Underwater Classification

Mario Lezzi

A B S T R A C T

Il seguente studio propone un approccio basato sul machine learning per distinguere tra suoni bioacustici marini e quelli di origine antropogenica legati al mondo marino, con particolare attenzione all'ottimizzazione della classificazione e all'identificazione dettagliata delle sottoclassi antropogeniche. L'obiettivo è contribuire al monitoraggio degli ecosistemi marini e mitigare l'impatto delle attività umane attraverso un sistema automatico di analisi dei segnali acustici subacquei. La ricerca si basa su un dataset preesistente di registrazioni marine. Dopo una fase iniziale di pre-processing dei dati, sono state estratte caratteristiche audio rilevanti e sono stati addestrati modelli di machine learning. L'approccio si articola in due fasi principali: una prima fase di classificazione binaria, in cui il modello distingue tra suoni antropogenici e bioacustici naturali, e una seconda fase di classificazione multiclasse, in cui i suoni antropogenici vengono ulteriormente suddivisi in sottoclassi specifiche per identificarne l'origine. Gli esperimenti hanno incluso il test di diversi modelli, l'analisi di molteplici caratteristiche e l'esplorazione di frequenze di campionamento variabili per ottimizzare le prestazioni. I risultati ottenuti evidenziano il potenziale di questi approcci per migliorare la gestione acustica degli ambienti marini, supportando strategie di conservazione più efficaci.

Underwater Classification

1. Introduzione

Con l'aumento della raccolta di dati audio, le sfide legate alla gestione e all'analisi di grandi volumi di suoni sono diventate sempre più complesse, soprattutto nel contesto marino. L'oceano è un ambiente ricco di suoni naturali e antropici, e le tecnologie moderne di registrazione subacquea hanno facilitato l'acquisizione di dati acustici per scopi scientifici, di conservazione e monitoraggio delle attività umane.
Un aspetto cruciale è la capacità di distinguere tra suoni generati da attività umane (Target) e quelli prodotti dalla fauna marina (Non Target). Tale distinzione è fondamentale per proteggere gli ecosistemi marini, promuovere la ricerca scientifica e sviluppare soluzioni innovative per il monitoraggio ambientale. L'impiego di tecniche di machine learning per l'analisi dei segnali acustici offre un'opportunità fondamentale per gestire e comprendere meglio gli ecosistemi marini.
Gli obiettivi principali di questo studio sono i seguenti:

1. **Analizzare la letteratura esistente** sul riconoscimento dei segnali acustici per identificare le caratteristiche fondamentali dei segnali audio in questo contesto. Questo consentirà di creare un template delle proprietà essenziali per distinguere i suoni prodotti dalla fauna marina da quelli antropogenici.
2. **Estrarre le caratteristiche audio** dai segnali acustici, prendendo spunto da studi precedenti in letteratura. Questo passaggio permetterà di identificare e selezionare le feature più rilevanti per il riconoscimento dei suoni, ottimizzando l'efficacia del modello di classificazione.
3. **Sviluppare modelli di classificazione** capaci di separare efficacemente i suoni bioacustici marini, distinguendo i segnali biologici da quelli di origine antropica. Successivamente, si punterà a migliorare i modelli per riconoscere con precisione non solo i suoni interclasse, ma anche intraclasse.
4. **Condurre esperimenti con diversi sample rates** per analizzare l'impatto delle variazioni nella frequenza di campionamento sulle performance dei modelli. Questo esperimento fornirà indicazioni utili per ottimizzare l'analisi acustica, tenendo conto delle diverse risoluzioni temporali nei segnali audio acquisiti.

Questi obiettivi rappresentano un passo significativo verso lo sviluppo di strumenti efficaci per l'analisi dei segnali acustici marini, con ricadute importanti sia per la conservazione ambientale che per il progresso della ricerca scientifica.

2. Stato dell'arte

L'analisi bioacustica è un campo in rapida espansione sostenuto dall'integrazione di nuove tecnologie,come l'intelligenza artificiale, che sta rivoluzionando ogni aspetto dell'analisi dei dati. Prima dell'avvento dell'IA, l'analisi bioacustica veniva eseguita mediante una combinazione di tecniche manuali e strumenti acustici tradizionali. Questi metodi erano spesso laboriosi e richiedevano un intervento umano significativo per la raccolta, l'analisi e l'interpretazione dei dati sonori. È evidente che, prima dell'introduzione delle nuove tecnologie, ogni dato bioacustico doveva essere analizzato singolarmente, trattato come caso di studio e sottoposto all'esame di esperti del settore, rallentando notevolmente le fasi di analisi e classificazione dei segnali.(6)(2)(Gibb et al., 2019; Blumstein et al., 2011). Nel contesto dell'analisi e della classificazione, i passaggi fondamentali possono essere così riassunti in:
Raccolta dati: Questa fase veniva realizzata mediante la registrazione degli audio e il campionamento manuale, richiedendo la presenza fisica degli esperti per distinguere accuratamente i segnali bioacustici da raccogliere.
Pre-processing dei dati : Durante questa fase, si procedeva al filtraggio e alla pulizia delle fonti per ridurre il rumore di fondo e migliorare la qualità del suono registrato. La segmentazione veniva eseguita manualmente per

selezionare solo gli elementi rilevanti per il contesto di applicazione.
Analisi del suono: I segnali raccolti venivano analizzati uno per uno, analizzando la frequenza e effettuando misurazioni manuali.
Classificazione e identificazione : In questa fase, venivano utilizzate chiavi dicotomiche per identificare e distinguere le diverse specie animali basandosi sulle vocalizzazioni, che presentano caratteristiche specifiche per ciascuna specie. Inoltre, veniva effettuato un confronto diretto con samples noti appartenenti a determinate specie.
Documentazione e archiviazione: Infine, si procedeva alla catalogazione, associando le registrazioni alle varie specie animali conosciute; in caso di assenza di corrispondenza, i segnali venivano classificati come componenti bioacustiche inconsistenti o sospettati di appartenere a sottospecie già esistenti.
Le limitazioni delle analisi pre-IA includevano il tempo impiegato e le ingenti risorse necessarie per ogni caso di studio. Inoltre, vi era una componente di soggettività e interpretazione personale da parte degli esperti, nonché una capacità limitata di elaborazione, poiché gli esperti di un determinato settore potevano lavorare solo su una parte delle ricerche alla volta.(6)(2) (Gibb et al., 2019; Blumstein et al., 2011).
Oggi, con l'introduzione dell'IA, sono state sviluppate diverse soluzioni per automatizzare le fasi precedentemente descritte. Nell'ambito dell'analisi dei segnali bioacustici, si utilizza un nuovo approccio basato sull'impiego dell'IA per automatizzare e analizzare empiricamente le fonti audio. In questo contesto, sono descritti diversi progetti su cui si basa il nostro lavoro di ricerca e sviluppo di un nuovo sistema in grado di eseguire una distinzione interclasse e intraclasse dei segnali bioacustici in ambiente sottomarino.(9) (Kahl et al., 2021).

2.1. Opere correlate

2.1.1. Automatic fish sounds classification

L'articolo (11) (Malfante et al 2018) presenta un avanzato metodo per il monitoraggio passivo della vitalità degli oceani, con un focus particolare sulle popolazioni di pesci. Lo studio sviluppa un modello discriminativo basato su tecniche di machine learning supervisionato, in particolare Random Forest (RF) e Support Vector Machines (SVM), per classificare i suoni dei pesci. Il modello si distingue per l'uso di caratteristiche estratte dai domini temporale, frequenziale e cepstrale dei segnali acustici. Testato su suoni reali registrati in diverse aree marine, il sistema ha raggiunto un'accuratezza di classificazione del 96,9%, superando significativamente i risultati ottenuti con metodi tradizionali. Un aspetto particolarmente rilevante dello studio è l'approccio dettagliato all'estrazione delle caratteristiche. Gli autori esplorano l'impatto delle caratteristiche provenienti dai diversi domini, dimostrando che sebbene ciascun dominio contenga informazioni discriminative utili, la combinazione delle caratteristiche estratte dai tre domini offre prestazioni superiori. Per ottimizzare la classificazione, è stato adottato un metodo di selezione delle caratteristiche basato sui pesi delle caratteristiche nel modello RF. Sono stati identificati due sottoinsiemi di caratteristiche: Most Valuable Features (MVF) e Valuable Features(VF). Il MVF comprende tre caratteristiche chiave: l'energia kurtosis dal dominio frequenziale (F28), la kurtosis media dal dominio temporale (T7) e il tasso di attraversamento della soglia dal dominio temporale (T15) e ha raggiunto un'accuratezza media del 91,5% (RF) e del 91,3% (SVM). Il set VF, che include il MVF e altre 16 caratteristiche aggiuntive, ha migliorato l'accuratezza globale al 95,6% (RF) e al 94,7% (SVM). Questi risultati suggeriscono che, sebbene l'uso di tutte le caratteristiche possa non essere necessario per ottenere risultati di alta qualità, la selezione mirata delle caratteristiche consente di mantenere l'efficacia del sistema di classificazione. Questo è particolarmente utile per applicazioni in tempo reale con risorse computazionali limitate. Inoltre, l'analisi evidenzia che l'importanza delle caratteristiche può variare a seconda della classe, sottolineando la necessità di una selezione accurata per ottimizzare la classificazione in base ai requisiti specifici delle diverse caratteristiche.(11)(Malfante et al. 2018) In sintesi, l'integrazione dell'intelligenza artificiale ha rivoluzionato l'analisi bioacustica, migliorando l'efficienza e la precisione nella classificazione dei segnali sonori.

