Tuttavia, questo metodo richiede un notevole sforzo lavorativo ed è soggetto a variazioni nella percezione umana, interferenze del rumore di fondo e alla necessità di validare le descrizioni di nuovi tipi di suoni prodotti dalla stessa specie [4].

Le tecniche di deep learning stanno rapidamente guadagnando popolarità nella scienza e nella ricerca marina, rivoluzionando il modo in cui analizziamo e comprendiamo gli ecosistemi marini. Uno dei settori in cui il deep learning ha avuto un impatto significativo è la biologia marina, dove viene utilizzato per l'identificazione e il monitoraggio delle specie marine, il rilevamento di habitat critici come le barriere coralline e le praterie di alghe, e la valutazione dei rischi ambientali come l'inquinamento e la perdita di habitat [5].

Nel contesto dello stato dell'arte nel campo della classificazione automatica dei rumori sottomarini, l'adozione di tecniche di preprocessing dei dati audio gioca un ruolo fondamentale per garantire l'accuratezza e l'affidabilità dei modelli di machine learning. Questi approcci includono l'utilizzo di spettrogrammi avanzati, che consentono la rappresentazione visuale e l'analisi delle caratteristiche temporali e frequenziali dei segnali audio.

Il lavoro di Mark Thomas et al. del 2020 [6] si basa su un dataset ottenuto tramite tecniche di monitoraggio acustico passivo, concentrato sulle vocalizzazioni di tre specie di balene: balene blu, balenottere comuni e balenottere boreali, registrate lungo il Scotian Shelf nel 2015 e 2016. Le registrazioni sono state standardizzate per catturare vocalizzazioni a bassa e alta frequenza, poi sottoposte a filtraggio per eliminare rumori indesiderati. Successivamente, sono stati generati spettrogrammi utilizzando variabili parametri STFT, con sovrapposizione della finestra FFT e limiti di frequenza definiti. In particolare, hanno valutato le prestazioni di due architetture CNN comunemente utilizzate, ossia: ResNet-50 e VGG-19 con normalizzazione batch.Le CNN sono state addestrate per un totale di 100 epoche. I risultati per ResNet-50, utilizzando 3 canali, indicano un'accuratezza del 95.3%, con una precisione dell'88.7%, una recall dell'87.1%, e un F-1 score dell'87.8%. Per quanto riguarda il modello VGG-19, anch'esso a tre canali, si registra un'accuratezza del 96.1%, con una precisione del 90.6%, una recall dell'89.2%, e un F-1 score dell'89.9%. Il DCS ha dimostrato la capacità di apprendere rappresentazioni generalizzabili, permettendo l'inclusione di altre specie senza la necessità di riaddestrare gli strati di estrazione delle caratteristiche convoluzionali della CNN. Il classificatore addestrato ha raggiunto un'accuratezza del 94.8%, una precisione dell'88.4%, e una recall dell'87.1%.

Parallelamente, lo studio di Cai et al. [7] si concentra sull' implementazione di un modello di classificazione parallela per i suoni dei mammiferi marini, basato su estrazione di caratteristiche multidimensionali e aumento dei dati. Il dataset principale utilizzato è il Watkins Marine Mammal Sound Database. Per quanto riguarda il preprocessing dei dati, inizialmente, viene standardizzato il tasso di campionamento di tutti i segnali audio a 10 kHz. Successivamente, viene estratta casualmente una porzione di 3 secondi di audio, che comprende la vocalizzazione o parte di essa. In aggiunta, vengono utilizzate tecniche di estrazione di caratteristiche, tra cui la Short-Time Fourier Transform (STFT) per trasformare i segnali audio dal dominio temporale al dominio frequenza-tempo e l'estrazione di diverse caratteristiche dal dominio della frequenza, come i coefficienti cepstrali della frequenza Mel (MFCC) e i Linear Frequency Cepstral Coefficients (LFCC). Per migliorare l'accuratezza della classificazione, i ricercatori hanno applicato il Data Augmentation ai dati sbilanciati, rendendo la distribuzione più equilibrata tra le specie. Il modello MDF-PNet ha utilizzato l'apprendimento del trasferimento per ottenere i pesi iniziali e una convergenza più rapida. Nella fase di selezione delle caratteristiche acustiche, il modello ha combinato multiple caratteristiche complementari nel dominio della frequenza, riqualificando un livello di connessione completo nel back-end per connettere i risultati della previsione dei diversi rami. MDF-PNet è stato testato mediante convalida incrociata k-fold e confrontato ripetutamente con ricerche simili. I risultati mostrano che utilizzando MDF-PNet insieme alla tecnica di Data Augmentation, la media dell'accuratezza su cinque test, che include accuratezza, loss, precisione, recall e F-1 Score, ha raggiunto un notevole 95.21%.

Nel campo della classificazione dei segnali acustici dei pesci, J, Amal R. et al. [8] si concentrano sull'utilizzo di un descrittore tempo-frequenza basato sulla wavelet per la classificazione automatica dei segnali acustici. Il dataset utilizzato comprende registrazioni acustiche di 102 diverse specie di pesci provenienti da fonti come FishBase e DOSITS, con varie condizioni di registrazione. Il preprocessing dei dati coinvolge la standardizzazione dei suoni in formato WAV e l'applicazione di un filtro Butterworth per rimuovere i rumori di fondo. Successivamente, le registrazioni sono segmentate utilizzando il metodo di segmentazione delle sillabe di Harma. La conversione dei dati audio prevede la trasformazione in spettrogrammi utilizzando la Trasformata di Fourier e l'estrazione di diverse caratteristiche dal dominio della frequenza, inclusi i coefficienti cepstrali della frequenza Mel (MFCC) e i Linear Frequency Cepstral Coefficients (LFCC). Estratte le sillabe, LWAP e coefficienti di approssimazione utilizzando DWT, essi vengono combinati insieme per formare la funzione vettore. DA, k-NN e SVM vengono utilizzati nella fase di classificazione. Dai risultati sperimentali emerge che l'uso di DA e SVM garantisce una maggiore precisione per il sistema proposto, con valori di accuratezza pari al 89.82% e al 56.48%, rispettivamente.

Inoltre, il lavoro di Noda et al. [9] si concentra sull'uso delle registrazioni acustiche dei pesci per la classificazione tassonomica automatica. Utilizzando registrazioni provenienti da FishBase e DOSITS, i dati sono standardizzati e sottoposti a preprocessing, incluso il campionamento uniforme a 44,1 kHz e il denoising con un filtro passa-basso. Le registrazioni vengono automaticamente segmentate in sillabe, consentendo l'isolamento dei suoni biologici dai rumori di fondo. La conversione dei dati audio include l'estrazione di caratteristiche come i coefficienti cepstrali della frequenza Mel (MFCC) e i Linear Frequency Cepstral Coefficients (LFCC) tramite la Trasformata di Fourier. Gli MFCC sono risultati più efficienti nel modellare i segnali acustici a basse frequenze, con un miglioramento significativo delle prestazioni aggiungendo informazioni sulle alte frequenze tramite LFCC. Incorporare informazioni temporali come l'entropia di Shannon e la lunghezza della sillaba ha rafforzato il sistema di classificazione. Le prestazioni del sistema sono risultate simili utilizzando diversi tipi di algoritmi di apprendimento automatico come KNN, RF e SVM. Tuttavia, è stato dimostrato che l'SVM è più efficace nel riconoscimento bioacustico grazie ai dati

che mostrano relazioni non lineari. Dividendo il dataset in due sottoinsiemi, uno per i suoni naturali e uno per i suoni innaturali, si sono ottenuti risultati medi di accuratezza del 95.21% per i suoni innaturali e del 98.17% per i suoni naturali, considerando tre metriche: precisione, recall e F-1 score.

Lo studio di Liu et al. [10] si concentra sulla classificazione dei fischi dei cetacei utilizzando reti neurali convoluzionali (CNN). Utilizzando il Watkins Marine Mammal Sound Database, i dati sono standardizzati con un tasso di campionamento a 48 kHz e viene effettuato il denoising per eliminare i rumori di bassa frequenza. Le registrazioni vengono segmentate e convertite in spettrogrammi tramite la Short-Time Fourier Transform (STFT). Successivamente, una CNN a sette strati viene addestrata per classificare i fischi dei cetacei in diverse categorie, utilizzando gli spettrogrammi come ingresso. Basandosi sulla superiorità delle CNN nella classificazione delle immagini, questo documento applica tale tecnologia alla classificazione dei fischi. I risultati sperimentali dimostrano che questo metodo può classificare automaticamente i fischi in modo efficace. La scelta dei parametri della rete neurale è stata cruciale per il risultato. Per quanto riguarda la funzione di attivazione, è stata scelta ReLU. I risultati dei test sono stati notevoli: su un totale di 8 test effettuati, la funzione ReLU ha registrato una percentuale di accuratezza del 100% in 4 test e del 93,3% in un altro test. Questi risultati dimostrano l'efficacia di ReLU nel contesto specifico di questo studio.

Nel lavoro di Demir et al. [11], il preprocessing inizia con il ricampionamento dei dati audio, standardizzando ogni campione a un tasso di campionamento di 4 kHz e segmentando i file audio in segmenti di 2 secondi. Successivamente, i dati audio vengono convertiti in spettrogrammi tramite la Trasformata di Fourier, sperimentando diverse lunghezze di finestra di Tukey per modulare il trade-off tra risoluzione temporale e frequenziale. Dopo aver pre elaborato i dati, utilizzano EfficientNet-B3 ed EfficientNet-B7 come modelli di estrazione delle funzionalità. Entrambi i modelli hanno ottenuto ottime prestazioni per le classi soft-negative e positive. Il modello EfficientNet-B3 ha registrato una precisione del 93% e un F1-score del 94%, mentre il modello EfficientNet-B7 ha ottenuto sia una precisione che un F1-score del 93%.

