UNDERWATER CLASSIFICATION

ABSTRACT

Questo studio si concentra sulla distinzione tra suoni generati dagli esseri umani(Target) e quelli prodotti dai pesci(Non Target), con l'obiettivo di migliorare il monitoraggio degli ecosistemi marini e mitigare l'impatto delle attività antropogeniche. Utilizzando tecniche avanzate di machine learning e analisi del segnale, abbiamo sviluppato un sistema di classificazione automatica in grado di identificare e separare con precisione i suoni antropogenici dai suoni della fauna marina. La metodologia comprende la raccolta di un ampio dataset di registrazioni audio subacquee, l'estrazione di caratteristiche distintive come spettri di frequenza e MFCC, e l'addestramento di modelli di Machine Learning. L'addestramento è stato suddiviso in due fasi principali: nella prima fase, abbiamo implementato una classificazione binaria, in cui il modello doveva determinare se ogni registrazione audio contenesse suoni generati da esseri umani o da pesci. Nella seconda fase, abbiamo applicato una classificazione multiclasse per i suoni generati dai pesci. In questa fase, il sistema è stato addestrato a riconoscere a quale sottoclasse specifica appartenesse ogni audio.

1. Introduzione

Nell'era dell'informazione, la quantità di dati audio generati e raccolti è in continua crescita. L'audio rappresenta una fonte ricca e variegata di informazioni. Tuttavia, la gestione e l'analisi di questi dati possono rappresentare una sfida significativa, soprattutto quando il volume di dati diventa ingente. Ci sono diversi campi dove ampie quantità di dati vengono gestite e analizzate: Ad esempio l'oceano è un ambiente acusticamente ricco e complesso, dove una varietà di suoni naturali e artificiali si mescolano continuamente. Con l'avvento delle tecnologie di registrazione subacquea, la raccolta di dati audio marini è diventata una pratica comune per molteplici scopi, tra cui la ricerca scientifica, la conservazione della fauna marina e il monitoraggio delle attività umane. Diventa interessante comprendere e analizzare a fondo questa tematica in modo tale da poter classificare gli audio marini per distinguerli tra suoni generati da esseri umani (**Target**) e suoni prodotti dalla fauna marina, in particolare i pesci (**Non Target**). Questa distinzione è cruciale per vari motivi: proteggere e preservare gli ecosistemi marini, migliorare la ricerca scientifica, sviluppare nuove tecnologie e strumenti di monitoraggio. Utilizzando tecniche avanzate di machine learning e analisi del segnale, il nostro approccio contribuirà a una gestione più sostenibile e consapevole degli ecosistemi marini.

2. Struttura del Dataset

Il dataset utilizzato per questo studio è stato organizzato in due principali sottocartelle: Target e Non Target. La cartella Target contiene i suoni antropogenici mentre la cartella Non Target include i suoni prodotti dai pesci. All'interno di ciascuna di queste due principali categorie, i dati sono ulteriormente suddivisi in sottocategorie specifiche per una classificazione più dettagliata. Nella cartella Target, i dati sono suddivisi 16 categorie, ognuna rappresentante una tipologia di suono umano differente, ad esempio Bombe sigillate, esplosioni, pescatori. Nella cartella Non Target, le registrazioni audio sono organizzate in 96 categorie basate sui diversi tipi di suoni marini identificati, ad esempio Tamburo nero, Delfino climene, Leone marino. Le registrazioni audio nel dataset variano notevolmente in durata. Alcuni file audio sono di breve durata, mentre altri possono estendersi su periodi più lunghi.

3. Analisi

Il primo passo nel nostro studio è stato esaminare attentamente il dataset originale. Abbiamo iniziato con la ricerca e la rimozione dei duplicati dal dataset, assicurandoci che ogni audio fosse unico. Per ogni file audio, abbiamo calcolato e registrato diverse proprietà fondamentali:
Ampiezza che indica l'intensità del suono cercando di comprendere al meglio le dinamiche sonore presenti nelle registrazioni.
Numero dei canali in modo tale da determinare se i file audio fossero mono o stereo.
Bit Depth che rappresenta la risoluzione dei dati audio.
Frequenza di campionamento valori che rappresentano il numero di campioni al secondo.
Durata abbiamo misurato la lunghezza temporale di ciascuna registrazione audio, poiché la durata può variare significativamente tra i file.
Questi calcoli ci hanno fornito una panoramica dettagliata delle caratteristiche del dataset e hanno costituito la base per le fasi successive di pre-processing.