2.1.2. BirdNet

BirdNet è un progetto di ricerca che impiega tecniche di deep learning, per l'identificazione delle specie di uccelli attraverso l'analisi dei loro canti e richiami. Sviluppato da un team internazionale che include esperti dell'Università di Cambridge e della Czech Academy of Sciences, BirdNet si propone di distinguere tra le varie specie di uccelli a partire dai suoni che emettono. Il modello di machine learning su cui si basa è stato addestrato utilizzando un ampio database di registrazioni acustiche provenienti da diverse specie di uccelli, permettendo così di ottenere una classificazione precisa e affidabile.
Il sistema è in grado di analizzare registrazioni audio contenenti suoni di uccelli e riconoscere le specie in base a caratteristiche acustiche distintive. L'accuratezza del modello può variare a seconda della qualità del segnale e del rumore di fondo, ma in generale, BirdNet è progettato per affrontare le sfide legate all'analisi di ambienti acustici complessi. Questa tecnologia ha numerose applicazioni in vari ambiti, come il monitoraggio della biodiversità, la ricerca comportamentale sugli uccelli, e le iniziative di conservazione, specialmente per le specie minacciate. Inoltre, grazie alla disponibilità di versioni online e di applicazioni mobile, BirdNet è accessibile a ricercatori, professionisti e appassionati di ornitologia, consentendo loro di caricare registrazioni audio per ottenere un'identificazione automatica delle specie.
In un contesto di ricerca più ampio, come quello delle acustiche marine o di altre analisi ambientali, approcci simili a quelli di BirdNet potrebbero essere impiegati per

distinguere tra suoni naturali e antropogenici, contribuendo così a migliorare la gestione e la conservazione degli ecosistemi.

2.1.3. A Survey on Audio Feature Extraction for Automatic Music Genre Classification

Il lavoro (19) (Dhamodaran et al 2023), offre una panoramica completa sulle tecniche di estrazione delle caratteristiche audio utilizzate nella classificazione automatica dei generi musicali, un compito fondamentale per applicazioni come il recupero di informazioni musicali e la gestione di grandi archivi digitali. Gli autori analizzano le principali caratteristiche audio che giocano un ruolo cruciale in questo contesto, suddividendole in temporali, spettrali e tonali. Tra queste, spiccano gli MFCC (Mel-Frequency Cepstral Coefficients), che rappresentano il segnale audio in termini di frequenze percepite dall'orecchio umano, il centroide spettrale, che misura la brillantezza di un suono, e i cromogrammi, utili per catturare informazioni armoniche e melodiche. Il paper discute anche il ruolo delle tecniche di preprocessing e trasformazione dei segnali, essenziali per garantire che i dati audio grezzi siano adeguati ai modelli di classificazione. Vengono inoltre esaminati diversi approcci di machine learning applicati alla classificazione dei generi, come SVM, KNN, reti neurali profonde e metodi ensemble come Random Forest, mettendo in luce l'importanza della selezione accurata delle caratteristiche per migliorare le performance. Gli autori evidenziano i principali dataset disponibili, come il noto GTZAN, e gli strumenti software utilizzati per l'estrazione delle caratteristiche, tra cui Librosa ed Essentia. Infine, il paper riflette sulle sfide aperte, come l'ambiguità nella definizione dei generi musicali e la mancanza di dataset rappresentativi delle variazioni culturali, e sottolinea la necessità di sviluppare metodi più robusti e generalizzabili. Questo lavoro rappresenta un contributo essenziale nello stato dell'arte, fornendo una guida dettagliata per i ricercatori interessati a migliorare la classificazione musicale attraverso tecniche avanzate di estrazione e analisi delle caratteristiche audio.

Progetti come BirdNET e l'approccio descritto da (11)Malfante et al. dimostrano che l'uso combinato di tecniche avanzate e selezione mirata delle caratteristiche consente risultati superiori nella rilevazione e classificazione dei suoni, sia in contesti terrestri che marini. In parallelo, il lavoro di (19)Dhamodaran et al., sebbene focalizzato sulla classificazione dei generi musicali, fornisce un'importante base teorica e pratica sull'estrazione di caratteristiche audio e sull'applicazione di algoritmi di apprendimento automatico. Il loro studio analizza tecniche come gli MFCC, il centroide spettrale e i cromogrammi, evidenziandone l'efficacia nella rappresentazione di dati audio complessi. Questi principi trovano applicazione non solo nella classificazione musicale, ma anche in ambiti come l'analisi bioacustica, dove la selezione accurata delle caratteristiche audio è cruciale per migliorare l'accuratezza dei modelli. L'integrazione delle metodologie presentate da Dhamodaran con approcci mirati come BirdNET evidenzia il potenziale dell'IA nel trasformare l'analisi e la classificazione dei suoni in contesti naturali, sfruttando tecniche adattabili a diverse tipologie di segnali acustici.

3. Struttura del Dataset

Il dataset utilizzato in questo studio, contenente 2663 samples, è suddiviso in due principali categorie: **Target** (16 sottoclassi) e **Non Target** (96 sottoclassi). La directory Target include segnali audio di origine antropogenica, mentre la directory Non Target contiene segnali bioacustici. Le registrazioni per la creazione di questo dataset sono state estratte da fonti disponibili online, con l'obiettivo di costruire un dataset composito da utilizzare come base per lo sviluppo del modello di intelligenza artificiale.
Essendo un dataset eterogeneo, le caratteristiche tecniche dei segnali audio variano considerevolmente: il campionamento presenta un range di frequenza che va da 600 Hz a 384.000 Hz, con segnali sia in formato mono che stereo. I file audio sono in diversi formati, tra cui `.wav`, `.mp3` e `.flac`. L'ampiezza del segnale è variabile, mentre per la profondità in bit (bit depth) abbiamo 650 samples a 32 bit float, 67 in MPEG Layer III, 11 in Microsoft ADPCM, 1884 in Signed 16 bit PCM, 23 in Signed 24 bit PCM, 6 in Signed 32 bit PCM e infine 17 in Unsigned 8 bit PCM. La durata dei segnali audio è anch'essa estremamente variabile, con registrazioni che durano da pochi secondi fino a oltre 30 minuti, con una mediana totale di 3.53 minuti. Le manipolazioni e gli interventi necessari per bilanciare e preparare i dati saranno descritti in dettaglio nelle sezioni successive.

