IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

Classification of Marine Mammal Species from Underwater Audio Using MFCC Features

L Ksheeraja[1], S. Poornima[2]

Dept. of CSE (Emerging Technologies), SRM Institute of Science and Technology,

Vadapalani Campus, Chennai, India[1]

Assistant Professor (Jr.G), Project Guide, Dept. of CSE (Emerging Technologies),

SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India[2]

Abstract: Passive acoustic monitoring produces far more underwater recordings than experts can review by hand, which motivates automatic species identification. This paper evaluates how much species information is carried by 13 time-averaged Mel-Frequency Cepstral Coefficients (MFCCs) on a 32-species marine mammal dataset of 1,697 recordings derived from the Watkins Marine Mammal Sound Database. A Support Vector Machine (SVM) and a Random Forest (RF) are trained on two representations, one vector per whole recording and one vector per 1-second segment, using the dataset's own train/test split. The RF is the better classifier in both cases, reaching 85.59 % accuracy on whole recordings (SVM: 73.53 %) and 78.56 % on segments (SVM: 74.16 %). Shorter segments give about eight times more training examples but do not improve the best model. K-Means clustering with 32 clusters gives an Adjusted Rand Index of 0.2542 and a normalised mutual information of 0.5703, showing real but overlapping species structure, and the first three MFCCs carry about 34 % of the RF importance. Fin whale, minke whale and leopard seal are classified almost perfectly, whereas sperm whale, Clymene dolphin and bearded seal remain difficult.

Keywords: Marine mammal classification, underwater acoustics, MFCC, support vector machine, random forest, K-Means clustering, passive acoustic monitoring.

I. INTRODUCTION

Sound is the main sensory channel of marine mammals. Baleen whales produce low-frequency moans and songs, toothed whales produce broadband clicks and tonal whistles, and seals produce pulses, trills and barks. Passive acoustic monitoring (PAM), in which hydrophones record the soundscape without disturbing the animals, has become a standard way to study presence, distribution and behaviour [2], but it generates volumes of audio that are slow and costly to review manually. Automatic species classification is therefore an active topic in applied machine learning and bioacoustics.

MFCCs summarise the shape of the short-time spectrum on a perceptually motivated mel scale. Originally developed for speech recognition [11], they offer a light and interpretable baseline when combined with classical classifiers. It is still unclear how well such a baseline separates many species at once, whether describing a recording by one vector is better than describing it by many short segments, and whether the MFCC space has a natural structure that matches species.

This paper addresses these questions on a 32-species dataset. Its contributions are:

- a documented MFCC baseline comparing an RBF-kernel SVM and a Random Forest on whole recordings and on 1-second segments;
- a per-species analysis that identifies the easiest and hardest classes;
- a label-free view of the same features through K-Means clustering and PCA, measured with ARI, NMI and silhouette; and
- an analysis of which MFCC coefficients the Random Forest relies on.

II. RELATED WORK

Supervised classification of marine mammal sounds has been studied with several feature and model combinations. Zhao et al. (2023) [4] extracted MFCC features and combined SVM, k-nearest neighbour, decision tree and naive Bayes classifiers through stacking; stacking improved recognition by 3.30 percent and worked for all six species considered. Three studies from 2024 used acoustic features with a random forest: Cominelli et al. [5] combined them with UMAP to explore marine soundscapes and found that the features separated species and locations, although low and high frequencies

were harder to classify; Machado et al. [6] found duration and frequency features of whistles informative for delphinid classification; and Rankin et al. [7] showed that the BANTER framework classifies beaked whale echolocation pulses effectively. Hamard et al. (2024) [8] applied Faster R-CNN to spectrograms for multi-species detection and classification but focused on two species, and Solsona-Berga et al. (2024) [9] found that taxonomic family information improved the classification of difficult beaked whale events.

Unsupervised methods have also been explored. Liang et al. (2025) [10] combined acoustic features with unsupervised clustering to classify five cetacean vocal-signal categories, Merchan et al. (2024) [11] used wavelet features with PaCMAP and HDBSCAN to identify individual Greater Caribbean manatees, and Frasier (2021) [12] used clustering to discover signal categories in echolocation clicks before classification. Nadir et al. (2020) [13] showed that fusing MFCC with acoustic binary patterns improved classification accuracy.