D'altra parte, nel lavoro di Tao Lu e Baokun Han e Fanqianhui Yu [12] concerne l'applicazione di AlexNet con l'apprendimento del trasferimento per rilevare e classificare i suoni di tre mammiferi marini, il preprocessing inizia con il ricampionamento dei dati audio a un tasso uniforme di 10 kHz. Le registrazioni audio vengono quindi convertite in spettrogrammi utilizzando la trasformata rapida di Fourier non uniforme (NFFT) con finestra di Hamming, generando immagini RGB che fungono da ingresso per il modello di classificazione. Successivamente, gli spettrogrammi vengono suddivisi in sottospettogrammi di lunghezza specifica. Questo studio ha approfondito l'applicazione di AlexNet con l'apprendimento del trasferimento per rilevare e classificare i suoni di tre mammiferi marini. Sfruttando la CNN, ha trasformato il compito di classificare questi segnali in un problema basato sulla visione artificiale, sfruttando l'abilità della CNN. Questo approccio elude i limiti dei tradizionali metodi basati su funzionalità fisse. I risultati hanno evidenziato le notevoli prestazioni del metodo, caratterizzate da efficienza e affidabilità. In particolare, i modelli di rilevamento hanno raggiunto un'accuratezza massima del 99,96%, mentre i modelli di classificazione hanno ottenuto un'accuratezza del 97,42%.

Infine, la ricerca di Dexin Duan et al. [13] si concentra sull' identificazione in tempo reale delle chiamate dei mammiferi marini utilizzando reti neurali convoluzionali (CNN). Questo documento propone un sistema basato su CNN per identificare le chiamate dei mammiferi marini in tempo reale, consentendo un monitoraggio più efficace delle popolazioni di mammiferi marini e una risposta più rapida alle emergenze ambientali. L'uso di CNN permette una maggiore precisione e affidabilità nell'identificazione delle chiamate dei mammiferi marini, migliorando così le pratiche di conservazione e protezione marina. Il modello sviluppato è stato testato utilizzando i dati del test e ha raggiunto una precisione superiore al 91% e un tasso di richiamo superiore all'84%. Inoltre, il modello è così veloce che l'identificazione può avvenire in tempo reale.

In contrapposizione ai lavori analizzati finora, lo studio di Ibrahim, Ali K et al. [14] descrive un approccio che utilizza gli scalogrammi come tecnica per migliorare la classificazione dei suoni delle cernie. I dati acustici, raccolti da varie fonti come database di registrazioni sonore marine, contengono suoni di cernia, suoni ambientali e rumori antropogenici. Il preprocessing ha coinvolto la rilevazione della soglia di rumore, impiegata per eliminare frequenze al di sotto del range minimo rilevabile da microfoni subacquei con la frequenza massima catturabile fissata a 8 kHz.

Successivamente, gli scalogrammi sono stati ottenuti tramite la trasformata wavelet continua (CWT), che decomprime un segnale in funzione di onde dilatate e traslate, offrendo una rappresentazione ad alta risoluzione sia nel tempo che nella frequenza. Questi scalogrammi sono stati convertiti in immagini RGB utilizzando funzioni specifiche di MATLAB per essere utilizzati nei modelli di deep learning.

Per la classificazione, sono stati utilizzati modelli di reti neurali convoluzionali preaddestrate (CNN) come VGG16, VGG19, Inception V3, GoogleNet e AlexNet, che sono stati adattati attraverso l' apprendimento per trasferimento. Gli ultimi quattro strati di questi modelli sono stati sostituiti con nuovi strati specifici per la classificazione dei suoni delle cernie, includendo due strati completamente connessi, uno strato di dropout e uno strato softmax.

I risultati hanno dimostrato che le CNN preaddestrate offrono prestazioni superiori rispetto ai metodi basati su caratteristiche manuali, con accuratezze notevoli per specie specifiche. Sebbene non sia stata osservata una differenza significativa tra l'uso di scalogrammi e spettrogrammi, entrambi i tipi di rappresentazione si sono rivelati efficaci. I modelli VGG19, AlexNet, GoogleNet e Inception V3 hanno ottenuto valori elevati nella metrica di accuratezza nell'individuazione dei suoni acustici delle quattro specie di cernie. Utilizzando gli spettrogrammi, AlexNet ha registrato un'accuratezza del 95,6% per la specie Epinephelus guttatus e del 91,2% per la specie Mycteroperca venenosa; VGG19 ha ottenuto un'accuratezza dell'89,2% per la specie Epinephelus striatus; Inception V3 ha raggiunto il 100% di accuratezza per la specie Mycteroperca bonaci. Utilizzando gli scalogrammi, VGG19 ha registrato un'accuratezza del 95,1% per la specie Epinephelus guttatus, del 91,3% per la specie Epinephelus striatus e del 92,9% per la specie Mycteroperca venenosa; GoogleNet e Inception V3 hanno entrambi ottenuto un'accuratezza del 100% per la specie Mycteroperca bonaci.

2.2 Scalogramma

Gli scalogrammi rappresentano uno strumento fondamentale nell' analisi del segnale, offrendo una prospettiva unica rispetto agli spettrogrammi tradizionali. Mentre gli spettrogrammi si basano sulla trasformata di Fourier e forniscono una rappresentazione tempo-frequenza dei segnali, gli scalogrammi si basano sull'analisi Wavelet e offrono una visione più dettagliata della struttura temporale dei segnali.

L'analisi degli scalogramma si basa sulla Continuous Wavelet Transform (CWT), che è dimostrata essere una trasformazione isometrica, preservando l'energia del segnale, questo permette di visualizzare l'energia del segnale in funzione delle diverse scale temporali. La formula sottostante descrive la Continuous Wavelet Transform (CWT).

$$CWT(a,b) = \int_{-\infty}^{\infty} x(t) \cdot \psi^* \left(\frac{t-b}{a} \right) dt$$

Dove $x(t)$ è il segnale in ingresso, $\psi(t)$ è la Wavelet madre, a è il fattore di scala che controlla la larghezza della Wavelet e b è il fattore di traslazione che controlla la posizione temporale della Wavelet [14].

Come sottolineato da Grossmann e Morlet, gli scalogrammi forniscono una rappresentazione dettagliata della struttura temporale dei segnali, consentendo di catturare variazioni temporali sottili e di rappresentare dettagli a diverse risoluzioni temporali. A differenza degli spettrogrammi, dove l'area di influenza di una determinata frequenza rimane costante, negli scalogrammi questa area varia in base alla frequenza stessa, grazie all'analisi della scala temporale logaritmica.

Inoltre, gli scalogrammi presentano una distribuzione più localizzata dell'energia del segnale attorno alle piccole scale temporali, fornendo una visione più chiara dei dettagli temporali del segnale. Questa caratteristica è particolarmente utile quando si analizzano segnali complessi, come quelli marini, dove possono esserci variazioni rapide e sottili nel tempo.

Nonostante gli scalogrammi e gli spettrogrammi forniscono entrambi una rappresentazione visiva bidimensionale dei segnali, gli scalogrammi offrono un livello di dettaglio superiore nella rappresentazione della struttura temporale dei segnali. Questo aspetto è cruciale nell'analisi di segnali complessi come quelli marini, dove la comprensione dei dettagli temporali può fornire informazioni preziose sulla natura e sull'origine del segnale [15].

3 DESCRIZIONE DEL DATASET

I suoni Target e Non Target utilizzati in questo studio sono stati selezionati da vari dataset, inclusi il database "Watkins Marine Mammal Sound" [16]. Questo ampio database raccoglie registrazioni acustiche risalenti dagli anni '40 agli anni 2000.

Oltre al Watkins Marine Mammal Sound, sono stati impiegati altri dataset come "A Collection of Sounds from the Sea", "SanctSound", "Sounds in the Ocean: Mammals" della NOAA [17], "Sounds Recorded in Glacier Bay" del National Park & Preserve Alaska [18], registrazioni Avisoft [19], "Marine Mammals Bioacoustic of Australia and Antarctica" [20], "Ocean Conservation Research" [21], e "Deepship" [22]. Questi dataset forniscono una vasta gamma di suoni marini, suoni ambientali e rumori specifici delle attività umane subacquee. Il dataset utilizzato per l'analisi è stato suddiviso in due categorie principali: "Non Target" e "Target", la Tabella 1 mostra la categorizzazione, con una lista dettagliata dei dataset di provenienza.

Categoria	Lista dei dataset da cui provengono i file
Non Target	Watkins Marine Mammal Sound A Collection of Sounds from the Sea Sounds Recorded in Glacier Bay SanctSound Ocean Conservation Research Avisoft Sounds in the Ocean: Mammals Marine Mammals Bioacoustic of Australia and Antarctica
Target	A Collection of Sounds from the Sea Sounds Recorded in Glacier Bay SanctSound Ocean Conservation Research Deepship

Table 1: Origine dei File nel Dataset

La categoria "Non Target" include 2000 file audio rappresentando una vasta gamma di suoni ambientali e prodotti da numerose specie marine e attività umane subacquee. Tra i suoni ambientali troviamo uragani, terremoti, eruzioni vulcaniche, vento, onde, e pioggia. Le specie marine rappresentate comprendono mammiferi come balene, delfini, foche e altri cetacei come il capodoglio e l'orca. Inoltre, sono inclusi suoni di altri animali marini come dugonghi, manati e pesci.