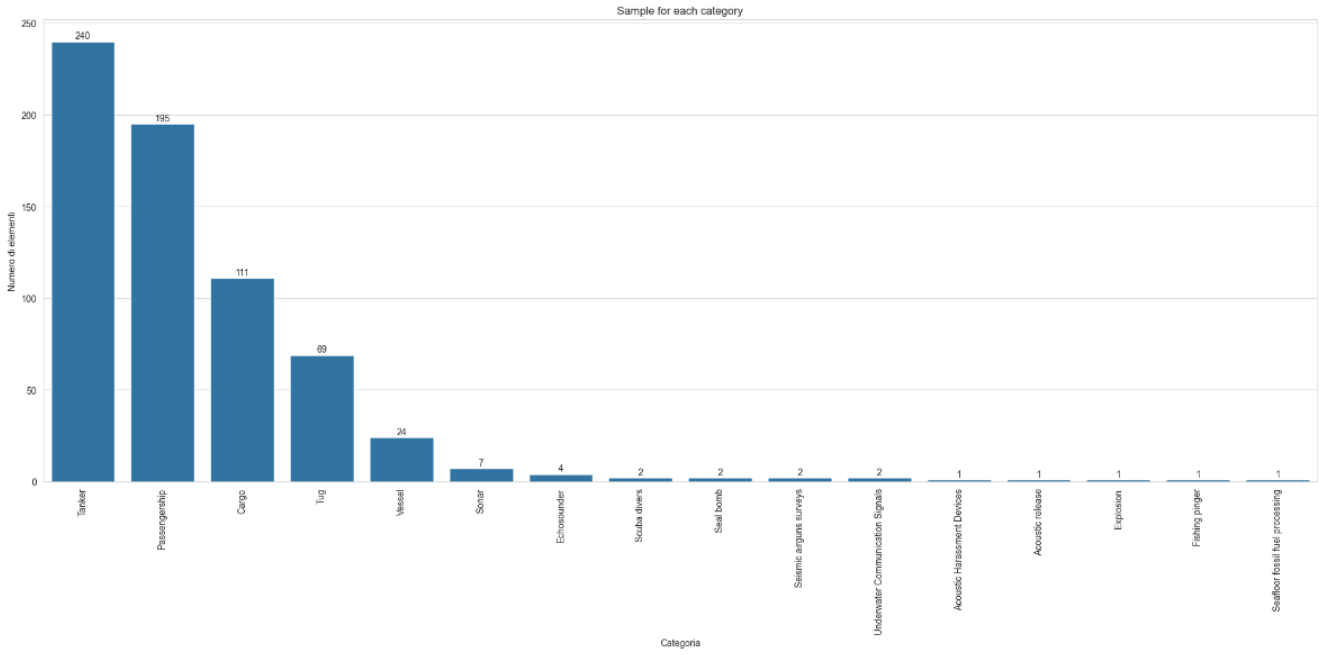

Figure 1: Samples in Target dopo rimozione duplicati

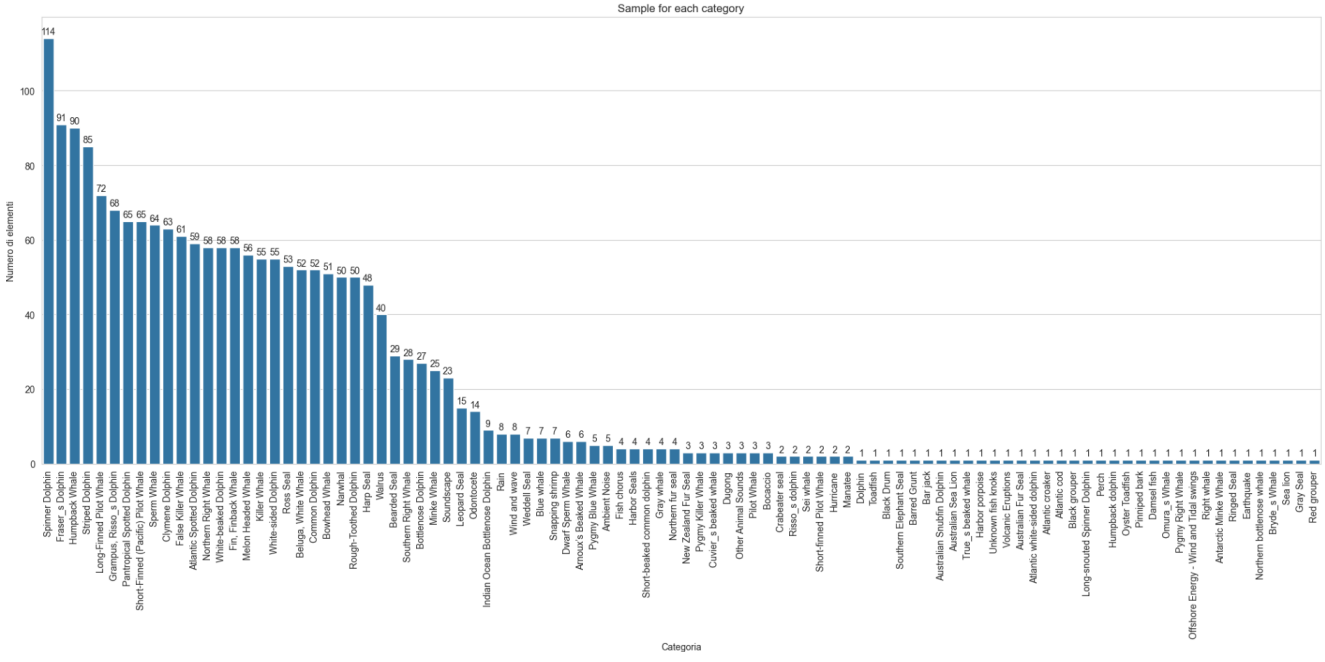

Figure 2: Samples in Non-Target dopo rimozione duplicati

4. Analisi dataset

Il primo passo dello studio ha riguardato un'accurata analisi preliminare del dataset. In questa fase è stata condotta una ricerca approfondita con la conseguente rimozione di 50 file duplicati, garantendo l'unicità di ciascun segnale audio e riducendo il numero di samples a 2613. Successivamente, per ogni sample, sono state calcolate le seguenti proprietà fondamentali:

- **Ampiezza** rappresenta l'intensità del segnale audio ed è un indicatore chiave delle dinamiche sonore presenti nelle registrazioni.L'ampiezza è stata calcolata come il valore massimo assoluto del segnale, che fornisce una misura della massima energia raggiunta nel file audio. Questa scelta permette di valutare e confrontare le registrazioni in termini di potenza sonora massima. Inoltre, l'uso del valore assoluto garantisce che l'analisi sia focalizzata sull'intensità complessiva del segnale, indipendentemente dalla polarità (positiva o negativa), semplificando così il confronto tra diversi segnali e assicurando una rappresentazione uniforme durante le fasi di pre-processing. [3]

- **Bit Depth**: indica la risoluzione dei dati audio, ossia la quantità di informazioni utilizzate per rappresentare ogni campione del segnale audio. Maggiore è il valore del bit depth, maggiore è la precisione con cui ogni campione viene registrato, consentendo una rappresentazione più dettagliata dell'intensità del suono. Un bit depth più alto permette di ridurre il rumore e aumenta la gamma dinamica del segnale, ma implica anche una maggiore dimensione del file. Ad esempio, un bit depth di 16 bit è standard per l'audio CD, mentre valori più alti, come 24 bit, sono comuni in ambito professionale per garantire una qualità superiore. [4]

- **Sample Rate**: Il sample rate (o frequenza di campionamento) rappresenta il numero di campioni acquisiti al secondo per registrare il segnale audio, ed è un parametro cruciale per definire la qualità temporale del segnale. Più alto è il sample rate, maggiore è la capacità di catturare dettagli nel dominio del tempo e di preservare la fedeltà del suono originale. Ad esempio, un sample rate di 44.1 kHz, utilizzato per i CD audio, significa che il segnale viene campionato 44.100 volte al secondo, mentre valori più alti, come 96 kHz o 192 kHz, sono utilizzati in applicazioni professionali per una qualità superiore. Tuttavia, l'uso di un sample rate più elevato aumenta anche la dimensione dei file audio, senza necessariamente migliorare la qualità percepibile. [5]

- **Durata**: La durata rappresenta la lunghezza temporale di ciascun segnale audio e indica quanto tempo dura una registrazione. Questo parametro è importante poiché la durata dei segnali audio può variare significativamente all'interno di un dataset. La gestione della durata è cruciale in fase di pre-processing, poiché segnali con durate molto diverse possono influenzare l'addestramento dei modelli. Ad esempio, segnali più lunghi potrebbero introdurre complessità aggiuntive nel processo di apprendimento, mentre segnali troppo corti potrebbero non contenere informazioni sufficienti. È possibile che la durata venga normalizzata o segmentata in modo che tutti i campioni abbiano una lunghezza temporale omogenea, migliorando la capacità del modello di generalizzare su dati di diverse lunghezze. [6]

- **Numero di canali**: si riferisce alla quantità di tracce audio presenti in un segnale, utilizzato per determinare se il file audio è in formato mono, stereo o multicanale. Un segnale mono (con un solo canale) rappresenta il suono in un'unica traccia, mentre un segnale stereo (con due canali) riproduce il suono su due tracce separate, consentendo di creare un effetto di spazializzazione sonora. Il numero di canali è un aspetto importante nel pre-processing, poiché i modelli di machine learning spesso richiedono un formato omogeneo per l'elaborazione. In molti casi, i segnali stereo possono essere convertiti in formato mono per semplificare l'elaborazione e ridurre la complessità computazionale senza perdere informazioni rilevanti. [7]

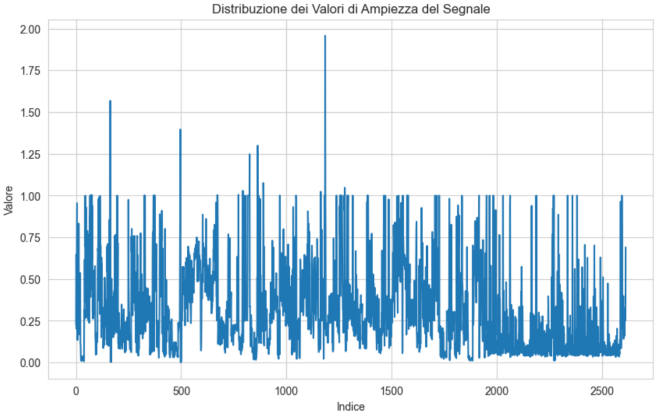

Figure 3: Ampiezza segnali

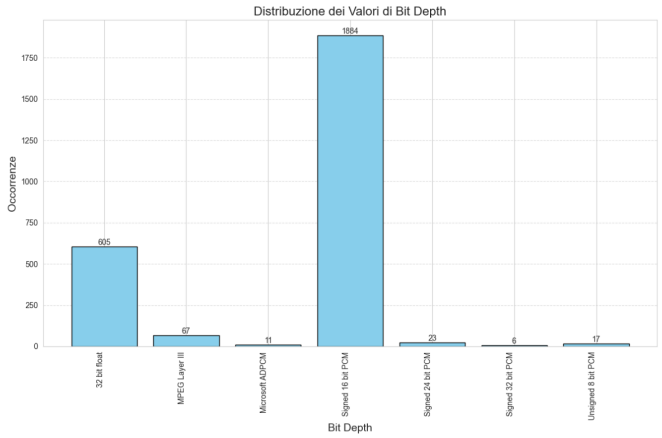

Figure 4: Distribuzione Valori Bit Depth in Target

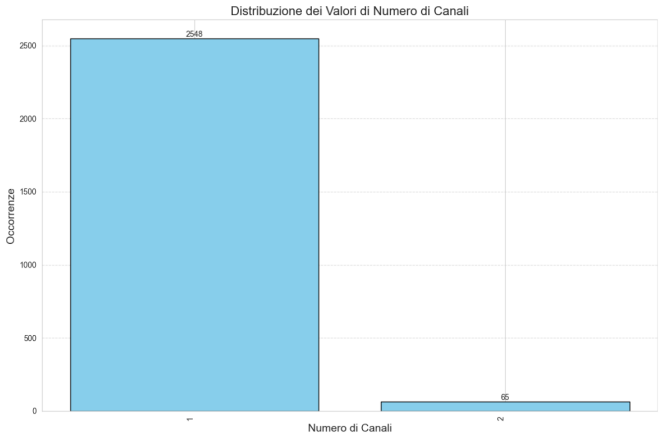

Figure 7: Distribuzione dei valori di Numero canali

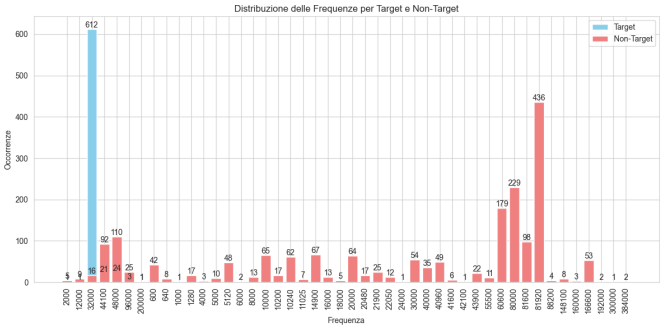

Figure 5: Distribuzione dei Sample Rate

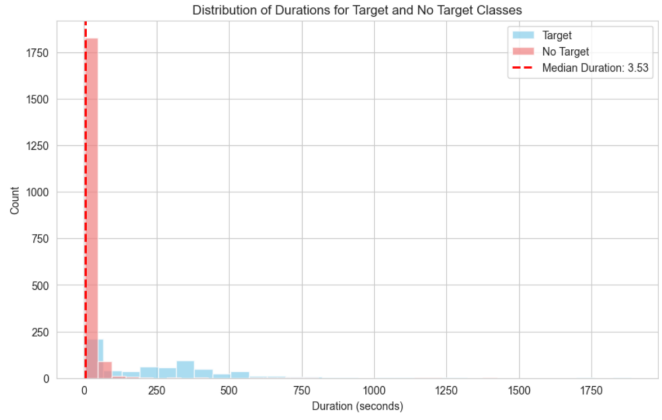

Figure 6: Distribuzione di Durata

Questi parametri hanno fornito una panoramica dettagliata delle caratteristiche del dataset e hanno costituito la base per le successive fasi di pre-elaborazione dei dati.