Most of these studies address few species or one signal type, use different methods, and rarely combine species classification with acoustic grouping in the same feature space. The present work studies both on 32 species and additionally measures the effect of segment length.

III. DATASET AND PREPROCESSING

A. Dataset

The Sangkrishna Underwater_Audio dataset [24], derived from the Watkins Marine Mammal Sound Database [23], contains **1,697 labelled WAV recordings of 32 species**. It ships with a fixed split of 1,357 training and 340 test recordings, used unchanged here. No values are missing, no rows are duplicated and no file failed feature extraction. Recording length is very uneven: the median is 1.70 s and 75 % of files are shorter than 3.23 s, but the mean is 10.43 s because 43 files exceed one minute (Fig. 1; the longest lasts 21 minutes; mostly fin, humpback and sperm whale). Only 31 species appear in the test set, since Weddell seal has no test recording, and test recordings per species range from 2 (leopard seal) to 23 (spinner dolphin).

TABLE II
DATASET SUMMARY

Property	Value
Recordings	1,697 (1,357 train / 340 test)
Species	32 (31 in test set)
Native sampling rates	10.2 kHz to 80 kHz
Duration range	0.05 s to 1,260 s
Mean / median duration	10.43 s / 1.70 s
Std. dev. of duration	71.24 s

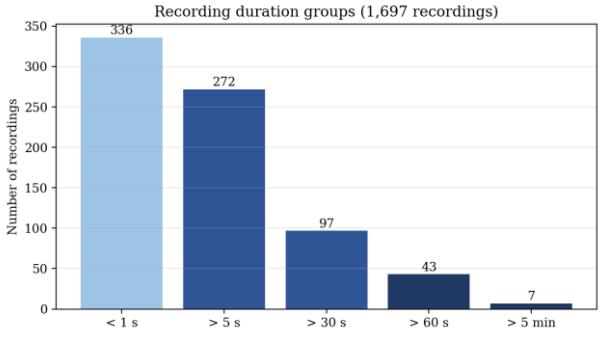


Fig. 1. Number of recordings grouped by duration; a few very long files dominate the total audio.

B. Preprocessing

Audio was decoded from raw WAV bytes with soundfile, multi-channel files were averaged to mono, and every recording was resampled to 16 kHz so that mel bands are comparable across native rates. Labels were integer-encoded and features standardised with a scaler fitted on the training data only, which prevents test information from leaking into

the scaling statistics. For the segment-level experiment each segment inherits the split of its parent recording, so no recording contributes to both sets.

IV. METHODOLOGY

The system is organised as five sequential layers, from raw recordings to evaluation and interpretation (Fig. 2). Both feature representations share the data and preprocessing layers and are passed to the same supervised and unsupervised models.

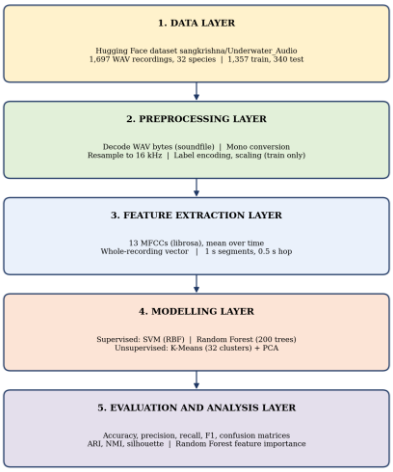


Fig. 2. System architecture: data, preprocessing, feature extraction, modelling and evaluation layers.

A. MFCC Feature Extraction

MFCCs were computed with librosa [21] at 16 kHz using a 2,048-sample window (128 ms), a hop of 512 samples (32 ms), 128 mel filters and 13 retained coefficients. Each frame is windowed, transformed to a power spectrum, passed through triangular filters spaced on the mel scale [12], log-compressed and decorrelated with a discrete cosine transform. A common form of the mel scale is

$$m = 2595 \log_{10}(1 + f / 700) \quad (1)$$

where f is frequency in Hz.

B. Feature Representations

Two 13-dimensional representations were built. In the **whole-recording** representation the coefficients are averaged over all frames, giving 1,697 vectors. In the **segment-level** representation each recording is cut into 1-second windows with a 0.5-second hop and the coefficients are averaged per window; recordings shorter than one second are zero-padded and at most 100 evenly spaced windows are kept per recording. This gave **17,182 segments** (14,159 training and 3,023 test), about 10 per recording on average.