La categoria "Target" comprende 663 file audio, che si focalizzano principalmente su suoni generati da una varietà di dispositivi e attività umane subacquee. Tra questi, ci sono dispositivi di disturbo acustico, dispositivi di rilascio acustico, sonar, esplosioni, e sonar da pesca. Inoltre, comprende suoni prodotti da diverse tipologie di navi, come navi cargo, navi passeggeri, petroliere e rimorchiatori. Sono inclusi anche suoni provenienti da sommozzatori, segnali di comunicazione subacquea e attività di elaborazione dei combustibili fossili sul fondo marino. Altri suoni "Target" rilevanti sono le indagini sismiche con airgun e le seal bomb. Questi suoni rappresentano una vasta gamma di rumori antropogenici subacquei, utili per l'analisi delle interferenze umane sull'ambiente marino. La Tabella 2 riassume il numero totale di file audio per ciascuna categoria.

Categoria	Numero di File Audio
Non target	2.000
Target	663

Table 2: Distribuzione dei File Audio all'interno del Dataset

3.1 Rimozione Duplicati

Durante la fase di analisi del dataset, sono stati identificati 25 file audio duplicati nella categoria "Non Target" e provenienti dal Dataset Watkins Marine Mammal Sound, con ciascun duplicato contenente una copia ridondante. Pertanto, un totale di 50 file è stato rimosso dal dataset. Questi duplicati sono stati eliminati poiché erano presenti in più sottoclassi della categoria "Non Target" e non

era possibile determinare con certezza a quali sottocategorie appartenessero. Questa ambiguità avrebbe potuto introdurre errori e dubbi nell'analisi, compromettendo la validità dei risultati.

Dopo la rimozione, il numero di file nella categoria "Non Target" è diminuito da 2000 a 1950. Conteggiando anche i 663 file della categoria "Target", il dataset totale utilizzato per l'analisi è risultato essere composto da 2613 file audio.

La Tabella 3 sintetizza la distribuzione finale dei file audio all'interno del dataset utilizzato per l'analisi, mentre la Figura 1 mostra il numero di file contenenti in ogni sottocategoria.

Categoria	Numero di File Audio
Non target	1.950
Target	663
Totale	2.613

Table 3: Distribuzione dei File Audio nel Dataset con la Rimozione dei Duplicati

4 DATA PREPROCESSING

Nel processo di preprocessing dei dati audio, sono state adottate una serie di decisioni strategiche per garantire una rappresentazione accurata ed efficiente dei segnali audio nel dataset.

4.1 Ricampionamento

Per garantire la cattura di tutti i dettagli essenziali del segnale senza spreco di risorse computazionali e di archiviazione è stata effettuata un'analisi approfondita dei file audio per effettuare il ricampionamento.

È stata scelta una frequenza massima di 43200 kHz siccome la maggior parte dei file audio ha una frequenza inferiore, come illustrato nella Figura 2, di conseguenza, è stata selezionata una frequenza di campionamento di 86400 kHz adottando il criterio di Nyquist per evitare l'aliasing e assicurare una rappresentazione accurata del segnale nel dominio digitale del dataset [23].

Per l'analisi spettrale localizzata nel tempo, è stata impiegata la trasformata di Fourier a breve termine (STFT), definita come:

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau) \cdot w(t-\tau) \cdot e^{-i2\pi f\tau} d\tau$$

dove $x(\tau)$ rappresenta il segnale audio, $w(t-\tau)$ è la finestra temporale, e f è la frequenza [10]. L'impiego della STFT ha migliorato significativamente la nostra capacità di analizzare e interpretare i segnali audio, consentendo una migliore comprensione delle loro caratteristiche spettrali e temporali.

4.2 Ridimensionamento a lunghezza fissa

Standardizzare la durata dei file audio assicura uniformità nei dati di input, fondamentale per il modello di apprendimento, poiché migliora la coerenza dei risultati. Per questo motivo, è stata condotta un'analisi, illustrata nella Figura 3, su tutti i file audio compresi tra 0 e 15 secondi, per determinare una durata ottimale di segmentazione.

L'analisi della distribuzione della durata ha mostrato una mediana di 3 secondi, che rappresenta la durata tipica degli audio brevi presenti nel dataset. Questo significa che metà dei file audio ha una durata inferiore o uguale a 3 secondi, mentre l'altra metà ha una durata superiore. Utilizzare la mediana come riferimento ci ha permesso di rappresentare equamente la maggior parte dei file audio nel dataset, garantendo una rappresentazione equilibrata e migliorando l'uniformità dei dati di input.

Per tutti i file la cui durata risultava essere minore di 3 secondi, si è scelto di aggiungere una porzione randomica di audio agli stessi.

4.3 Conversione del formato dei file audio

Per avere coerenza nel processo di gestione ed elaborazione dei file audio, garantendo una qualità audio ottimale per le applicazioni in cui vengono impiegati, è stato adottato il procedimento di conversione dei file MP3 in formato WAV. Il formato MP3 utilizza una tecnica di compressione lossy, che comporta una riduzione della qualità del suono poiché rimuove informazioni considerate non essenziali dall'audio. In contrasto, i file WAV utilizzano una compressione lossless, che mantiene tutte le informazioni audio originali, garantendo una fedeltà superiore del suono. La Tabella 4 illustra la distribuzione dei file audio in base al formato di origine. Anche se la qualità persa con la compressione MP3 non può essere completamente recuperata, la conversione in formato WAV impedisce ulteriori perdite di qualità e standardizza il formato dei file audio.

Formato	Numero di File Audio
.wav	2.546
.mp3	67

Table 4: Distribuzione dei File Audio secondo il Formato

4.4 Profondità di bit

La profondità di bit audio determina quanti valori di ampiezza possono essere registrati per ogni campione audio. Un aumento della profondità di bit consente di catturare un numero maggiore di valori di ampiezza per ciascun campione, migliorando così la precisione nella riproduzione del segnale audio originale [24].

Il PCM (Pulse Code Modulation) è invece un metodo standard per la rappresentazione digitale dei segnali audio, che trasforma segnali analogici in forma digitale tramite il campionamento del segnale a intervalli regolari nel tempo e la conversione di ciascun campione in un valore numerico [25].

Per migliorare la qualità del suono, garantire uniformità e aumentare l'efficienza computazionale, è stato deciso di standardizzare la profondità di bit a 16 bit PCM, una scelta che offre numerosi vantaggi. Innanzitutto, la maggior parte dei file audio già utilizza questo formato, come mostrato nella Tabella 5. Inoltre, i 16 bit PCM offrono una qualità del suono adeguata per la maggior parte delle applicazioni. Questo livello di risoluzione consente una rappresentazione dettagliata del segnale audio senza occupare troppo spazio di archiviazione. L'uso di 16 bit migliora anche l'efficienza nei processi di elaborazione audio, riducendo la quantità di dati da elaborare.

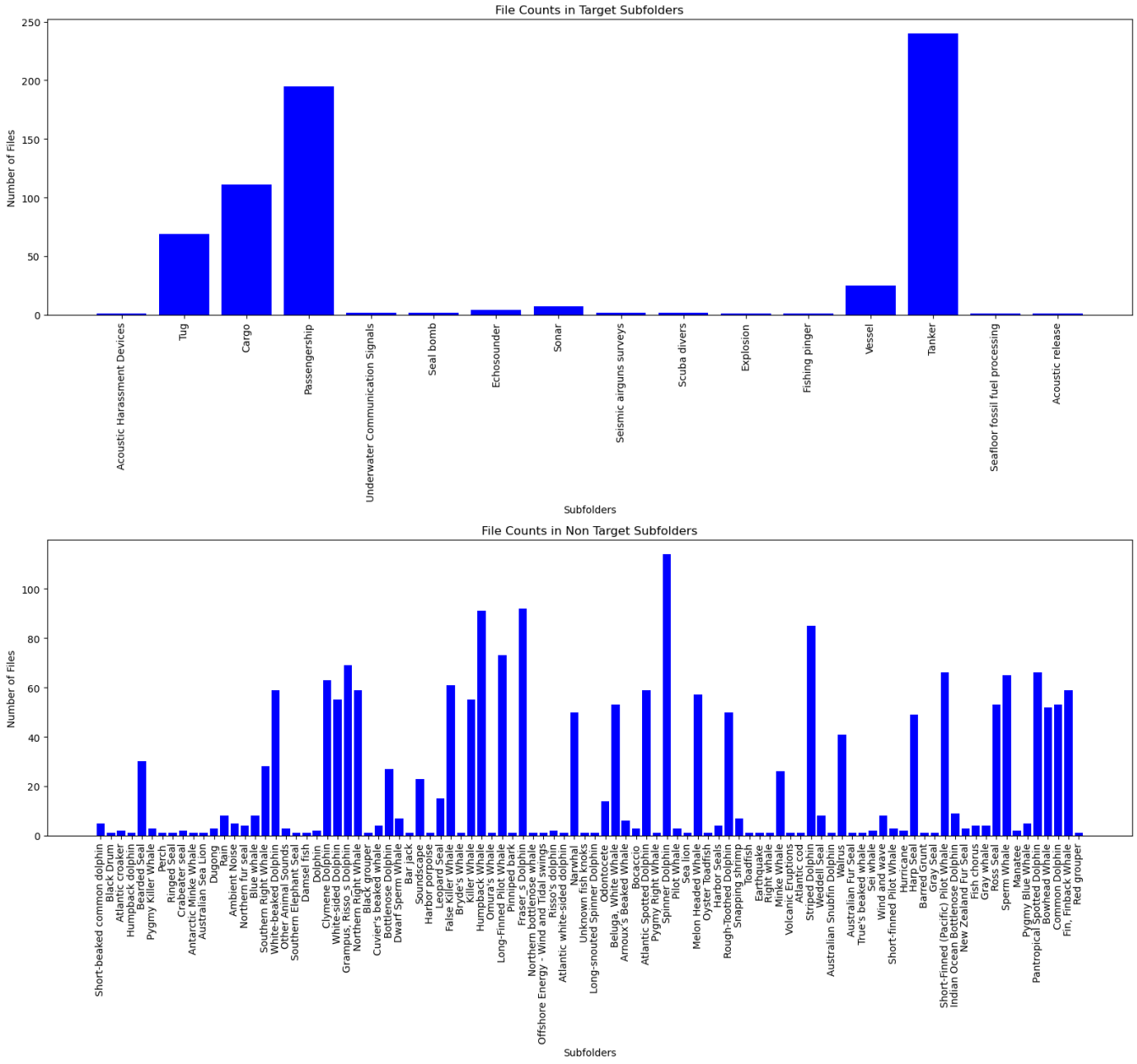

Figure 1: Distribuzione dei File Audio nelle Rispettive Sottocategorie

Bit Depth	Numero di File Audio
8 bit PCM	17
16 bit PCM	1.895
24 bit PCM	23
32 bit PCM	6
32 bit float	605