5. Pre-Processing

Dopo aver completato l'analisi preliminare, è stato eseguito il pre-processing dei samples per prepararli all'addestramento del modello. Questo passaggio è stato fondamentale per garantire che i dati fossero coerenti, comparabili e privi di anomalie che avrebbero potuto compromettere le prestazioni del modello. In primo luogo, i file audio, originariamente provenienti da diverse fonti e in formati variabili (come `.mp3`, `.flac` e `.wav`), sono stati convertiti tutti in formato `.wav`. Questa scelta è stata dettata dalla necessità di utilizzare un formato non compresso, capace di preservare la qualità del segnale audio e di evitare perdite di informazioni che avrebbero potuto influire negativamente durante l'elaborazione e l'addestramento.
In seguito, è stata effettuata una normalizzazione dei segnali, scalando tutti i campioni affinché il valore massimo assoluto dell'ampiezza fosse uguale a uno. [8] Questo processo ha avuto un duplice obiettivo: ridurre le variazioni estreme di volume tra i diversi file e rendere i segnali più omogenei e comparabili, migliorando così la robustezza del modello. Una volta normalizzati, tutti i file sono stati convertiti a una profondità di 16 bit[10. Questa risoluzione rappresenta uno standard consolidato nell'elaborazione audio digitale, offrendo una qualità uniforme e compatibile con molte librerie e strumenti di machine learning.
L'utilizzo di una bit depth uniforme ha permesso di evitare problemi legati a formati di dati eterogenei, garantendo una standardizzazione che facilita il processo di apprendimento del modello.
Un altro aspetto chiave del pre-processing è stata la conversione di tutte le registrazioni in mono canale. [9] Questa scelta ha consentito di ridurre la complessità computazionale eliminando una traccia ridondante, poiché i file stereo contengono due canali distinti (destro e sinistro) che, nel contesto di questo studio, non aggiungono informazioni rilevanti ai fini dell'addestramento. La conversione in mono ha inoltre garantito che tutti i file

avessero la stessa struttura, rendendo il dataset più uniforme e facilmente elaborabile dal modello.

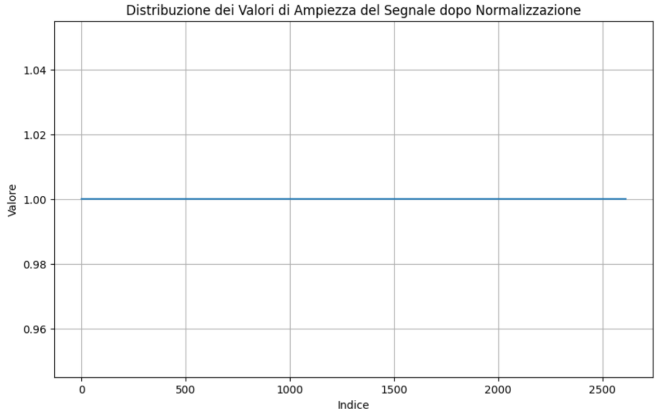

Figure 8: Ampiezza segnali dopo la normalizzazione

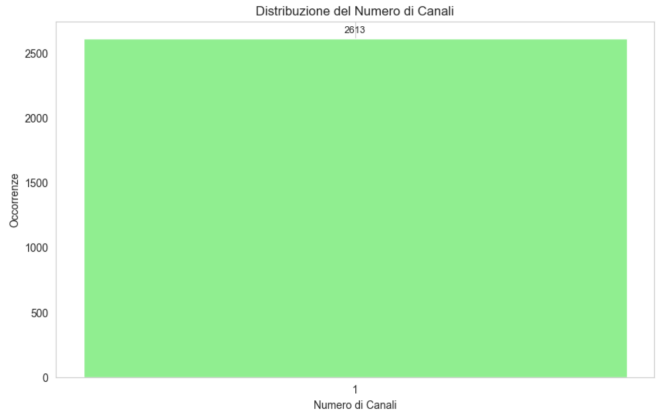

Figure 9: Distribuzione monocanale

Figure 10: 16 bit PCM Bit Depth

Inoltre, la frequenza di campionamento di tutti i samples è stata uniformata a un valore standard di 96 kHz.[11] Questo passaggio era fondamentale per assicurare la confrontabilità tra le diverse registrazioni e garantire che il modello ricevesse input coerenti.

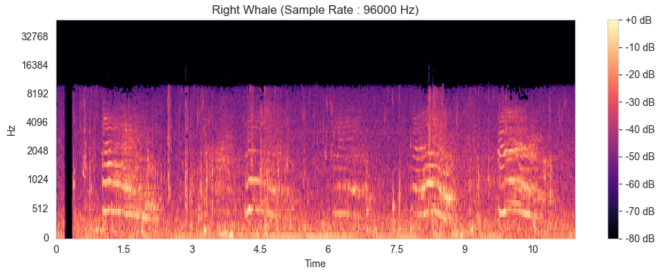

Figure 11: Ricampionamento a 96000 kHz

I samples sono stati poi tagliati a 4 secondi, generando sub-samples di durata uniforme di 4 secondi[12], un'operazione che semplifica l'elaborazione e l'analisi dei dati. Nel caso di registrazioni più brevi, sono stati aggiunti periodi di silenzio[13] per raggiungere la lunghezza desiderata. Il risultato prodotto è un dataset contentente 50993 sub-samples.

Queste operazioni di pre-processing, dalla normalizzazione all'aggiunta di silenzio, hanno permesso di standardizzare i dati in modo da ottimizzare l'addestramento del modello. L'obiettivo di questa fase è stato garantire che i segnali fossero rappresentativi delle loro caratteristiche, evitando che differenze tecniche nei formati o nei livelli di qualità dei file potessero influenzare l'apprendimento del modello più delle informazioni effettivamente rilevanti.

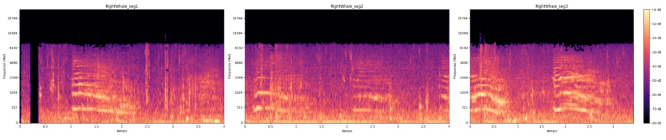

Figure 12: Spettrogrammi sample audio segmentato

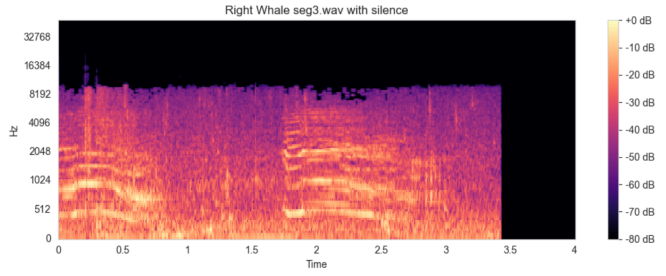

Figure 13: Aggiunta silenzio

6. Metodologie

L'estrazione delle features dai segnali audio rappresenta un passo fondamentale nella costruzione di modelli di intelligenza artificiale per l'analisi bioacustica. Ogni feature numerica selezionata consente di descrivere specifiche proprietà del segnale, facilitando la distinzione tra suoni bioacustici e antropogenici, nonché la classificazione delle sorgenti sonore. In questo contesto, le features spettro-temporali forniscono informazioni dettagliate sul contenuto spettrale e sulla struttura temporale dei segnali, che sono fondamentali per una corretta identificazione e classificazione.

- **Spectral Centroid Mean** Il centroide spettrale rappresenta la media ponderata delle frequenze presenti in un segnale, e può essere considerato come un indicatore della "luminosità" del suono. In termini matematici, il centroide spettrale è definito come:

$$\frac{1}{E}\sum_i i \cdot E_i$$

dove i rappresenta la posizione o l'indice di un elemento (ad esempio, un frame o un bin di frequenza), Ei rappresenta l'energia associata a quella particolare posizione i, E è l'energia totale del segnale, calcolata come somma di tutte le Ei. Nelle applicazioni di segnale e audio, questa può essere vista come una misura di dove si trova il "baricentro energetico" in un determinato dominio (come il dominio delle frequenze, ad esempio nello spettro audio).

- **Spectral Bandwidth RMS** La larghezza di banda spettrale RMS misura la dispersione delle frequenze attorno al centroide spettrale e riflette la complessità del segnale:

$$RMS_i = \sqrt{\frac{1}{E}\sum_i i^2 E_i - i^{-2}}$$

Questa feature è cruciale per distinguere tra suoni semplici e complessi, dove i suoni antropogenici tendono ad avere una maggiore larghezza di banda rispetto ai segnali bioacustici (10)(Lerch, 2012).

- **Standard Deviation** La deviazione standard dello spettro descrive la variazione delle frequenze rispetto al centroide spettrale:

$$\sigma_s = \sqrt{\frac{1}{n-1}\sum_i (s[i] - \mu_s)^2}$$

Questa feature permette di valutare la stabilità spettrale del segnale, utile per identificare suoni con fluttuazioni frequenti tipiche di alcuni rumori antropogenici(16) (Sharma et al., 2020).

- **Skewness** L'asimmetria (Skewness) dello spettro quantifica la simmetria della distribuzione delle frequenze attorno al centroide:

$$\frac{1}{n}\sum_i \left(\frac{s[i] - \mu_s}{\sigma_s}\right)^3$$

Valori positivi indicano una prevalenza di frequenze più alte rispetto alla media, mentre valori negativi indicano il contrario. Questa feature è essenziale per distinguere suoni con distribuzioni spettrali atipiche, come quelli prodotti da alcune specie marine (1)(Bishop, 2006).

- **Kurtosis** La curtosi misura la "puntosità" della distribuzione delle frequenze:

$$\frac{1}{n}\sum_i \left(\frac{s[i] - \mu_s}{\sigma_s}\right)^4$$

Una curtosi elevata indica la presenza di picchi spettrali accentuati, tipici di suoni impulsivi come i clic dei cetacei, utili per identificare eventi sonori specifici (5)(Figueroa et al., 2015).