C. Classifiers

The **SVM** [13] uses an RBF kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2) \quad (2)$$

with one-versus-one multi-class voting, class_weight = "balanced" and the scikit-learn [20] defaults $C = 1$ and γ = "scale". The **Random Forest** [14] uses 200 trees, balanced class weights and seed 42; it splits on Gini impurity $G = 1 - \Sigma p_k^2$ and yields impurity-based feature importances. Both classifiers use identical features, scaling and split, and no hyperparameter tuning or cross-validation was performed.

D. Clustering and Metrics

K-Means [15] with $k = 32$ (the number of species), ten initialisations and seed 42 was run on the standardised features without labels, and PCA [19] provided two-dimensional views. Classification is scored with accuracy and weighted precision, recall and F1 (macro F1 where noted). Clustering is scored with the Adjusted Rand Index (ARI) [16],

Normalized Mutual Information (NMI) [18] and the silhouette score [17] on a random sample of 1,000 points. Whole-recording scores use 340 test recordings; segment-level scores use 3,023 test segments, each scored individually.

V. RESULTS

A. Classification Performance

Table III and Fig. 3 give the main results. The Random Forest outperformed the SVM in both representations. On whole recordings it classified 291 of 340 recordings correctly versus 250 for the SVM, a margin of 12.06 points. At segment level the margin fell to 4.40 points (2,375 versus about 2,242 of 3,023 segments).

TABLE III
TEST-SET CLASSIFICATION RESULTS

Input	Model	Acc. (%)	W-Prec.	W-Rec.	W-F1
Whole rec.	SVM	73.53	0.7805	0.7353	0.7251
Whole rec.	RF	**85.59**	0.8630	0.8559	0.8532
1-s segments	SVM	74.16	0.8136	0.7416	0.7294
1-s segments	RF	78.56	0.8546	0.7856	0.7889

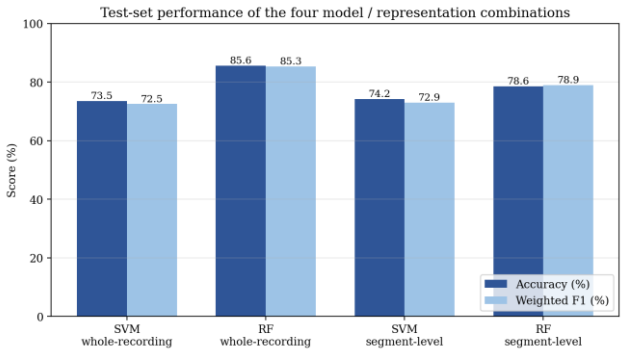


Fig. 3. Accuracy and weighted F1 (%) of the four model and representation combinations.

The confusion matrix of the Random Forest on whole recordings (Fig. 4) has a strong diagonal with few, light off-diagonal cells. The SVM matrix is more scattered; for example some Fraser's and Risso's dolphin recordings were assigned to melon-headed whale, and sperm-whale recordings were spread across several dolphin and pilot-whale classes.

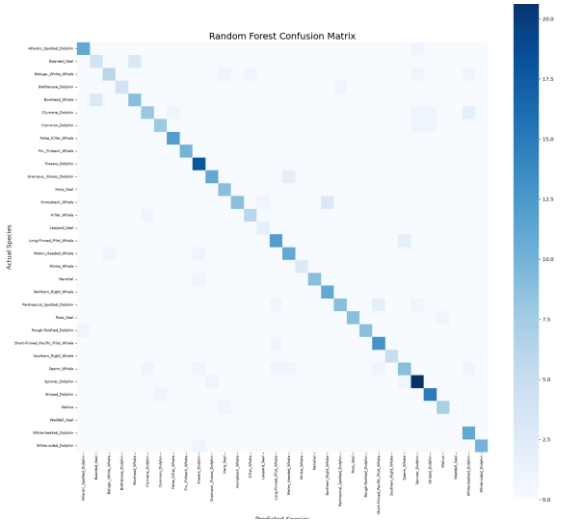


Fig. 4. Random Forest confusion matrix on the 340 test recordings (whole-recording features).