Table 5: Distribuzione dei File Audio secondo la Profondità di Bit

4.5 Conversione dei file audio in monocanale

Nella Tabella 6 è possibile osservare la distribuzione dei file, che ha guidato la decisione di convertire i dati audio stereo in formato monocanale, considerando che la maggior parte dei dati disponibili era già in questa forma.

Questa conversione da stereo a monocanale assicura coerenza nel trattamento dei dati, riducendo contemporaneamente la complessità computazionale e le risorse necessarie per l'allenamento del modello.

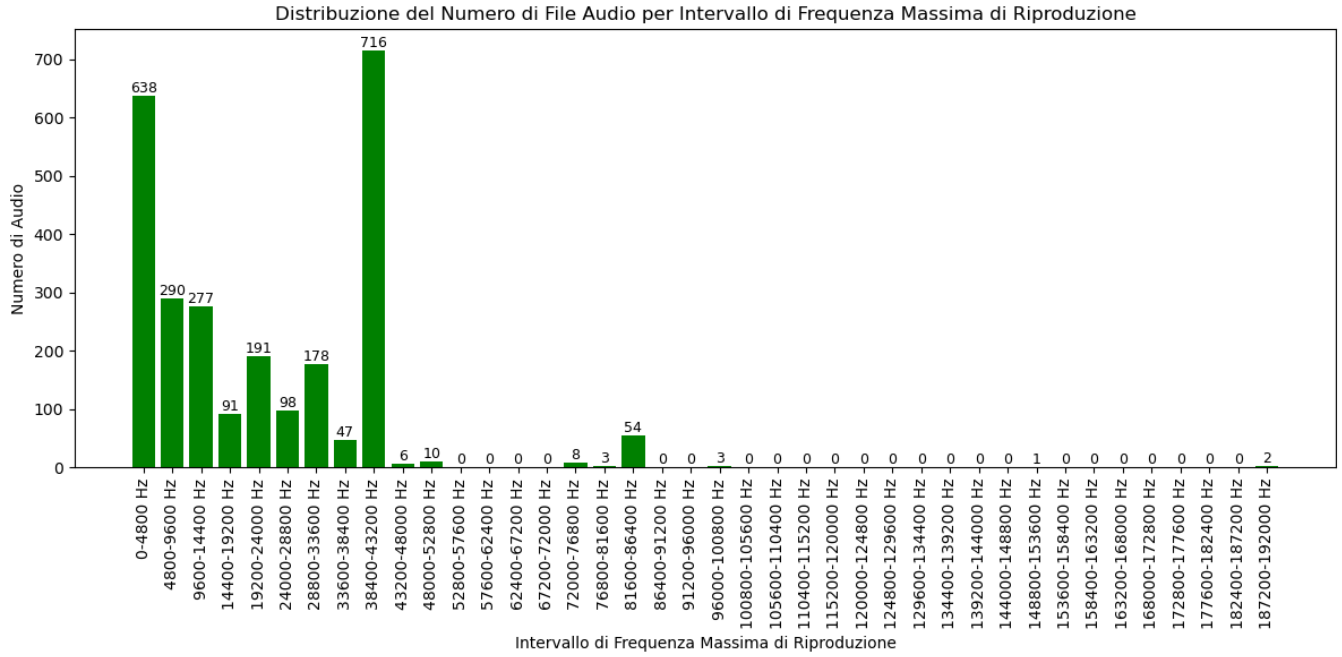

Figure 2: Distribuzione delle Frequenze Massime di Riproduzione degli Audio nel Dataset

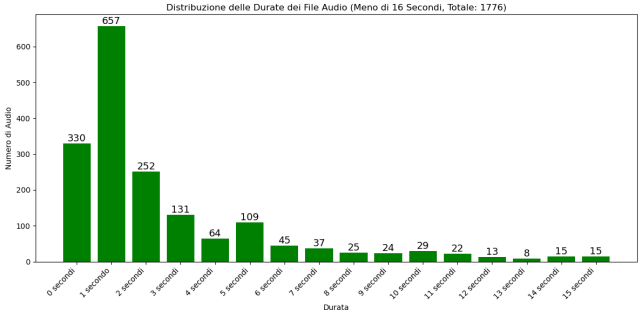

Figure 3: Distribuzione delle Durate degli Audio nel Dataset

Formato	Numero di File Audio
Monocanale	2.548
Stereo	65

Table 6: Distribuzione dei File Audio secondo il Formato dei Canali

4.6 Generazione dei scalogrammi

Nella fase di preprocessing dei dati audio, è stato deciso di generare scalogrammi per rappresentare visivamente il contenuto spettrale dei segnali. Motivo per cui, è stato utilizzata la Trasformata Wavelet Continua (CWT) con la Wavelet Morlet e scale variabili da 2 a 250. Questa scelta di scale è stata attentamente adattata alle potenze di calcolo disponibili.

Inizialmente, è stato scelto di sperimentare con una scala massima di 120, ma il risultato era inutilizzabile per il modello a causa della perdita di dettagli essenziali. La Figura6 rappresenta uno dei scalogrammi generati con la scala in questione.

Al contrario, una scala massima di 500 avrebbe fornito un dettaglio maggiore, ma incontrava limiti hardware significativi. Pertanto, la gamma di scale da 2 a 250 garantisce una rappresentazione dettagliata e accurata delle caratteristiche temporali e frequenziali dei segnali audio, pur rimanendo all'interno delle capacità computazionali disponibili.

I file audio sono stati caricati preservando la frequenza di campionamento scelta per mantenere le caratteristiche del segnale.

Gli scalogrammi sono stati salvati in formato PNG con dimensioni di 10x5 pollici per assicurare un equilibrio ottimale tra qualità visiva e gestione efficiente dello spazio di archiviazione.

Nella Figura 4, è possibile osservare uno dei scalogrammi generati della categoria "Target", sottocategoria "Cargo". Nella Figura 5 è rappresentato uno scalogramma della categoria "Non Target", sottocategoria "Blue Whale".

Terminate le fasi di preprocessing, il dataset finale è risultato essere composto da 66.824 scalogrammi.

5 ESPERIMENTI

Gli esperimenti proposti sono stati sviluppati utilizzando tre diverse workstation: un MacBook Pro 2021 con chip Apple M1, un MacBook Pro con chip Apple M3 Pro e un HP Spectre x360 dotato di scheda grafica NVIDIA® GeForce RTX™ 4050 e CPU Intel i7 Ultra. Queste macchine, grazie alle loro specifiche tecniche avanzate, hanno permesso di sfruttare le capacità computazionali necessarie per l'addestramento del modello.

Per affrontare efficacemente le due sfide principali di classificazione binaria e multiclasse, sono stati condotti esperimenti per

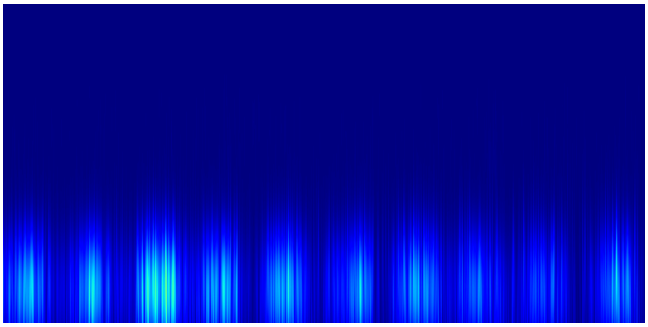

Figure 4: Scalogramma di un segnale audio rappresentato con la Trasformata Wavelet Continua (CWT) della sottocategoria "Cargo" - Target con scala 2 a 250

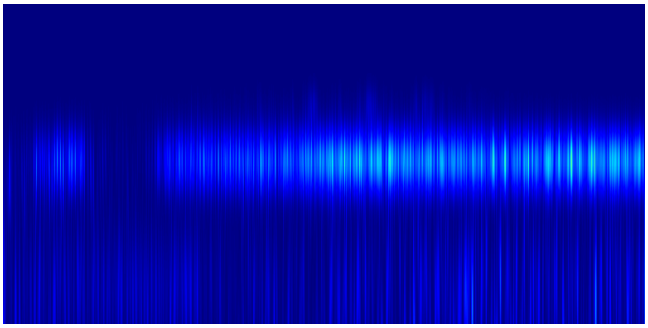

Figure 5: Scalogramma di un segnale audio rappresentato con la Trasformata Wavelet Continua (CWT) della sottocategoria "Blue Whale" - Non Target con scala 2 a 250

Figure 6: Scalogramma di un segnale audio rappresentato con la Trasformata Wavelet Continua (CWT) della sottocategoria "Blue Whale" - Target con scala 2 a 120

valutare come i modelli apprendono dal set di dati proposto. A tale scopo, sono stati creati parallelamente due subset a partire dal dataset originale. Il primo subset è stato generato utilizzando i sottoscalogrammi ottenuti durante il preprocessamento dei dati, mentre il secondo subset è stato costruito basandosi sui file audio originali, non preprocessati.