- **Shennon Entropy** L'entropia di Shannon misura la quantità di informazione o complessità del segnale:

$$-\sum_j p(s_j)\log_2(p(s_j))$$

Questa feature è indicativa della complessità del segnale, con valori più alti che suggeriscono una maggiore variabilità, utile per differenziare tra suoni complessi e semplici(4)(Cover & Thomas, 2006).

- **Renyi Entropy** l'entropia di Renyi permette di analizzare segnali che presentano una distribuzione di probabilità complessa o non uniforme, il che è comune in molte applicazioni, come la compressione dei segnali e il riconoscimento dei modelli. A differenza dell'entropia di Shannon, che considera solo le probabilità degli eventi, l'entropia di Renyi introduce un parametro (α) che consente di pesare in modo differente gli eventi. L'entropia di Renyi è definita come segue:

  $$\frac{1}{1-\alpha}\log_2\left(\sum_j p(s_j)^\alpha\right)$$

  Utilizzando l'entropia di Renyi, si possono confrontare segnali di diversa complessità e determinare la loro informazione contenuta. Ad esempio, valori elevati di α enfatizzano eventi rari, mentre valori più bassi si concentrano su eventi più probabili, rendendo questa misura utile per applicazioni in cui è necessario rilevare anomalie o variazioni significative nel segnale.

- **Rate of Attack** Il rate of attack è una misura che quantifica la velocità con cui un segnale aumenta in ampiezza in un intervallo di tempo specifico. Nella progettazione di sistemi audio e nel trattamento dei segnali, questo parametro è particolarmente importante per comprendere come un segnale o un suono, evolve nel tempo.

  $$\max_i\left(\frac{s[i]-s[i-1]}{n}\right)$$

- **Rate of decay** Il Rate of Decay è un parametro importante nell'analisi dei segnali audio e della loro percezione, che indica la velocità con cui un suono perde la sua intensità dopo aver raggiunto il picco.

  $$\min\quad\frac{s[i]-s[i+1]}{n}$$

  Quando un suono viene prodotto, la sua intensità iniziale è massima. Tuttavia, nel tempo, la pressione sonora inizia a diminuire a causa di vari fattori, come l'assorbimento del suono nell'ambiente e la dispersione. Il Rate of Decay descrive quanto rapidamente avviene questo processo.

- **Silence Ratio** Il silence ratio quantifica la proporzione di "silenzio" nel segnale, definito come la percentuale di frame al di sotto di una certa soglia:

  $$\frac{\#(s \text{ where } s < \text{threshold})}{\sum_i^n s[i]}$$

  Questa feature è utile per discriminare tra suoni continui e suoni intermittenti, spesso associati a fenomeni naturali e antropogenici rispettivamente (7)(Giannakopoulos & Pikrakis, 2014).

- **Threshold Crossing Rate** Il Threshold Crossing Rate misura la frequenza con cui il segnale supera una determinata soglia di ampiezza:

  $$\frac{\#(\text{Threshold Crossing})}{n}$$

  Questa feature è utile per identificare eventi sonori distinti come clic o picchi, frequenti in suoni bioacustici come quelli emessi dai mammiferi marini (14)(Popescu et al., 2009).

- **Mean** La media fornisce un indicatore del livello generale del segnale.

  $$\frac{1}{n}\sum_{i=1}^{n} s[i]$$

- **Max over mean** La Max over Mean è una misura statistica utilizzata nell'analisi dei segnali per descrivere la relazione tra il valore massimo di un segnale e la sua media. Questa misura è particolarmente utile per valutare la variabilità e la distribuzione dei valori di un segnale nel tempo. La formula matematica è così definita:

  $$\text{Max over Mean} = \frac{\max(s[i])}{\frac{1}{n}\sum_i s[i]}$$

- **Min over mean** La Min over Mean è una misura statistica utilizzata nell'analisi dei segnali e dell'audio. Essa rappresenta il rapporto tra il valore minimo di un segnale e la sua media. Questa misura può fornire informazioni utili sulla distribuzione dei valori nel segnale e sulle sue caratteristiche generali. La formula per calcolare la Min over Mean è:

  $$\text{Min over Mean} = \frac{\min(s[i])}{\frac{1}{n}\sum_i s[i]}$$

- **Energy measurements** Le Energy Measurements (misurazioni dell'energia) si riferiscono a tecniche utilizzate per quantificare l'energia contenuta in un segnale audio o in un'onda sonora, sia essa discreta o continua. Possono rappresentare l'energia totale del segnale o la sua energia media, che è normalizzata rispetto alla lunghezza del segnale.

  L'energia fornisce informazioni utili sull'intensità complessiva del segnale, mentre l'energia media è una misura relativa, spesso utilizzata per confrontare segnali di diversa durata. La formula è:

  $$E = \sum_{n=0}^{N-1} |x(n)|^2$$

- **MFCC: Mel-Frequency Cepstral Coefficients** Questi coefficienti sono ampiamente utilizzati nell'analisi audio per rappresentare la forma d'onda in modo che si adatti meglio alla percezione umana delle frequenze. I MFCC catturano le caratteristiche spettrali del segnale, rendendoli particolarmente utili per identificare timbri e texture sonore. La formula per calcolare il n-esimo MFCC è:

$$c_n = \sum_{k=1}^{K} \log(S(k)) \cdot \cos\left(\frac{n\pi(k-\frac{1}{2})}{K}\right)$$

- **ZCR: Zero-Crossing Rate** Questa misura indica la frequenza con cui un segnale attraversa lo zero, fornendo informazioni utili sulla complessità temporale del suono. Viene calcolata come:

$$ZCR = \frac{1}{N}\sum_{n=1}^{N-1} |\text{sgn}(x[n]) - \text{sgn}(x[n-1])|$$

- **Spectral Centroid** Questa caratteristica rappresenta il centro di massa dello spettro di potenza, indicativo della tonalità predominante nel segnale audio. La formula per calcolarlo è:

$$C = \frac{\sum_f f \cdot |X(f)|^2}{\sum_f |X(f)|^2}$$

- **Spectral Bandwidth** Questa caratteristica misura la dispersione delle frequenze rispetto al centroid spettrale e si calcola come:

$$BW = \frac{\sum_f (f-C)^2 \cdot |X(f)|^2}{\sum_f |X(f)|^2}$$

- **Chroma Features** rappresentano le intensità delle dodici note musicali suonate in un brano, permettendo di analizzare l'armonia e la tonalità. Queste caratteristiche sono particolarmente preziose per comprendere le relazioni armoniche tra le note nel contesto dei suoni marini, contribuendo così a un'analisi più ricca e stratificata del dataset. La formula per calcolare la k-esima feature cromatica $C_k(t)$ al tempo t è:

$$C_k(t) = \sum_{f \in F_k} |X(f,t)|^2$$

- **Spectral Contrast** Il contrasto spettrale è una misura che descrive le differenze di intensità tra le bande di frequenza in uno spettro audio. Esso cattura l'interazione tra frequenze adiacenti e fornisce una rappresentazione della struttura armonica di un segnale.Si calcola con la seguente formula:

$$SC_k = \frac{1}{N}\sum_{n=1}^{N} \max(0, |X(f_n)|^2 - |X(f_k)|^2)$$

- **Tonnetz** Il Tonnetz è una rappresentazione delle relazioni armoniche in un brano musicale, basata su intervalli fondamentali come terze, quinte e seste. Esso descrive come queste relazioni evolvono nel tempo, fornendo un'analisi della struttura tonale. Nel calcolo, vengono estratte sei caratteristiche che riflettono le variazioni armoniche legate a questi intervalli. In questo studio, il Tonnetz viene calcolato sull'intero brano audio, restituendo la media di queste sei caratteristiche, che sintetizzano le principali relazioni armoniche senza considerare la variazione temporale della tonalità.

- **Tempo** Il tempo indica la velocità con cui si susseguono i battiti all'interno di un brano. Viene solitamente espresso in battiti per minuto (BPM), che rappresenta il numero di battiti che si verificano in un minuto. Serve per determinare il tempo musicale di un brano, utilizzando l'intensità degli onset (cambiamenti ritmici) per stimare il valore dei battiti al minuto (BPM). Questo è utile per analizzare la velocità e il ritmo di un brano musicale in contesti come la classificazione musicale, la sincronizzazione o altre analisi audio.

7. Integrazione delle Features

La combinazione delle caratteristiche descritte in precedenza si è dimostrata efficace nel migliorare l'accuratezza della classificazione dei segnali bioacustici rispetto all'uso di singole caratteristiche. Studi recenti hanno evidenziato che la selezione e l'ottimizzazione di caratteristiche come le MFCC (Mel-frequency cepstral coefficients), combinate con centroidi spettrali e skewness, possono incrementare significativamente le performance di modelli di machine learning come Support Vector Machines e Random Forests (Malfante et al., 2018). L'integrazione di queste caratteristiche consente di costruire un modello robusto, in grado di distinguere non solo tra suoni antropogenici e bioacustici, ma anche di classificare con precisione le sottoclassi coivolte. Questo approccio fornisce un sistema di monitoraggio acustico avanzato e altamente affidabile.
Sono state inoltre utilizzate numerose caratteristiche numeriche tratte dal progetto BirdNet, tra cui gli MFCC, lo Zero Crossing Rate e il Centroide Spettrale. Queste caratteristiche sono state selezionate per la loro capacità di catturare importanti aspetti acustici, risultando efficaci nella classificazione dei segnali. Successivamente, è stato considerato anche un altro set di caratteristiche proposto da Dhamodaran et al. nello studio "A Survey on Audio Feature Extraction for Automatic Music Genre Classification", ampliando ulteriormente la gamma di caratteristiche analizzate per migliorare le prestazioni dei modelli.
Le caratteristiche individuate sono state suddivise in tre set di dati distinti, ciascuno appartenente a un progetto diverso: BirdNet, Malfante e Dhamodaran. [Table 1]