 This work is licensed under a Creative Commons Attribution 4.0 International License

B. Effect of Segmentation

Segmentation increased training examples from 1,357 to 14,159 but did not help the best model: SVM accuracy changed by +0.63 points while RF accuracy dropped by 7.03 points. Three factors explain this. First, the evaluation units differ: fin whale alone supplies 785 test segments (26 %) and fin, humpback and sperm whale together make up about 51 % of the segment test set, so segment accuracy is weighted toward species with long recordings. Second, a one-second window may contain only noise or a fragment of a call. Third, segments from one recording are highly correlated, so the effective gain in information is smaller than the 10-fold increase in examples suggests. The two accuracies are therefore not strictly comparable.

C. Per-Species Results

Table IV lists the easiest and hardest species. Fin whale reached an F1-score of 1.00 in every experiment and minke whale in all three that report per-class results. Baleen whales and seals therefore tend to have distinctive MFCC profiles. The hardest classes were sperm whale, Clymene dolphin, bearded seal, beluga and common dolphin, several of which are dolphins whose calls overlap acoustically with other species. At segment level the RF had very low precision for bearded seal (0.10) and the Clymene dolphin recall was only 0.17 (Fig. 5).

TABLE IV
SELECTED PER-SPECIES F1-SCORES

Species	SVM whole	SVM seg.	RF seg.
Fin whale	1.00	1.00	1.00
Minke whale	1.00	1.00	1.00
Leopard seal	0.80	1.00	1.00
Humpback whale	0.76	0.97	0.92
Sperm whale	0.38	0.29	0.59
Clymene dolphin	0.47	0.29	0.28
Bearded seal	0.53	0.29	0.16
Beluga	0.62	0.22	0.30
Common dolphin	0.59	0.12	0.43

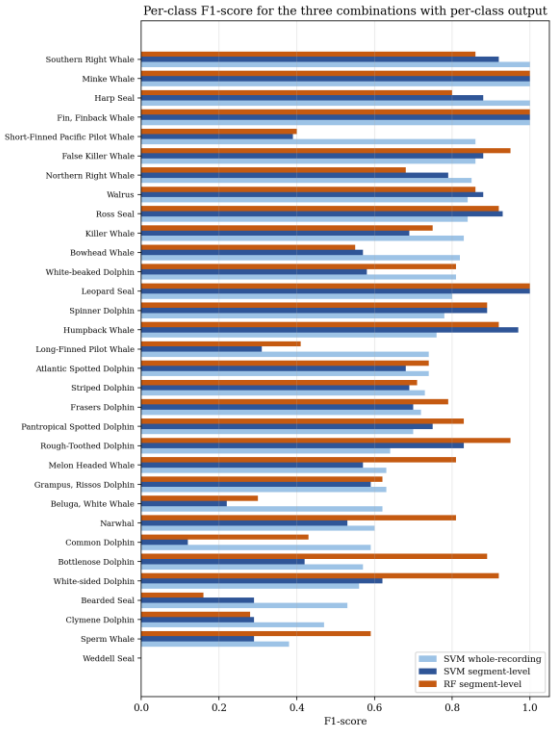


Fig. 5. Per-class F1-score for the whole-recording SVM and the segment-level SVM and Random Forest.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

D. Structure of the Feature Space

K-Means with 32 clusters on the 1,697 whole-recording vectors gave the scores in Table V. The first two principal components explain 57.1 % of the variance, so Fig. 6 shows most but not all of the structure. Some compact, isolated groups appear next to a large dense central region where clusters overlap, and some species such as fin whale fall almost entirely into one cluster while others are spread across several. This agrees with the supervised results, where distinctive species were classified well and overlapping dolphins were confused.

TABLE V
K-MEANS EVALUATION (WHOLE RECORDINGS)

Metric	Value	Interpretation
Adjusted Rand Index	0.2542	Above chance, modest
Normalized Mutual Information	0.5703	Moderate shared information
Silhouette score	0.2079	Weakly separated

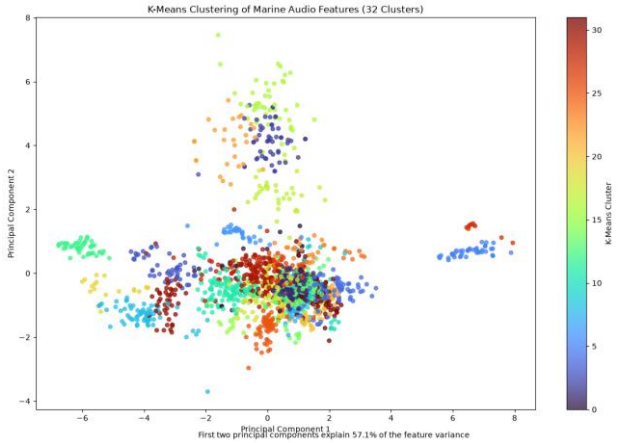


Fig. 6. K-Means clusters (k = 32) of the whole-recording MFCC features in PCA space.