ORCID(s):

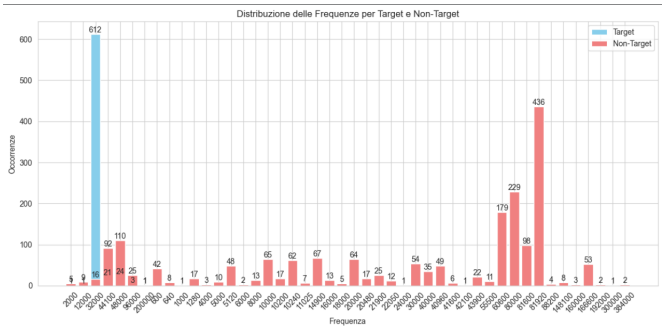

Figure 1: Grafico che mostra la distribuzione delle frequenze per i target e non target

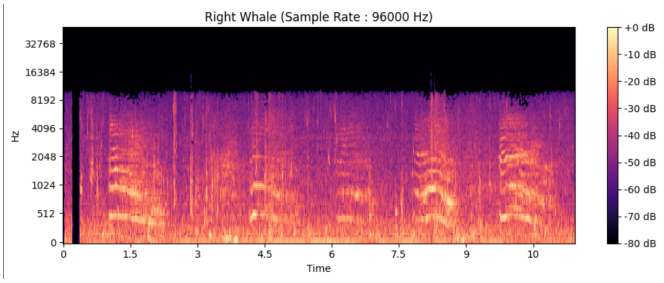

Figure 2: Grafico che mostra il ricampionamento a 96000 Hz

Spectral Centroid Mean	Spectral Bandwidth RMS	Standard Deviation	Skewness	Kurtosis
2585.774149596942	3545.4413815320154	0.22685716	0.29066348916425234	0.8920958313062881
3084.4909614115686	3774.6354112851127	0.10862513	-0.005662303418967790	-0.5694383533004904
2374.5189388747103	3489.082003330513	0.24789372	-0.27642137167332030	0.6855441477563251
2821.1442961363573	3698.989681063645	0.1568869	0.5795425154366439	1.3941836469104008
2649.5152568905096	3629.1870319804143	0.18223394	-0.08800976259032169	-0.08303857799051473
2705.7919751532813	3680.0295434591885	0.15845428	0.31480971820940057	-0.04925358302473359
3291.7467829495417	3868.125248656747	0.090920165	0.22710260526563683	0.4665684448611071
2636.179619396162	3111.6410886508465	0.082634844	-0.49164232971477534	0.4711344494880101
4411.094768089833	4513.896661066632	0.039391544	0.16898514976439629	0.4107143552745578
4392.384409699665	4516.783312927988	0.038142923	-0.022091882462507473	0.2585874714627416
4372.748284728635	4511.524573081631	0.036317635	0.024017650821905980	-0.2112536037762016
4419.744803021355	4523.684624415015	0.036478736	-0.008105412898455731	-0.1970419803699741
4488.227203380546	4519.846658069024	0.034533694	-0.011797477646959176	5.155498653023676
4427.584061932453	4520.654120391711	0.036474977	-0.013095787863391570	0.018800809752125947
4449.017725897941	4526.951856667042	0.03430764	-0.008333705983940917	-0.3454474886821952
3594.646116690734	3658.112066994046	0.03436695	-0.41628670212326060	-0.2309991262033932
4821.859800961919	4004.3551367044424	0.13998	0.030719191389485277	0.1163312469661881
4806.280412930921	3998.7664280100744	0.1430181	0.039713602651217796	0.1630207846583498
4799.999643858767	3996.7435239849615	0.14001971	-0.024103938715580482	0.0436778328983558
4786.276201287277	3994.794484946212	0.1431333	-0.000283439921752783	-0.011341295888555702
4811.120582660897	4015.11889660129	0.14109512	0.004171813044432215	0.0328810151306413
4831.188363881321	4008.2373816715826	0.13669215	0.007663820847590646	0.042180286373737461
4827.908276992953	4012.559671970624	0.13854362	-0.005246129976262719	0.013539777094082162
3828.871950378165	3221.256100432458	0.13215795	-0.26540642955109750	0.6787590830273849
5532.179289620277	4457.577169194846	0.19759922	-0.24129016821691918	0.0859738034862713
5678.509422501125	4414.533823133348	0.17195883	0.07026497653859424	0.1286496662594913
5718.164985088628	4427.411386340597	0.15495048	-0.16751030399718653	0.1859777555251192
5706.244325830912	4458.495159463547	0.16502361	-0.033861720907948464	-0.07374409076858646

Figure 3: Estrazione features numeriche da ogni audio

4. Pre-Processing

Una volta completata l'analisi iniziale, abbiamo proceduto con il pre-processing dei dati audio per prepararli all'addestramento del modello. Abbiamo normalizzato i segnali audio per uniformare l'ampiezza. La normalizzazione è stata effettuata per ridurre le variazioni estreme nelle ampiezze, che possono influire negativamente sulle prestazioni del modello. Questo processo ha reso i segnali audio più omogenei e comparabili tra loro. Per standardizzare la qualità delle registrazioni audio, tutti i file audio sono stati convertiti a una profondità di bit di 16-bit. Questa standardizzazione è importante perché garantisce che tutti i dati abbiano la stessa risoluzione, facilitando il lavoro del modello di machine learning. Abbiamo uniformato la frequenza di campionamento dei file audio a un valore standard di 96 kHz. Questo passaggio è stato essenziale per garantire che tutte le registrazioni fossero confrontabili e che il modello di machine learning ricevesse input omogenei. Successivamente, abbiamo suddiviso tutte le registrazioni audio in segmenti di 4 secondi. Questo passaggio è stato importante per creare segmenti di dati di durata uniforme, facilitando l'elaborazione e l'analisi. Per le registrazioni audio che erano più brevi di 4 secondi, abbiamo aggiunto periodi di silenzio alla fine per raggiungere la durata desiderata. Questi passaggi di pre-processing sono stati cruciali per assicurare che i dati fossero in un formato coerente e di alta qualità, pronti per l'addestramento dei modelli di machine learning.