Table 1
Set di Dati Comparativi

Primo Set (Malfante)	
Spectral Centroid Mean	Spectral Bandwidth RMS
Standard Deviation	Skewness
Kurtosis	Shannon Entropy
Renyi Entropy	Rate of Attack
Rate of Decay	Threshold Crossings
Silence Ratio	Mean
Max Over Mean	Min Over Mean
Energy-Measurements	
Secondo Set (BirdNet)	
MFCC 1	MFCC 2
MFCC 3	MFCC 4
MFCC 5	MFCC 6
MFCC 7	MFCC 8
MFCC 9	MFCC 10
MFCC 11	MFCC 12
MFCC 13	ZCR
Spectral Centroid	Spectral Bandwidth
Chroma 1	Chroma 2
Chroma 3	Chroma 4
Chroma 5	Chroma 6
Chroma 7	Chroma 8
Chroma 9	Chroma 10
Chroma 11	Chroma 12
Terzo Set (Dhamodaran)	
MFCC 1	MFCC 2
MFCC 3	MFCC 4
MFCC 5	MFCC 6
MFCC 7	MFCC 8
MFCC 9	MFCC 10
MFCC 11	MFCC 12
MFCC 13	ZCR
Spectral Contrast	Tonnetz 1
Tonnetz 2	Tonnetz 3
Tonnetz 4	Tonnetz 5
Tonnetz 6	Chroma 1
Chroma 2	Chroma 3
Chroma 4	Chroma 5
Chroma 6	Chroma 7
Chroma 8	Chroma 9
Chroma 10	Chroma 11
Chroma 12	Tempo

8. Split & Oversampling

Inizialmente è stata condotta una classificazione binaria per distinguere i suoni antropogenici e bioacustici. Questa prima fase ha consentito di sviluppare un modello semplice, in grado di identificare con accuratezza i due gruppi principali di suoni. Successivamente il problema è stato esteso a una classificazione multiclasse, con l'obiettivo di categorizzare ulteriormente i suoni antropogenici (Target) in diverse sottoclassi. Questo approccio ha fornito una comprensione più profonda e dettagliata delle caratteristiche audio presenti nel dataset.
Per addestrare e testare il modello, il dataset è stato splittato in tre set: l'80% dei dati è stato dedicato al training, il 10% alla validazione e il 10% al test. Questa divisione ha garantito che il modello fosse addestrato su una porzione sufficientemente ampia di dati, mantenendo al contempo un set indipendente per valutare le prestazioni durante l'addestramento e per testare la generalizzazione su dati mai visti.
Prima di effettuare lo split, il dataset è stato filtrato in modo da garantire che ci fossero almeno 10 samples distinti per ogni sottoclasse. Per la classificazione multiclasse si aggiunge un'ulteriore filtro, che permette di lavorare solo sui sub-samples Target.
Si passa da un totale di 50993 sub-samples a un totale di 49429 per la classificazione binaria, mentre per quanto riguarda la multiclasse, il numero di sub-samples è ridotto a 43089.
Per garantire che i sub-samples di uno stesso sample non finissero in set diversi durante lo split, i nomi dei file sono stati privati dell'estensione *.wav*. Successivamente, i nomi dei file dei sub-samples sono stati raggruppati per nome del file principale, creando una nuova colonna nel dataframe, denominata `Parent`, che contiene i nomi dei file principali dei samples.
Lo split è stato effettuato sulla colonna `Parent` utilizzando la funzione `GroupShuffleSplit` del package `sklearn`, con i parametri `n_splits = 1` e `random_state=42`.
Prima dell'applicazione di SMOTE, il set di addestramento per la classificazione binaria contava 33872 samples per la classe Target e 4917 per la classe Non-Target.
Dopo l'uso di SMOTE, entrambe le classi sono state bilanciate, bilanciando prima tra di loro le sottoclassi e successivamente le classi, ottenendo 44150 sub-samples sia per Target che per Non-target. [Table 2]
Per quanto riguarda la classificazione multiclasse, SMOTE è stato utilizzato per bilanciare tutte le sottoclassi della classe Target in training, portandole a 8830 samples ciascuna. A questo punto, la classe Target è composta da un set di addestramento con 5 sottoclassi (tutte bilanciate), uno di validazione con 5 sottoclassi e uno di test con 5 sottoclassi [Table 3]. Durante l'applicazione di SMOTE, il parametro `k_neighbors` è stato testato con diversi valori per determinare l'adeguato numero di vicini da considerare nella generazione dei samples sintetici. È stato scelto un valore relativamente basso di `k_neighbors`, poiché alcune delle sottoclassi erano poco rappresentate nel dataset.

	Target	Non-Target	Tot
Train	33872	4917	38789
Train Balanced	44150	44150	88300
Test	4814	810	5624
Validation	4403	613	5016

Table 2
Classificazione binaria

	0	1	2	3	4	Tot
Train	7802	8830	8744	8213	283	33872
Train Balanced	8830	8830	8830	8830	8830	44150
Test	1007	1607	1201	993	6	4814
Validation	862	1291	1276	956	18	4403

Table 3
Classificazione multiclasse

Inoltre, le sotto-classi con un solo sub-sample sono state rimosse, in quanto non consentivano il corretto funzionamento di SMOTE. Questo approccio ha preservato meglio le caratteristiche di ciascuna sottoclasse minoritaria, evitando la creazione di samples sintetici troppo distanti dai dati reali.
Questo ha contribuito a mantenere la diversità dei segnali bioacustici e ha consentito al modello di apprendere caratteristiche distintive per ciascuna categoria (3)(Chawla et al., 2002).

9. Addestramento

Per l'addestramento del modello, sono stati utilizzati tre algoritmi di machine learning: Random Forest, Support Vector Machine (SVM) e LightGBM. Ognuno di questi modelli è stato scelto per le proprie caratteristiche distintive e la capacità di affrontare problemi di classificazione in contesti complessi come quello dei segnali bioacustici.

- **Random Forest**: Questo modello di machine learning è un ensemble di alberi decisionali che utilizza il metodo del bagging per migliorare la precisione predittiva e ridurre il rischio di overfitting. La Random Forest è particolarmente adatta per dataset con molte variabili, poiché consente di catturare relazioni complesse e non lineari nei dati. Nel caso specifico, il modello è stato configurato utilizzando una ricerca a griglia (*Grid Search*) per ottimizzare i principali iperparametri. La griglia includeva diverse combinazioni di valori per il numero di alberi (*n_estimators*), impostato tra 100, 150 e 200, la profondità massima degli alberi (*max_depth*), variata tra 5, 7 e 9, e la dimensione minima per una suddivisione interna di un nodo (*min_samples_split*), con valori pari a 10 e 15. Inoltre, è stata considerata la dimensione minima di campioni necessaria per formare una foglia (*min_samples_leaf*), impostata tra 4 e 5.

Per ogni combinazione di parametri, è stata eseguita una validazione incrociata a cinque fold per garantire che il modello selezionato fosse robusto e generalizzabile. La funzione ha utilizzato la strategia *max_features="sqrt"* per selezionare casualmente un sottoinsieme di caratteristiche alla costruzione di ogni albero, riducendo la correlazione tra gli alberi e migliorando la diversità del modello. Inoltre, è stato impostato un valore di *random seed* pari a 42 per garantire la riproducibilità dei risultati.

L'addestramento del modello è stato eseguito sui dati di training e il modello ottimale selezionato dalla ricerca a griglia è stato valutato sui set di validazione e test.

- **Support Vector Machine (SVM)**: Le SVM (Support Vector Machine) sono uno strumento potente per la classificazione, progettato per trovare l'iperpiano ottimale che separa le classi nel piano multidimensionale. Sono particolarmente utili in scenari ad alta dimensione, dove possono essere utilizzate diverse funzioni di kernel per gestire la non linearità. In questo studio, l'SVM è stato configurato con il kernel Radial Basis Function (RBF), che è il valore predefinito nella libreria `scikit-learn` mentre il parametro *random_state* è impostato su `42` per garantire la riproducibilità dei risultati.
- **LightGBM**: Questo algoritmo di boosting è progettato per essere efficiente in termini di tempo e spazio. Utilizza un approccio di gradient boosting basato su alberi e si distingue per la capacità di gestire grandi dataset, riducendo il numero di dati necessari per l'addestramento grazie all'ottimizzazione delle risorse. Il modello di LightGBM è inizializzato con il parametro *random_state* impostato a un valore fisso (42) per garantire la riproducibilità dei risultati.

  L'addestramento viene eseguito utilizzando il metodo `lgb.train`, che sfrutta il set di addestramento e di validazione per ottimizzare il modello in un numero predefinito di iterazioni (`num_boost_round=1000`). Durante ogni iterazione, il modello apprende dalle caratteristiche dei dati, cercando di minimizzare la funzione di perdita *multi_logloss*.

10. Esperimenti con sample rate diversi

Gli addestramenti sono stati ripetuti utilizzando diversi sample rate per analizzare l'impatto della risoluzione temporale dei dati sull'accuratezza dei modelli di classificazione. In particolare, sono stati utilizzati tre sample rate differenti, diversi dal 96 kHz, ovvero 192000 Hz (192 kHz), 44100 Hz (44.1 kHz) e 384000 Hz (384 kHz). Il primo, 44100 Hz (44.1 kHz), è un valore standard utilizzato in molti contesti audio, inclusi i formati musicali, e rappresenta una risoluzione temporale adeguata per la maggior parte delle applicazioni sonore. Il secondo, 192000 Hz (192 kHz), offre una risoluzione superiore rispetto al classico 44.1 kHz, ed è utilizzato in applicazioni che richiedono una qualità audio più alta, come nella registrazione audio professionale. Infine, il sample rate di 384000 Hz (384 kHz) è stato scelto per esplorare il comportamento dei modelli su dati con una risoluzione molto alta, al fine di valutare come l'alta frequenza di campionamento influenzi la capacità del modello di catturare dettagli sottili nei segnali audio. Questi esperimenti permettono di esaminare la robustezza dei modelli rispetto alle variazioni nel campionamento e di identificare eventuali trade-off tra risoluzione e prestazioni.