The same procedure on all 17,182 segment vectors (Fig. 7) is more compact in places. On the 3,023 test segments the ARI between species and clusters was 0.4063, but this is computed on a different, species-skewed sample and is not directly comparable with 0.2542. Grouping test segments by cluster, the RF was 100 % accurate in 12 of the 30 clusters containing test segments, whereas accuracy was lowest in clusters 9 (8.9 %), 30 (20.4 %) and 14 (33.3 %), the regions where species overlap and most errors occur.

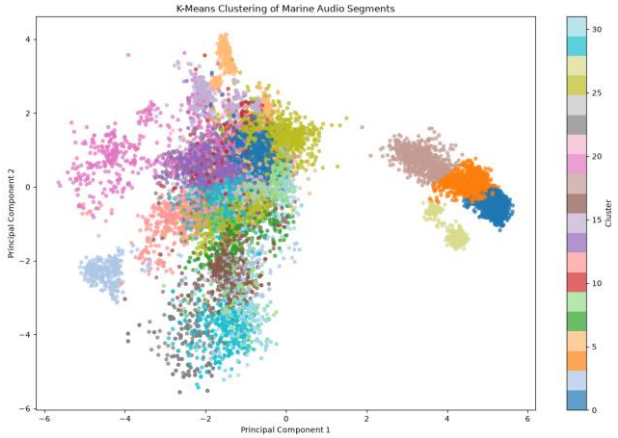


Fig. 7. K-Means clusters (k = 32) of the segment-level features in PCA space.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

E. MFCC Coefficient Importance

In the segment-level forest MFCC 2 was the most important coefficient (0.129), followed by MFCC 1 (0.109) and MFCC 3 (0.106) (Fig. 8). The first three account for about 34 % of the total importance and the first six for about 59 %, so the broad spectral envelope carries most of the information, while MFCC 7–13 each contribute 0.053–0.065 (about 41 % together). Importance is spread fairly evenly, with no single dominant feature. These impurity-based scores indicate reliance by the forest, not biological meaning, and can be biased for correlated features.

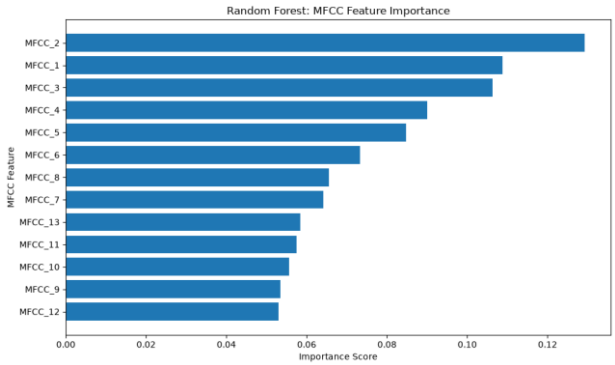


Fig. 8. Random Forest importance of the 13 MFCC coefficients (segment-level model).

VI. DISCUSSION AND LIMITATIONS

With only 13 time-averaged MFCCs and default settings, the best configuration (RF on whole recordings) classified 85.6 % of test recordings correctly across 31 evaluated species, so MFCCs carry substantial species information and the approach is fast and interpretable. The SVM was left at default C and γ, and a tuned SVM could narrow the gap. Several limitations apply:

- all audio was resampled to 16 kHz, so content above 8 kHz, including many dolphin clicks and whistles, is discarded;
- only the mean of each coefficient is kept, so variance and temporal call structure are lost;
- one train/test split without tuning or cross-validation gives no confidence intervals;
- segment predictions were scored per segment, not per recording, and long recordings dominate that score;
- test classes are uneven (2 to 23 recordings) and Weddell seal has none, so macro averages include a class that cannot be evaluated; and
- recordings of one species may share a site or hydrophone, so the model may partly learn the recording environment.