Entrambi i subset sono stati accuratamente creati e divisi in tre gruppi separati: uno per il training, uno per la validazione e uno per il test. Queste suddivisioni seguono una distribuzione del 70%

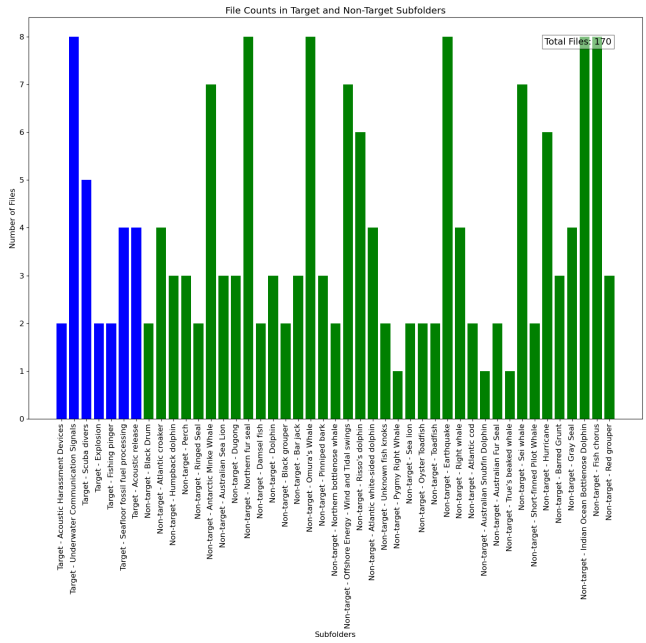

Figure 7: Distribuzione dei File Audio Preprocessati nelle Sottocategorie Minoritarie

per il training, del 15% per la validazione e del restante 15% per il test. Questa strategia è stata pianificata per ottimizzare il processo di sviluppo del modello, consentendoci di valutare accuratamente le sue prestazioni su dati precedentemente non osservati durante le fasi di addestramento e validazione.

Durante l'analisi della distribuzione dei file di entrambi i subset creati, è emerso che alcune sottocategorie contenevano meno di 10 file, identificandosi così come sottocategorie minoritarie. Per mitigare questa carenza di dati, è stata valutata l'opzione di generare ulteriori file, garantendo almeno 10 file per sottocategoria tramite la creazione di nuovi dati e l'introduzione di vari livelli di rumore. Tuttavia, questo approccio avrebbe potuto compromettere la rappresentatività dei dati.

Di conseguenza, in seguito all'analisi del subset contenente i sottoscalogrammi, sono stati eliminati in totale 170 di essi. Nella Figura 7 è possibile osservare la distribuzione dei scalogrammi appartenenti alle sottocategorie minoritarie. Tale scelta è stata motivata dal fatto che, su un totale di 66.824 scalogrammi, l'aggiunta di ulteriori dati per le sottocategorie minoritarie avrebbe potuto introdurre una distorsione dei dati, replicando continuamente gli stessi scalogrammi e minando l'integrità del dataset.

Nella Tabella 7 è possibile osservare la distribuzione dei sottoscalogrammi nei tre sottoinsiemi distinti per la validazione, l'addestramento e il test.

Partizione	Numero di scalogrammi
Addestramento	46.629
Validazione	10.028
Test	9.997

Table 7: Distribuzione dei File Audio Preprocessati in ciascuna Partizione

Per adattare il subset alle necessità dei modelli, la categoria "Target" è stata inizialmente bilanciata partendo dalla sottocategoria "Passengership", identificata come la sottocategoria con il maggior numero di campioni, con un totale di 10.934 scalogrammi come mostra la Figura 8. Tutte le altre sottocategorie della stessa categoria sono state poi regolate allo stesso numero di campioni tramite tecniche di Data Augmentation descritte nel paragrafo 5.1, raggiungendo così un totale di 98.406 file per la categoria in questione.

Successivamente, la categoria "Non Target" è stata bilanciata partendo dalla sottocategoria "Humpback Whale", identificata come la sottocategoria con il maggior numero di campioni, con un totale di 1.395 scalogrammi, come mostrato nella Figura 8. Tutte le altre sottocategorie sono state poi regolate allo stesso numero di campioni tramite tecniche di Data Augmentation descritte nel paragrafo 5.1, raggiungendo così un totale di 79.515 file per la categoria in questione.

In seguito, tutti i sottoscalogrammi sono stati accorpati alle categorie di provenienza "Target" e "Non Target" eliminando tutte le sottocategorie.

Nella figura 9 è possibile osservare la distribuzione dei scalogrammi all'interno delle due categorie principali.

Per non influenzare negativamente le prestazioni del modello, è stato deciso di omologare la categoria "Non Target" alla categoria "Target", applicando le stesse tecniche di Data Augmentation sui scalogrammi originali.

Nel secondo approccio adottato, è stato invertito l'ordine di preprocessamento dei dati rispetto alla fase di suddivisione del dataset. Pertanto, è stata eseguita la suddivisione utilizzando direttamente i file audio non preprocessati (ovvero quelli appartenenti al dataset originale), prima di procedere alla fase successiva di preprocessamento.

Per quanto riguarda la suddivisione del subset creato rispetto ai file audio non preprocessati, è emerso che le sottocategorie minoritarie contavano 180 audio su un totale di 2.613 file. Anche in questo caso, si è scelto di eliminare le sottocategorie mostrate in Figura 10 per conservare l'integrità del subset.

Partizione	Numero di scalogrammi
Addestramento	1.685
Validazione	381
Test	367

Table 8: Distribuzione dei File Audio Non Preprocessati in ciascuna Partizione

Per adattare il subset alle necessità dei modelli, è stato prima bilanciato il numero di campioni nella categoria "Target", partendo dalla sottocategoria "Passengership", che è stata identificata come quella con il maggior numero di campioni, totalizzando 11.589 scalogrammi come mostrato nella Figura 11. Le altre sottocategorie della stessa categoria sono state successivamente equiparate al medesimo numero di campioni utilizzando tecniche di Data Augmentation descritte nel paragrafo 5.1, portando così il totale a 57.945 file per la categoria considerata.

In seguito, è stato bilanciato il numero di campioni nella categoria "Non Target", partendo dalla sottocategoria "Humpback Whale",

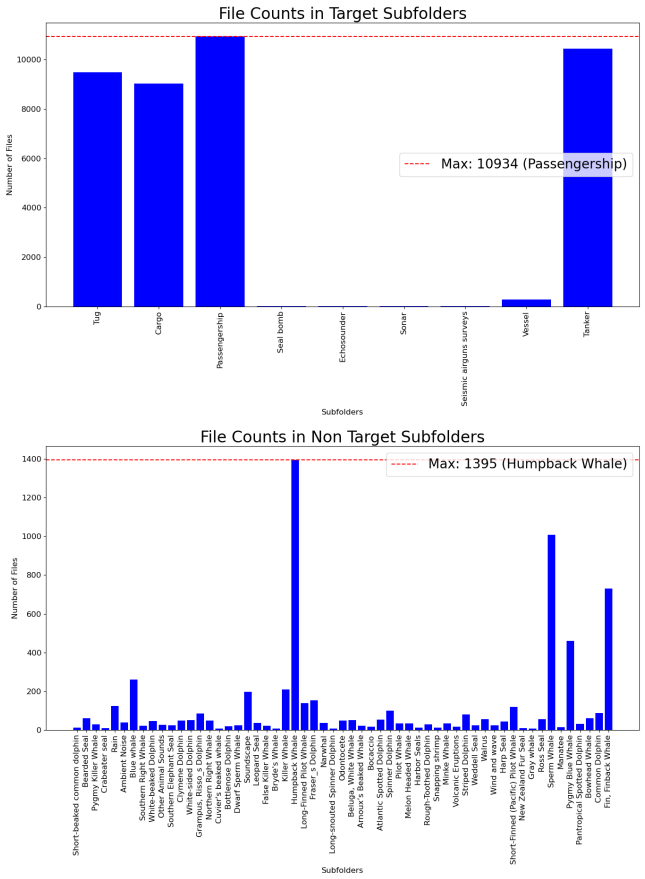

Figure 8: Distribuzione dei File nel Subset suddiviso rispetto i File Audio Preprocessati