- **44100 Hz (44.1 kHz)**: È il sample rate standard per l'audio musicale, utilizzato nei CD e in molte applicazioni di registrazione audio professionale. È adeguato per la maggior parte delle analisi audio, garantendo una qualità elevata senza un'elevata richiesta computazionale.
- **192000 Hz (192 kHz)**: Viene utilizzato in applicazioni che richiedono una qualità audio superiore rispetto al 44.1 kHz, come la registrazione audio di alta qualità e in contesti professionali dove è importante preservare dettagli acustici con una risoluzione maggiore.
- **384000 Hz (384 kHz)**: Questo sample rate è utilizzato in applicazioni ad alta fedeltà, come nella registrazione audio professionale di altissima qualità, nelle analisi scientifiche o nelle applicazioni che richiedono una risoluzione temporale molto elevata per catturare dettagli minimi nei segnali audio, come nelle analisi acustiche avanzate.

11. Risultati

In questa sezione vengono riportati i risultati ottenuti dai modelli di classificazione implementati. L'obiettivo di questa analisi è valutare l'efficacia di ciascun modello nel classificare correttamente il dataset, utilizzando tre diversi insiemi di feature selezionate e i sample rate differenti scelti per gli esperimenti. I modelli sono stati valutati sia sui dati di validazione che sui dati di test, al fine di fornire una visione completa delle loro prestazioni.
Le metriche di performance riportate considerano i risultati ottenuti durante la fase di validazione, utilizzata per ottimizzare i modelli, e quelli osservati sul set di test, che forniscono una stima delle prestazioni del modello su dati mai visti prima. Questo approccio permette di confrontare il comportamento dei modelli durante l'addestramento e la loro capacità di generalizzare a nuovi dati, garantendo che le performance siano robuste e generalizzabili.
Per misurare le prestazioni di ciascun modello, sono state utilizzate le seguenti metriche:

- **Precision**: rappresenta la proporzione di predizioni corrette tra quelle assegnate alla classe positiva. È particolarmente utile quando l'obiettivo è ridurre il numero di falsi positivi, ad esempio quando l'assegnazione errata a una classe positiva è particolarmente costosa.
- **Recall**: misura la capacità del modello di identificare correttamente tutti i campioni appartenenti alla classe positiva. In altre parole, quanti veri positivi il modello è riuscito a rilevare. È importante quando si desidera ridurre il numero di falsi negativi.
- **F1-score**: è la media armonica tra precisione e recall, combinando entrambe le metriche in un unico valore. È utile quando c'è uno squilibrio tra le classi e non è preferibile privilegiare né la precisione né il recall. Un F1-score alto indica che il modello bilancia bene entrambi i fattori.
- **Accuracy**: metrica globa rappresenta la percentuale complessiva di predizioni corrette rispetto al totale dei campioni. Sebbene sia una metrica comune, l'accuracy può essere fuorviante in caso di dataset sbilanciati, dove il modello potrebbe semplicemente predire la classe maggioritaria senza effettivamente imparare a discriminare correttamente.

Inoltre, è stata considerata la **macro-avg**, ovvero la media semplice delle metriche per ciascuna classe. La macro-avg calcola separatamente la precision, recall e F1 per ogni classe e successivamente prende la loro media, indipendentemente dalla distribuzione delle classi. Questo tipo di media è utile quando si vuole avere una valutazione equilibrata del modello, dando lo stesso peso a tutte le classi. Nel processo di valutazione, sono stati utilizzati sia un set di validazione che un set di test per stimare le capacità di generalizzazione dei modelli:

- **Set di validazione**: utilizzato durante il processo di addestramento per ottimizzare i modelli e selezionare i parametri migliori. Le metriche calcolate su questo set aiutano a monitorare l'adattamento dei modelli ai dati e a evitare fenomeni di overfitting, che si verificano quando un modello si adatta troppo strettamente ai dati di addestramento, perdendo la capacità di generalizzare. Il set di validazione viene utilizzato per regolare i parametri, come la profondità dell'albero nel caso di modelli ad albero decisionale, o il tasso di apprendimento nei modelli di gradient boosting.
- **Set di test**: utilizzato esclusivamente dopo l'addestramento e la selezione dei parametri. Serve per valutare l'efficacia finale del modello su dati completamente nuovi, fornendo una stima realistica delle sue prestazioni. Questo set non viene mai utilizzato durante la fase di addestramento, evitando così di influenzare la valutazione finale delle prestazioni del modello. Il set di test offre una misura indipendente e obiettiva della capacità di generalizzazione del modello su dati che non ha mai visto.

I risultati ottenuti dai modelli su ciascun set di feature sono riportati nelle tabelle che seguono, con le metriche calcolate su entrambi i set di validazione e di test. Le differenze tra le due valutazioni possono indicare la capacità del modello di generalizzare correttamente. Se il modello ha buone prestazioni sul set di addestramento e di validazione, ma scarse prestazioni sul set di test, potrebbe esserci un problema di overfitting, in cui il modello si è adattato troppo ai dati di addestramento. Se, invece, le prestazioni sono basse su entrambi i set, ciò potrebbe suggerire un problema di underfitting, in cui il modello non è stato in grado di imparare dai dati. Le tabelle riporteranno le performance per ciascuna classe, oltre alla media di queste metriche per tutte le classi, permettendo di analizzare la capacità del modello di discriminare correttamente tra le diverse categorie.

SR	Random Forest					SVM					LightGBM				
	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support
Set 1															
44100	0.97	0.99	0.98	0.9928	5016	0.90	0.97	0.93	0.9679	5016	0.97	1.00	0.98	0.9926	5016
96000	0.97	0.99	0.98	0.9926	5016	0.83	0.72	0.76	0.9135	5016	0.97	0.99	0.98	0.9928	5016
192000	0.97	0.99	0.98	0.9926	5016	0.87	0.88	0.87	0.9444	5016	0.97	0.99	0.98	0.9932	5016
384000	0.97	0.99	0.98	0.9922	5016	0.82	0.71	0.74	0.9093	5016	0.98	1.00	0.98	0.9934	5016
Set 2															
44100	0.98	1.00	0.99	0.9954	5016	0.96	0.97	0.96	0.9843	5016	0.99	1.00	0.99	0.9958	5016
96000	0.97	0.99	0.98	0.9908	5016	0.96	0.95	0.95	0.9805	5016	0.98	0.99	0.99	0.9942	5016
192000	0.97	0.99	0.98	0.9908	5016	0.95	0.95	0.95	0.9793	5016	0.98	0.99	0.99	0.9944	5016
384000	0.95	0.99	0.97	0.9864	5016	0.91	0.88	0.89	0.9557	5016	0.97	0.99	0.98	0.9922	5016
Set 3															
44100	0.98	0.99	0.99	0.9936	5016	0.97	0.98	0.97	0.9886	5016	0.98	1.00	0.99	0.9954	5016
96000	0.95	0.99	0.97	0.9870	5016	0.95	0.98	0.97	0.9846	5016	0.97	0.99	0.98	0.9924	5016
192000	0.95	0.99	0.97	0.9862	5016	0.96	0.97	0.96	0.9841	5016	0.98	0.99	0.98	0.9934	5016
384000	0.95	0.99	0.97	0.9846	5016	0.96	0.97	0.96	0.9839	5016	0.97	0.99	0.98	0.9908	5016

Table 4
Risultati validation classificazione binaria

SR	Random Forest					SVM					LightGBM				
	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support
Set 1															
44100	0.98	0.98	0.98	0.9918	5624	0.87	0.95	0.90	0.9452	5624	0.99	0.99	0.99	0.9945	5624
96000	0.98	0.98	0.98	0.9897	5624	0.86	0.86	0.86	0.9294	5624	0.99	0.98	0.98	0.9924	5624
192000	0.98	0.97	0.97	0.9860	5624	0.82	0.90	0.85	0.9180	5624	0.98	0.97	0.97	0.9868	5624
384000	0.97	0.97	0.97	0.9860	5624	0.84	0.84	0.84	0.9195	5624	0.97	0.97	0.97	0.9844	5624
Set 2															
44100	0.99	0.99	0.99	0.9945	5624	0.94	0.96	0.95	0.9749	5624	0.99	0.99	0.99	0.9947	5624
96000	0.98	0.98	0.98	0.9908	5624	0.94	0.97	0.97	0.9737	5624	0.99	0.98	0.98	0.9922	5624
192000	0.97	0.99	0.98	0.9883	5624	0.93	0.92	0.92	0.9632	5624	0.98	0.99	0.98	0.9911	5624
384000	0.99	0.99	0.99	0.9950	5624	0.93	0.93	0.93	0.9651	5624	0.99	0.98	0.99	0.9936	5624
Set 3															
44100	0.99	0.98	0.98	0.9913	5624	0.97	0.96	0.97	0.9838	5624	0.98	0.99	0.98	0.9924	5624
96000	0.97	0.98	0.98	0.9886	5624	0.97	0.96	0.96	0.9806	5624	0.99	0.99	0.99	0.9961	5624
192000	0.96	0.98	0.97	0.9858	5624	0.96	0.95	0.96	0.9792	5624	0.99	0.99	0.99	0.9950	5624
384000	0.98	0.99	0.99	0.9932	5624	0.96	0.93	0.94	0.9737	5624	0.99	1.00	0.99	0.9973	5624