VII. CONCLUSION AND FUTURE WORK

This work evaluated a 13-coefficient MFCC baseline for classifying 32 marine mammal species. The Random Forest beat the SVM in both representations (85.59 % versus 73.53 % on whole recordings; 78.56 % versus 74.16 % on segments), 1-second segmentation did not improve the best model, K-Means showed moderate but overlapping species structure (ARI 0.2542, NMI 0.5703, silhouette 0.2079), and the low-order coefficients MFCC 1–3 were the most important. Fin whale, minke whale and leopard seal were classified almost perfectly, while sperm whale, Clymene dolphin and bearded seal were consistently difficult.

Future work includes aggregating segment predictions per recording by majority vote or averaged probabilities so both representations are scored on the same unit; adding standard deviation, delta coefficients and spectral features; using higher sampling rates and log-mel or constant-Q spectrograms; tuning hyperparameters with cross-validation over several splits; training convolutional networks or pretrained audio embeddings; and splitting by recording source to rule out environment confounds.

ACKNOWLEDGMENT

The authors thank their course faculty and project guide, S. Poornima, Assistant Professor, Department of Computer Science and Engineering (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus, for her mentorship and support, and Dr. P. Chitra, Professor and Head of the Department, for her encouragement.

REFERENCES

[1] B. Zhao, W. Cai and J. Zhu, "Acoustic identification of marine mammals based on Stacking classification fusion," *J. Hangzhou Dianzi Univ.*, vol. 43, no. 4, pp. 7–13, 2023 (in Chinese).

[2] S. Cominelli, N. Bellin, C. D. Brown, V. Rossi and J. Lawson, "Acoustic features as a tool to visualize and explore marine soundscapes," *Ecology and Evolution*, vol. 14, no. 2, e10951, 2024.

[3] Q. Hamard, M.-T. Pham, D. Cazau and K. Heerah, "A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data," *Ecological Informatics*, vol. 84, 102906, 2024.

[4] R. B. Machado, T. O. S. Amorim, F. de Castro and A. Andriolo, "Use of whistles for acoustic classification of delphinids in the Western South Atlantic Ocean," *J. Acoust. Soc. Am.*, vol. 156, no. 2, pp. 1070–1080, 2024.

[5] S. Rankin *et al.*, "Open-source machine learning BANTER acoustic classification of beaked whale echolocation pulses," *Ecological Informatics*, vol. 80, 102511, 2024.

[6] A. Solsona-Berga *et al.*, "Machine learning with taxonomic family delimitation aids in the classification of ephemeral beaked whale events in passive acoustic monitoring," *PLOS ONE*, vol. 19, no. 6, e0304744, 2024.

[7] Y. Liang, Y. Wang, F. Chen, H. Yu, F. Ji and Y. Chen, "Automatic detection and unsupervised clustering-based classification of cetacean vocal signals," *Applied Sciences*, vol. 15, no. 7, 3585, 2025.

[8] F. Merchan, K. Contreras, H. Poveda, H. M. Guzman and J. E. Sanchez-Galan, "Unsupervised identification of Greater Caribbean manatees using scattering wavelet transform and hierarchical density clustering," *Frontiers in Marine Science*, vol. 11, 1416247, 2024.

[9] K. E. Frasier, "A machine learning pipeline for classification of cetacean echolocation clicks in large underwater acoustic datasets," *PLOS Comput. Biol.*, vol. 17, no. 12, e1009613, 2021.

[10] M. Nadir, S. M. Adnan, S. Aziz and M. U. Khan, "Marine mammals classification using acoustic binary patterns," *Archives of Acoustics*, vol. 45, no. 4, pp. 721–731, 2020.

[11] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[12] S. S. Stevens, J. Volkmann and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, 1937.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[16] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[17] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[18] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[19] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Phil. Trans. R. Soc. A*, vol. 374, 20150202, 2016.

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[21] B. McFee *et al.*, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf.*, 2015, pp. 18–25.

[22] Q. Lhoest *et al.*, "Datasets: A community library for natural language processing," in *Proc. EMNLP 2021: System Demonstrations*, 2021.

[23] L. S. Sayigh, N. Daher, K. Allen, H. Rosenthal *et al.*, "The Watkins Marine Mammal Sound Database: An online, freely accessible resource," *Proc. Meetings on Acoustics*, vol. 27, 040013, 2016.

[24] sangkrishna, "Underwater_Audio" dataset, Hugging Face Hub (sangkrishna/Underwater_Audio). Accessed 2026.