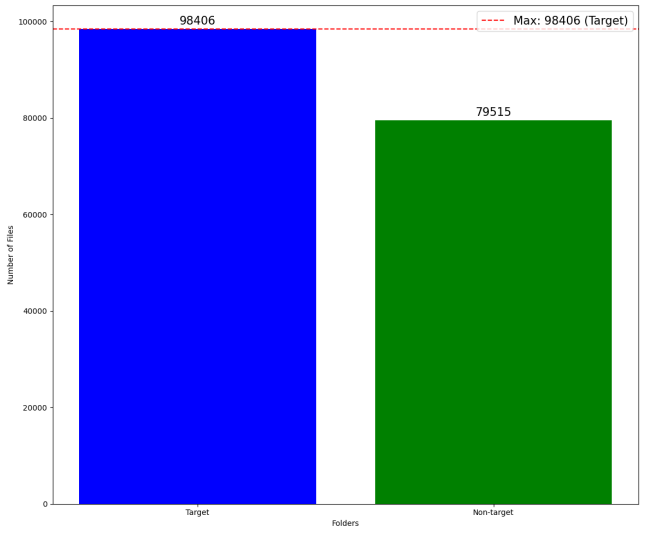

Figure 9: Distribuzione dei scalogrammi nelle Due Categorie Principali

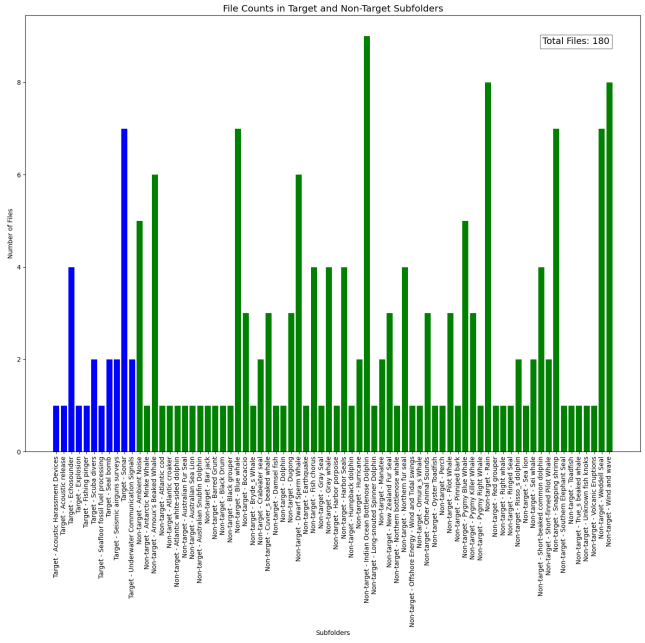

Figure 10: Distribuzione dei File Audio Non Preprocessati nelle Sottocategorie Minoritarie

che è stata identificata come quella con il maggior numero di campioni, totalizzando 1.454 scalogrammi come mostrato nella Figura 8. Le altre sottocategorie sono state successivamente equiparate al medesimo numero di campioni utilizzando tecniche di Data Augmentation descritte nel paragrafo 5.1, portando così il totale a 42.166 file per la categoria considerata.

Successivamente, tutti i scalogrammi sono stati aggregati nelle categorie "Target" e "Non Target", eliminando tutte le sottocategorie.

Nella Figura 12 è possibile vedere la distribuzione dei scalogrammi all'interno delle due categorie principali.

Per preservare le prestazioni dei modelli, è stata presa la decisione di uniformare la categoria "Non Target" alla categoria "Target", applicando le stesse tecniche di Data Augmentation ai scalogrammi originali.

Questo approccio ha assicurato un equo numero di campioni per entrambe le categorie, migliorando l'equilibrio del dataset e ottimizzando le performance dei modelli addestrati su di esso.

5.1 Data Augmentation

Per migliorare la capacità dei modelli di adattarsi a una vasta gamma di condizioni reali, sono state selezionate accuratamente una serie di trasformazioni durante il processo di Data Augmentation.

Oltre al flip orizzontale eseguito con una probabilità del 50%, che introduce variazioni nell'orientamento delle immagini, sono stati introdotti mascheramenti temporali e frequenziali con la stessa probabilità. Questi mascheramenti, eseguiti con un colore blu scuro, simulano variazioni nel contenuto delle immagini, aggiungendo complessità al dataset.

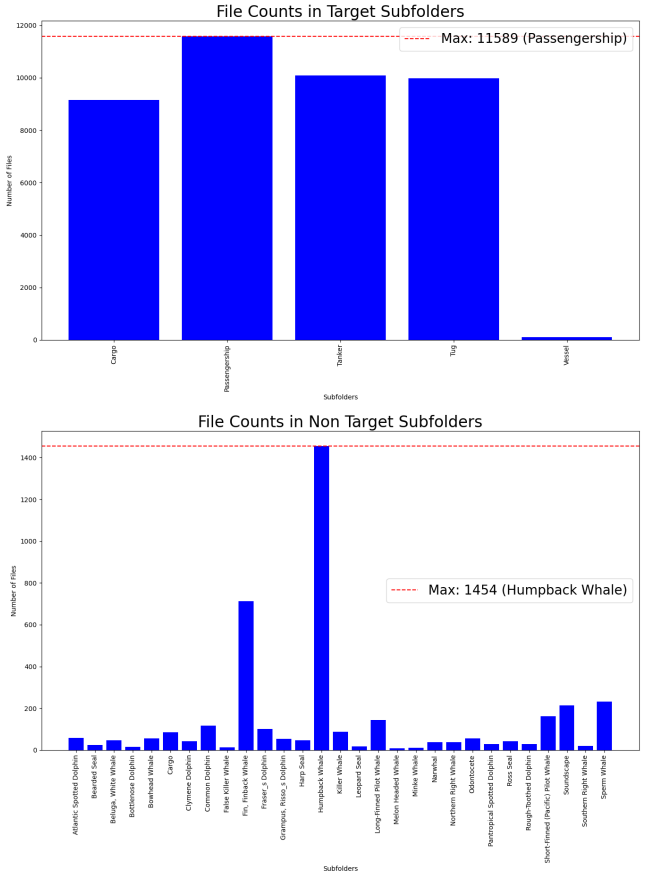

Figure 11: Distribuzione dei scalogrammi nel Subset suddiviso rispetto i File Audio Non Preprocessati

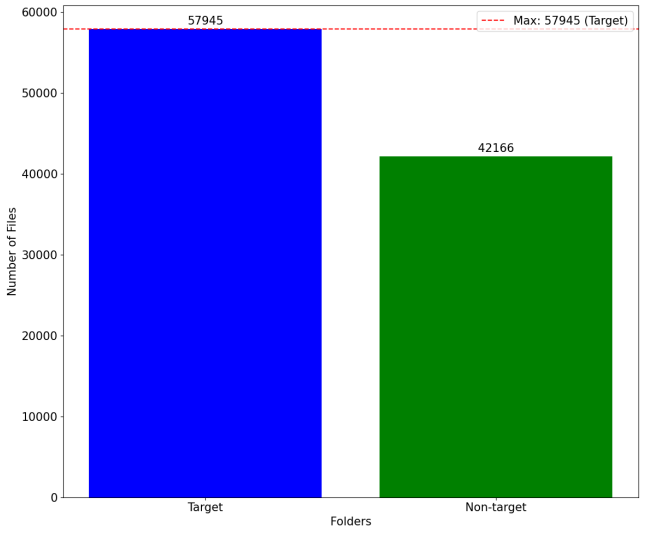

Figure 12: Distribuzione dei scalogrammi nelle Due Categorie Principali

Inoltre, sono stati inseriti livelli casuali di rumore compresi tra 5 e 15, con una probabilità del 50%, per introdurre variazioni casuali e migliorare la capacità di generalizzazione del modello.

In aggiunta, è stata adottata la tecnica del Time Shift per simulare possibili discrepanze nei tempi presenti nei dati reali. Questa tecnica introduce una variazione temporale casuale nelle immagini, generando un valore casuale rappresentante lo shift temporale e applicandolo agli array di pixel delle immagini. Questo effetto di "spostamento" simula variazioni nei momenti temporali rappresentati dalle immagini.

Per gestire le operazioni randomiche nelle diverse trasformazioni, è stato utilizzato un seed. Impostare un seed consente di ottenere risultati randomici coerenti ad ogni esecuzione del processo, garantendo la riproducibilità e facilitando il tracciamento delle modifiche apportate al dataset.

Un altro aspetto fondamentale è stato il mantenimento delle dimensioni originali delle immagini dopo le trasformazioni. Questo ha consentito di evitare qualsiasi perdita di risoluzione che potrebbe compromettere la qualità complessiva dei dati e di assicurare che i modelli siano esposti a una varietà di condizioni realistiche durante l'addestramento.

6 MODELLI DI MACHINE LEARNING

6.1 AlexNet

AlexNet è una rete neurale profonda con 60 milioni di parametri e 650.000 neuroni che migliora l'apprendimento aumentando la profondità e applicando strategie di ottimizzazione multiparametriche. Per superare il problema del vanishing gradient nelle reti profonde, è stata introdotto la funzione di attivazione non lineare chiamata "ReLU" e definita come

$$ReLU(x) = max(x, 0)$$

Inoltre, per prevenire l'overfitting e migliorare la generalizzazione, sono stati utilizzati strati di dropout dopo gli strati fully connected. La convoluzione in AlexNet permette di apprendere automaticamente le caratteristiche dalle immagini di addestramento, riducendo la complessità della rete attraverso la condivisione dei parametri. La convoluzione è così definita:

$$C(m, n) = (M * w)(m, n) = \sum_k \sum_l M(m - k, n - l)w(k, l)$$

Il pooling viene impiegato per ridurre le dimensioni delle caratteristiche, aggregando pixel vicini nella feature map. La normalizzazione locale tramite cross-channel è ispirata ai neuroni reali, normalizzando le feature map prima di passarle allo strato successivo. Nei layer fully connected, la funzione di attivazione softmax vincola l'output dei neuroni nell'intervallo (0,1) [12].

6.2 EfficientNet

EfficientNet-B3 è una rete che si basa sul concetto di "compound scaling". Questo approccio affronta il compromesso tra dimensione del modello, accuratezza ed efficienza computazionale, scalando tre dimensioni essenziali della rete neurale: larghezza, profondità e risoluzione. La larghezza riguarda il numero di canali in ogni strato della rete neurale, la profondità si riferisce al numero totale di strati nella rete e la risoluzione che implica l'adattamento della dimensione dell'immagine in ingresso [26].