Table 5
Risultati test classificazione binaria

SR	Random Forest					SVM					LightGBM				
	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support
Set 1															
44100	0.44	0.53	0.46	0.4113	4403	0.30	0.39	0.26	0.2246	4403	0.42	0.51	0.45	0.3891	4403
96000	0.42	0.53	0.44	0.4018	4403	0.29	0.37	0.23	0.2071	4403	0.42	0.51	0.45	0.3950	4403
192000	0.42	0.52	0.45	0.3961	4403	0.30	0.45	0.24	0.2914	4403	0.41	0.50	0.44	0.3797	4403
384000	0.41	0.52	0.44	0.3945	4403	0.27	0.47	0.27	0.3261	4403	0.41	0.50	0.44	0.3825	4403
Set 2															
44100	0.56	0.66	0.59	0.5816	4403	0.33	0.34	0.32	0.3770	4403	0.59	0.63	0.60	0.5358	4403
96000	0.52	0.63	0.54	0.5335	4403	0.32	0.38	0.31	0.3554	4403	0.48	0.61	0.50	0.5081	4403
192000	0.50	0.60	0.53	0.5074	4403	0.31	0.49	0.30	0.3573	4403	0.51	0.59	0.52	0.4897	4403
384000	0.46	0.56	0.48	0.4420	4403	0.27	0.47	0.26	0.3189	4403	0.49	0.54	0.50	0.4295	4403
Set 3															
44100	0.56	0.66	0.59	0.5816	4403	0.33	0.34	0.32	0.3770	4403	0.59	0.63	0.60	0.5358	4403
96000	0.52	0.62	0.55	0.5226	4403	0.53	0.65	0.55	0.5435	4403	0.50	0.60	0.52	0.5033	4403
192000	0.50	0.61	0.52	0.5124	4403	0.48	0.60	0.49	0.4840	4403	0.53	0.61	0.54	0.5126	4403
384000	0.47	0.57	0.49	0.4583	4403	0.38	0.50	0.39	0.3741	4403	0.47	0.56	0.48	0.4520	4403

Table 6
Risultati validation classificazione multiclasse

SR	Random Forest					SVM					LightGBM				
	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support	Precision	Recall	f1-score	Accuracy	Support
Set 1															
44100	0.37	0.56	0.39	0.4283	4814	0.23	0.39	0.20	0.2279	4814	0.37	0.52	0.40	0.3887	4814
96000	0.38	0.57	0.40	0.4377	4814	0.21	0.37	0.18	0.2015	4814	0.40	0.54	0.43	0.4082	4814
192000	0.35	0.55	0.37	0.4063	4814	0.25	0.44	0.21	0.2704	4814	0.34	0.53	0.37	0.3903	4814
384000	0.36	0.54	0.36	0.4084	4814	0.26	0.43	0.25	0.3392	4814	0.34	0.53	0.36	0.3968	4814
Set 2															
44100	0.52	0.60	0.54	0.4877	4814	0.31	0.51	0.32	0.3704	4814	0.55	0.60	0.57	0.4977	4814
96000	0.54	0.59	0.55	0.4846	4814	0.30	0.37	0.35	0.3693	4814	0.43	0.59	0.46	0.4842	4814
192000	0.63	0.63	0.62	0.5293	4814	0.30	0.51	0.29	0.3606	4814	0.60	0.60	0.60	0.4965	4814
384000	0.56	0.58	0.56	0.4566	4814	0.26	0.50	0.25	0.3319	4814	0.52	0.57	0.54	0.4695	4814
Set 3															
44100	0.52	0.60	0.54	0.4877	4814	0.31	0.51	0.32	0.3704	4814	0.55	0.60	0.57	0.4977	4814
96000	0.56	0.60	0.57	0.4861	4814	0.46	0.60	0.49	0.4886	4814	0.44	0.60	0.47	0.4915	4814
192000	0.61	0.62	0.61	0.5154	4814	0.43	0.56	0.46	0.4209	4814	0.60	0.60	0.60	0.4965	4814
384000	0.53	0.57	0.54	0.4514	4814	0.38	0.56	0.41	0.3741	4814	0.56	0.59	0.57	0.4809	4814

Table 7
Risultati test classificazione multiclasse

12. Conclusioni

I risultati ottenuti nei diversi esperimenti di classificazione binaria e multiclasse mostrano chiaramente che il modello LightGBM si è rivelato il più performante rispetto a Random Forest e SVM. Per quanto riguarda la classificazione binaria, LightGBM ha raggiunto i migliori risultati in termini di precision, recall, F1-score e accuracy. Durante la fase di test, il campionamento a 44100 Hz all'interno del Set 2 si è dimostrato particolarmente efficace, raggiungendo valori di precision pari a 0.97, recall di 1.00, F1-score di 0.99 e un'accuratezza complessiva di 0.9946. Anche il modello Random Forest ha mostrato prestazioni eccellenti, pur rimanendo leggermente inferiore rispetto a LightGBM in alcune configurazioni. Al contrario, SVM ha evidenziato maggiori difficoltà, in particolare con sample rate più bassi, dove i valori di precision e recall sono stati nettamente inferiori rispetto agli altri modelli.
Nel caso della classificazione multiclasse, LightGBM ha confermato nuovamente le sue prestazioni superiori. Il campionamento a 44100 Hz combinato con i dati del Set 2 ha garantito i risultati migliori, con un F1-score pari a 0.59 e un'accuratezza di 0.5538. Anche in questa configurazione, Random Forest ha fornito prestazioni adeguate, ma comunque inferiori rispetto a LightGBM. SVM, invece, si è dimostrato il modello meno performante, con difficoltà evidenti nella capacità di generalizzare i risultati.
Un aspetto rilevante dell'analisi è legato all'influenza del sample rate. Tra i vari valori testati, il campionamento a 44100 Hz si è dimostrato il più stabile e performante, indipendentemente dal tipo di classificazione. Anche la configurazione dei dati utilizzata ha avuto un impatto significativo: il Set 2 è risultato essere il più affidabile, garantendo le migliori prestazioni sia in validazione che in test, a prescindere dal modello utilizzato.
In conclusione, l'analisi evidenzia che il modello LightGBM, in combinazione con il sample rate di 44100 Hz e i dati del Set 2, rappresenta la configurazione ottimale per affrontare sia il problema di classificazione binaria che quello multiclasse. Questa combinazione consente di ottenere risultati accurati e robusti, rendendola particolarmente adatta per applicazioni pratiche nel contesto analizzato.

13. Sviluppi futuri

Gli sviluppi futuri di questo progetto possono essere orientati verso diversi aspetti strategici, finalizzati a migliorare l'accuratezza e l'efficacia del modello di classificazione.
In primo luogo, l'utilizzo di un dataset più ampio e diversificato rappresenta una priorità. La raccolta di suoni bioacustici e antropogenici provenienti da una varietà più estesa di habitat e condizioni ambientali consentirebbe di creare un dataset maggiormente rappresentativo. Questo approccio migliorerebbe la capacità del modello di generalizzare, garantendo prestazioni affidabili anche in contesti reali complessi e variabili.
Parallelamente, l'integrazione di tecniche avanzate per la selezione e l'estrazione delle caratteristiche acustiche potrebbe apportare significativi benefici. Una rappresentazione più dettagliata e descrittiva dei segnali audio potrebbe essere ottenuta mediante l'applicazione di metodologie come l'analisi delle componenti principali (PCA) o tecniche basate sul deep learning per il feature extraction. Questi strumenti aiuterebbero a ottimizzare la struttura dei dati in input, consentendo al modello di catturare con maggiore precisione le peculiarità distintive dei segnali bioacustici e antropogenici.
Infine, l'adozione di modelli di machine learning più avanzati potrebbe rappresentare un ulteriore passo avanti. Tecnologie come le reti neurali convoluzionali (CNN) o approcci basati sul deep learning sono particolarmente promettenti per la classificazione audio grazie alla loro capacità di apprendere rappresentazioni complesse direttamente dai dati grezzi. Inoltre, un approccio combinato, che integri diverse tipologie di modelli, potrebbe sfruttare la complementarità delle varie tecniche di classificazione, migliorando la robustezza e l'affidabilità del sistema.
Questi sviluppi, nel loro insieme, potrebbero offrire un contributo significativo al perfezionamento del sistema, supportando una gestione più sostenibile degli ecosistemi marini. Una più accurata identificazione dei segnali acustici sarebbe infatti cruciale per monitorare e mitigare l'impatto delle attività antropiche, favorendo una migliore conservazione della biodiversità e la protezione degli habitat naturali.

14. Bibliography

References

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Blumstein, D. T., Mennill, D. J., Clemins, P., Girod, L., Yao, K., Patricelli, G., ... & Kirschel, A. N. (2011). Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology, 48*(3), 758-767.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

[4] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons.

[5] Figueroa, L., Belloni, M., & Abascal-Mena, R. (2015). Statistical modeling of cetacean click properties for automatic classification. *Journal of the Acoustical Society of America, 137*(3), 1025-1034.

[6] Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution, 10*(2), 169-185.

[7] Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press.

[8] Hildebrand, J. A. (2009). Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series, 395*, 5-20.

[9] Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics, 61*, 101236.

[10] Lerch, A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. John Wiley & Sons.

[11] Malfante, M., Mars, J. I., Dalla Mura, M., & Gervaise, C. (2018). Automatic fish sounds classification in real-life marine environments: Performance evaluation and perspectives. *Journal of the Acoustical Society of America, 143*(5), 2834-2845.

[12] Mellinger, D. K., & Clark, C. W. (2000). Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America, 107*(6), 3518-3529.

[13] Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence, 116*, 374-388.

[14] Popescu, M., Ehrlich, R., & Xu, W. (2009). Detection and classification of acoustic events for in-home monitoring of an elder person living alone. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 314-319).

[15] Salamon, J., Bello, J. P., Farnsworth, A., & Rosenheim, K. (2017). Feature learning with convolutional neural networks for bioacoustics. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2662-2666). IEEE.

[16] Sharma, R., Agarwal, S., & Jain, R. (2020). Acoustic signal classification using hybrid features and machine learning techniques. *Journal of Ambient Intelligence and Humanized Computing, 11*(3), 1171-1180.

[17] Sueur, J., Aubin, T., & Simonis, C. (2008). Equipment review: Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics, 18*(2), 213-226.

[18] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5), 293-302.

[19] Dhamodaran, P., Balakrishnan, V., & Dharmavaram, A. (2023). A Survey on Audio Feature Extraction for Automatic Music Genre Classification. *International Journal of Recent Technology and Engineering, 12*(3), 45-52.