5. Metodologia

La metodologia adottata per questo studio si basa sull'estrazione di caratteristiche numeriche fondamentali dai dati audio, che servono come input per i modelli di machine learning. L'estrazione delle caratteristiche (features) è un passaggio cruciale poiché queste caratteristiche rappresentano le informazioni rilevanti del segnale audio, facilitando la distinzione tra suoni generati dagli esseri umani e quelli prodotti dai pesci.
Queste caratteristiche sono state calcolate per ogni segmento audio e utilizzate come input per i modelli di machine learning.

Shannon Entropy	MFCC Mean Energy	Threshold Crossings	Silence Ratio
4.972139013520225	-265.2887	4596	0.6117578125
5.162510481166423	-289.32742	4892	0.6347135416666667
4.945728632377005	-258.72095	2814	0.576921875
4.977809664470404	-274.20407	3842	0.6766302083333333
5.108433070455767	-271.14337	4086	0.5879791666666667
5.08587532485647	-275.1605	3393	0.5999583333333334
4.921533705934605	-289.36353	6318	0.6757630208333333
4.563852402377499	-352.29816	4837	0.7201623801873257
3.8824375924014065	-299.21637	14264	0.8494296875
3.606892373368884	-295.20447	14290	0.8622708333333333
4.08600136184417	-301.32236	14352	0.8780208333333334
4.005849695741916	-299.83688	14688	0.8800651041666666
3.016340798303224	-302.8276	15048	0.8831953125
3.8830241862421255	-301.61517	14512	0.8804270833333333
4.453632873549875	-302.19574	13796	0.9000677083333334
3.9505824010089556	-381.441	11706	0.9058736527990793
4.368866542093834	-151.47128	31192	0.6609036458333334
4.415885331978914	-150.22757	31663	0.633453125
4.766979676360853	-151.09906	31119	0.6826796875
4.8031417822397735	-153.22322	30504	0.6631380208333333
4.504979200701108	-153.303	30981	0.6733515625
4.698741940801668	-158.10602	31449	0.6850885416666667
4.716511109543882	-156.47066	31524	0.6744713541666667
4.162518838445577	-242.52832	24386	0.7354369500855387
4.812127990838438	-147.06218	34404	0.4397552083333333
4.798854411173082	-153.49849	39759	0.476
4.567529334739545	-157.16608	42899	0.44590625
4.730484142652613	-160.07945	40188	0.44793489583333335

Figure 4: Estrazione features numeriche da ogni audio

	Unnamed: 0	precision	recall	f1-score	support
0	0	0.957492	0.994481	0.975636	906.000000
1	1	0.998526	0.988328	0.993401	3427.000000
2	accuracy	0.989615	0.989615	0.989615	0.989615
3	macro avg	0.978009	0.991405	0.984518	4333.000000
4	weighted avg	0.989946	0.989615	0.989686	4333.000000

Figure 5: Risultati per la classificazione binaria utilizzando KNN

6. Addestramento

Per fare l'addestramento è stato fatto lo split del dataset in 80% in set di train, 10% di validation, 10% test. Inoltre ci siamo assicurati che segmenti dello stesso file audio fossero inseriti nello stesso set. Per garantire risultati omogenei è stato applicato un algoritmo di oversampling sul set di train. Sono state fatte due tipi di classificazioni: una binaria e una multiclasse.L'obiettivo della classificazione binaria è quello di addestrare il modello affinchè possa riconoscere la classe di appartenenza di un file audio, mentre per la multiclasse deve saper riconoscere la sottoclasse di appartenenza.

7. Risultati

Per la classificazione binaria abbiamo utilizzato i seguenti modelli: Gradient Boosting, KNN, Logistic Regression, Random Forest, XGBoost ed SVM. I risultati sono stati positivi dato che quasi tutti i modelli hanno riportati valori tra il 90% e 99%. Per la classificazione multiclasse abbiamo utilizzato Gradient Boosting, KNN, Logistic Regression, Random Forest ed SVM. I risultati sono stati intorno al 40%

8. Conclusioni

Il progetto svolto parte da un dataset che comprende file audio marini e antropogenici. Per uno studio più approfondito è stato svolta una fase di pre-processing di tutto il dataset con l'obiettivo di ottenere risultati più precisi. Il sistema di classificazione produce risultati soddisfacenti dato che è in grado di riconoscere se in un file audio sono presenti suoni antropogenici.

9. Sviluppi futuri

In futuro questo progetto potrebbe essere svolto partendo da un dataset migliore utilizzando combinazioni di features numeriche per ottenere risultati più precisi, utilizzando più modelli per l'addestramento.