6.3 ResNet

L'architettura ResNet-50 è una variante della Residual Network (ResNet), il modello è caratterizzato da una struttura di apprendimento residuo profondo, che introduce connessioni di salto per creare scorciatoie tra i livelli, consentendo alla rete di apprendere funzioni residue invece di mappature dirette. Le connessioni di scelta rapida saltano uno o più livelli, collegando direttamente i livelli precedenti a quelli successivi. Queste scorciatoie consentono al modello di preservare informazioni cruciali dagli strati iniziali, mitigando il problema del gradiente di scomparsa durante la propagazione all'indietro [27].

La Tabella 9 fornisce un confronto dettagliato tra i tre modelli.

Efficienza: EfficientNet B3 offre un'ottima efficienza in termini di parametri e FLOPs rispetto all'accuratezza ottenuta.

Profondità: ResNet permette di costruire reti molto profonde senza il problema della degradazione del gradiente, grazie ai blocchi residui.

Modello	Parametri	FLOPs
AlexNet	~ 60 milioni	~ 1.1 miliardi
EfficientNet B3	~ 12 milioni	~ 1.8 miliardi
ResNet-50	~ 25 milioni	~ 3.8 miliardi

Table 9: Confronto tra AlexNet, EfficientNet B3 e ResNet-50

6.4 Transfer learning

Con l'apprendimento per trasferimento, i modelli di addestramento preaddestrati vengono utilizzati come base mantenendo i parametri su ImageNet per una rapida convergenza dell'addestramento. Solo una parte del modello viene sostituita e addestrata su un nuovo compito, richiedendo un set di dati relativamente piccolo per il fine-tuning [12].

AlexNet: In AlexNet, il transfer learning consente di sfruttare i parametri preaddestrati su ImageNet, una vasta collezione di immagini etichettate, come base per un nuovo compito. Questo approccio accelera notevolmente la fase di addestramento poiché molte delle caratteristiche di base delle immagini sono già state apprese. Ad esempio, gli strati convoluzionali possono essere riutilizzati per estrarre le caratteristiche mentre solo gli ultimi strati fully connected vengono addestrati nuovamente sul nuovo dataset.

EfficientNet-B3: Anche EfficientNet-B3 beneficia del transfer learning utilizzando pesi preaddestrati su ImageNet.

ResNet-50: ResNet-50 utilizza il transfer learning in modo simile, partendo da una versione preaddestrata del modello su ImageNet. Grazie alla struttura a connessioni di salto, ResNet può preservare e trasferire le informazioni apprese in strati precedenti.

6.5 Metriche per la valutazione delle prestazioni dei modelli

Nel campo del deep learning, si utilizza generalmente la matrice di confusione per visualizzare le prestazioni del modello. Dalla

matrice di confusione è possibile calcolare le metriche comuni utilizzate per valutare le prestazioni del modello [12]. Pertanto, in questo studio sono stati utilizzati parametri comuni, tra cui accuracy, precision, recall e F1-score, per valutare le prestazioni dei modelli di classificazione.

7 CLASSIFICAZIONE BINARIA

Per la classificazione dei suoni in categorie "Target" e "Non Target" sono stati condotti tre esperimenti utilizzando due diverse architetture di reti neurali: AlexNet ed EfficientNet-B3. Ogni CNN è stata addestrata con il transfer learning, sostituendo l'ultimo strato della rete preaddestrata con un nuovo livello adattato al compito specifico, sfruttando così le conoscenze preesistenti del modello e migliorandone le prestazioni.

In tutti gli esperimenti, è stato applicato un doppio bilanciamento del dataset di allenamento, come descritto nel paragrafo 5. Per evitare di ottenere le stesse trasformazioni in entrambe le fasi di bilanciamento, il seed è stato impostato a 10 per il primo bilanciamento e a 1.017 per il secondo.

Per l'addestramento dei modelli è stato impiegato l'algoritmo Adaptive Moment Estimation (ADAM) con un tasso di apprendimento iniziale di 0,001 e un massimo di 50 epoche per addestramento. È stato applicato un meccanismo di early stopping per interrompere il processo di addestramento quando le metriche di performance iniziavano a peggiorare, evitando così l'overfitting. Inoltre, sono stati utilizzati checkpoint per il salvataggio dei parametri a ogni epoca, mantenendo la tracciabilità dei risultati. È stato utilizzato un batch size di 128 per garantire un equilibrio tra la velocità di addestramento e la stabilità del modello.

Nel primo esperimento, è stato utilizzato il subset creato secondo i file audio preprocessati e il modello di classificazione binaria basato su AlexNet. Nel secondo esperimento, è stata impiegata l'architettura EfficientNet-B3 sullo stesso subset. Negli ultimi due esperimenti, è stato utilizzato il subset generato utilizzando i file audio non preprocessati e sono state impiegate entrambe le architetture singolarmente, AlexNet ed EfficientNet-B3.

8 CLASSIFICAZIONE MULTICLASSE

Per la classificazione multiclasse, è stato utilizzato il subset generato in base ai file audio preprocessati, dove la categoria "Non Target" è stata eliminata.

Sono stati condotti quattro esperimenti utilizzando le architetture AlexNet e ResNet, tutti sullo stesso subset. Nel primo esperimento, è stata utilizzata l'architettura AlexNet senza meccanismo di early stopping. Tuttavia, si è osservato un significativo overfitting, che ha portato a prestazioni subottimali del modello.

Per affrontare questo problema, nel secondo esperimento è stato introdotto l'early stopping durante l'addestramento di ResNet-50. Questa modifica ha permesso di interrompere il processo di addestramento quando le metriche di performance iniziavano a peggiorare, migliorando così la generalizzazione del modello e prevenendo l'overfitting osservato nel primo esperimento.

Il terzo esperimento ha utilizzato l'architettura ResNet-50, combinata con il meccanismo di early stopping e una modifica alla tecnica di Data Augmentation. In particolare, sono state eliminate le maschere di frequenza e tempo, al fine di valutare l'impatto di questa variazione sulla performance del modello. L'early stopping ha continuato a prevenire l'overfitting, garantendo un addestramento più efficace.

Nel quarto esperimento, l'architettura ResNet-50 è stata nuovamente utilizzata con early stopping, ma il dataset è stato volutamente lasciato non bilanciato per analizzare l'effetto del bilanciamento dei dati sulle prestazioni del modello. Questo esperimento ha permesso di osservare come la distribuzione sbilanciata delle classi influenzi la capacità del modello di generalizzare correttamente.

Per l'addestramento dei modelli, è stato impiegato l'algoritmo Adaptive Moment Estimation (ADAM) con un tasso di apprendimento iniziale di 0,001 e un massimo di 50 epoche. Sono stati utilizzati checkpoint per salvare i parametri ad ogni epoca, garantendo la tracciabilità dei risultati. Anche in questo caso, è stato adottato un batch size di 128.

9 RISULTATI OTTENUTI

I risultati ottenuti dagli esperimenti descritti nel paragrafo 5 hanno evidenziato diverse dinamiche nel processo di classificazione sia binaria che multiclasse.

Nel contesto della classificazione binaria, l'impiego di due architetture di reti neurali differenti, AlexNet ed EfficientNet-B3, ha permesso di confrontare le performance dei modelli su un compito specifico.

I risultati ottenuti, come evidenziato nella Tabella 10, mostrano che nel primo esperimento con AlexNet, il modello si è fermato dopo 10 epoche, raggiungendo un'accuracy di 0.92, una precisione di 0.94, un recall di 0.97 e un F1-score di 0.95.

Nel secondo esperimento con EfficientNet-B3, il modello si è fermato dopo 8 epoche, ottenendo un'accuracy di 0.91, una precisione di 0.93, un recall di 0.96 e un F1-score di 0.95.

Nel terzo esperimento con AlexNet, il modello si è fermato dopo 7 epoche, con un'accuracy di 0.88, una precisione di 0.89, un recall di 0.96 e un F1-score di 0.92.

Infine, nel quarto esperimento con EfficientNet-B3, il modello si è fermato dopo 27 epoche, ottenendo un'accuracy e una precisione di 0.88, un recall di 0.97 e un F1-score di 0.93.

L'introduzione del meccanismo di early stopping ha chiaramente beneficiato entrambi i modelli, riducendo il rischio di overfitting e migliorando la generalizzazione. Nonostante AlexNet tenda a mostrare una leggera superiorità in termini di accuracy, precision e recall nei test iniziali, EfficientNet-B3 ha comunque dimostrato di mantenere un buon equilibrio complessivo delle metriche essendo anche computazionalmente meno oneroso. Questi risultati evidenziano la robustezza e l'efficacia dei modelli CNN utilizzati per l'analisi dei scalogrammi.

Nel caso della classificazione multiclasse, l'esperimento ha coinvolto l'utilizzo di due architetture diverse, AlexNet e ResNet, su un dataset generato in base ai sottoscalogrammi. Dalla Tabella 11 è possibile osservare come nel primo esperimento con AlexNet, il modello si sia fermato dopo 50 epoche, raggiungendo un'accuracy di 0.40, una precisione di 0.42, un recall di 0.40 e un F1-score di 0.39.

Nel secondo esperimento con ResNet, il modello ha terminato l'addestramento dopo 13 epoche, ottenendo un'accuracy di 0.45, una precisione di 0.46, un recall di 0.45 e un F1-score di 0.44.

	Epoche	Accuracy	Precision	Recall	F1-score
Test 1 (AlexNet)	10	**0.92**	**0.94**	**0.97**	**0.95**
Test 2 (EfficientNet-B3)	8	0.91	0.93	0.96	**0.95**
Test 3 (AlexNet)	7	0.88	0.89	0.96	0.92
Test 4 (EfficientNet-B3)	27	0.88	0.88	**0.97**	0.93

Table 10: Risultati ottenuti dalla classificazione binaria

	Epoche	Accuracy	Precision	Recall	F1-score
Test 1 (AlexNet)	50	0.40	0.42	0.40	0.39
Test 2 (ResNet)	13	**0.45**	0.46	**0.45**	**0.44**
Test 3 (ResNet)	13	0.44	**0.47**	0.44	**0.44**
Test 4 (ResNet)	12	0.43	0.43	0.43	0.42

Table 11: Risultati ottenuti dalla classificazione multiclasse

Nel terzo esperimento con ResNet, il modello si è fermato dopo 13 epoche, con un'accuracy di 0.44, una precisione di 0.47, un recall di 0.44 e un F1-score di 0.44.

Infine, nel quarto esperimento con ResNet, il modello si è fermato dopo 12 epoche, ottenendo un'accuracy di 0.43, una precisione di 0.43, un recall di 0.43 e un F1-score di 0.42.

Dai risultati si evince che ResNet ha superato AlexNet, anche se le prestazioni complessive rimangono inferiori rispetto alla classificazione binaria. Questo può essere attribuito alla complessità maggiore del compito multiclasse, che richiede al modello di distinguere tra un numero maggiore di sottocategorie.

Anche in questo caso l'early stopping ha dimostrato di essere una tecnica efficace per prevenire l'overfitting, come mostrato dal confronto tra il primo e il secondo esperimento nella classificazione multiclasse.

Tuttavia, la modifica alla tecnica di Data Augmentation nel terzo esperimento non ha portato alcun miglioramento significativo, suggerendo che le maschere di frequenza e tempo potrebbero non essere critiche per le prestazioni del modello.

Infine, il quarto esperimento nella classificazione multiclasse ha evidenziato l'importanza del bilanciamento del dataset. L'analisi dei risultati indica che un dataset sbilanciato può influire negativamente sulla capacità del modello di generalizzare correttamente, riducendo le metriche di performance.

10 CONCLUSIONI

In conclusione, l'analisi condotta attraverso esperimenti di classificazione binaria e multiclasse ha fornito importanti insight sulle prestazioni di diverse architetture di reti neurali e sull'impatto di tecniche come l'early stopping e il bilanciamento del dataset. I risultati evidenziano che, nel contesto della classificazione binaria, AlexNet ha mostrato prestazioni leggermente superiori rispetto ad EfficientNet-B3, con un vantaggio in termini di accuratezza, precisione e richiamo nei test iniziali. Tuttavia, EfficientNet-B3 ha mantenuto un buon equilibrio complessivo delle metriche ed è risultato meno oneroso dal punto di vista computazionale.

Nella classificazione multiclasse, ResNet ha dimostrato di superare AlexNet, sebbene le prestazioni complessive siano rimaste inferiori rispetto alla classificazione binaria a causa della maggiore complessità del compito. È emerso anche che l'early stopping si è rivelato efficace nel mitigare l'overfitting, mentre le modifiche alla tecnica di Data Augmentation hanno avuto un impatto meno significativo.

Inoltre, il bilanciamento del dataset si è dimostrato cruciale per garantire una corretta generalizzazione del modello. Pertanto un dataset bilanciato, con un numero equo di campioni per ciascuna classe, aiuta il modello ad apprendere in modo più efficace, evitando di favorire una sottocategoria rispetto ad un'altra.

Questi risultati sottolineano l'importanza di valutare attentamente le specifiche esigenze del problema e di adottare approcci ad hoc per massimizzare le performance del modello di machine learning, sfruttando le potenzialità delle diverse architetture e tecniche disponibili.

11 SVILUPPI FUTURI

Gli sviluppi futuri dovranno concentrarsi sull'ottimizzazione dei parametri modificati durante il preprocessing, al fine di migliorare ulteriormente le prestazioni dei modelli di classificazione. In particolare, sarà importante esplorare diverse tecniche di ricampionamento.

Sarà cruciale sperimentare varie strategie di segmentazione degli audio stessi per garantire che le caratteristiche rilevanti siano adeguatamente catturate e utilizzate dai modelli.

Inoltre, sarà necessario estendere il dataset, poiché i dati presenti nelle sottocategorie "Target" e "Non Target" non erano sufficienti per ottenere buone performance nella classificazione multiclasse. Un dataset più ampio e diversificato potrebbe fornire una base migliore per l'addestramento dei modelli, migliorando la loro accuratezza nelle classificazioni più complesse. Questi aggiustamenti mirati potrebbero fornire significativi benefici in termini di accuratezza, precisione, recall e F1-score, portando a una generalizzazione ancora migliore dei modelli sviluppati.

REFERENCES

[1] "At-risk marine biodiversity faces extensive, expanding, and intensifying human impacts," https://www.researchgate.net/publication/350569596_At-risk_marine_biodiversity_faces_extensive_expanding_and_intensifying_human_impacts.

[2] G. A. Petrossian, "Preventing illegal, unreported and unregulated (iuu) fishing: A situational approach," *Biological Conservation*, vol. 189, pp. 39–48, 2015,

detecting and Understanding Non-compliance with Conservation Rules. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006320714003140

[3] "La biodiversità marina e la preservazione degli habitat oceanici," https://iquadro.energy/la-biodiversita-marina-e-la-preservazione-degli-habitat-oceanici/#:~:text=L'importanza%20della%20biodiversitÃă%20marina&text=Lo%20stesso%20discorso%20vale%20anche,in%20crisi%20milioni%20di%20persone.

[4] A. K. I. H. Z. L. M. C. M. T. S.-U. R. S. N. N. E. A. M. Ali, "Transfer learning for efficient classification of grouper sound," *J. Acoust. Soc. Am.*, 2020.

[5] Z. X. L. C. X. L. L. Y. Jin Kangkang, Xu Jian, "An acoustic tracking model based on deep learning using two hydrophones and its reverberation transfer hypothesis, applied to whale tracking," *Frontiers in Marine Science*, vol. 10, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmars.2023.1182653

[6] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, "Marine mammal species classification using convolutional neural networks and a novel acoustic representation," in *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds. Cham: Springer International Publishing, 2020, pp. 290–305.

[7] W. Cai, J. Zhu, M. Zhang, and Y. Yang, "A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation," *Sensors*, vol. 22, no. 19, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/19/7443

[8] A. R. J, C. Sathish Kumar, and P. J. Anu Jose, "A wavelet based time-frequency descriptor for automatic classification of acoustic signals of fishes," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, vol. 1, 2019, pp. 1059–1063.

[9] J. J. Noda, C. M. Travieso, and D. Sánchez-Rodríguez, "Automatic taxonomic classification of fish based on their acoustic signals," *Applied Sciences*, vol. 6, no. 12, 2016. [Online]. Available: https://www.mdpi.com/2076-3417/6/12/443

[10] S. Liu, M. Liu, M. Wang, T. Ma, and X. Qing, "Classification of cetacean whistles based on convolutional neural network," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2018, pp. 1–5.

[11] J.-F. Laplante, M. A. Akhloufi, and C. Gervaise, "Deep learning for marine bioacoustics and fish classification using underwater sounds," in *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2022, pp. 288–293.

[12] T. Lu, B. Han, and F. Yu, "Detection and classification of marine mammal sounds using alexnet with transfer learning," *Ecological Informatics*, vol. 62, p. 101277, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121000686

[13] D. Duan, L. gang Lü, Y. Jiang, Z. Liu, C. Yang, J. Guo, and X. Wang, "Real-time identification of marine mammal calls based on convolutional neural networks," *Applied Acoustics*, vol. 192, p. 108755, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X22001293

[14] A. K. Ibrahim, H. Zhuang, L. M. Chérubin, M. T. Schärer-Umpierre, R. S. Nemeth, N. Erdol, and A. M. Ali, "Transfer learning for efficient classification of grouper sound," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. EL260–EL266, 09 2020. [Online]. Available: https://doi.org/10.1121/10.0001943

[15] O. Rioul and M. Vetterli, "Wavelets and signal processing," 1991.

[16] "Watkins marine mammal sound database," https://whoicf2.whoi.edu/science/B/whalesounds/about.cfm.

[17] "A collection of sounds from the sea," https://oceanexplorer.noaa.gov/explorations/sound01/background/seasounds/seasounds.html.

[18] "Sounds recorded in glacier bay," https://home.nps.gov/glba/learn/nature/soundclips.htm.

[19] "Avisoft," https://avisoft.com.

[20] "Marine mammal bioacoustics of australia and antarctica," http://cmst.curtin.edu.au/research/marine-mammal-bioacoustics/.

[21] "Ocean conservation research," https://ocr.org.

[22] "Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification," https://www.sciencedirect.com/science/article/pii/S0957417421007016.

[23] "Nyquist–shannon sampling theorem," https://en.wikipedia.org/wiki/NyquistâĂŞShannon_sampling_theorem.

[24] "Nozioni di base sull'audio digitale: frequenza di campionamento audio e profondità di bit," https://www.izotope.com/en/learn/digital-audio-basics-sample-rate-and-bit-depth.html#:~:text=What%20is%20audio%20bit%20depth,recreate%20the%20original%20audio%20signal.

[25] "Audio bit depth," https://en.wikipedia.org/wiki/Audio_bit_depth#:~:text=In%20digital%20audio%20using%20pulse,the%20resolution%20of%20each%20sample.

[26] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[27] D. V. D. Kirandeep, Ramanpreet Kaur, "Image recognition using resnet50."