Ecological Informatics 97 (2026) 103910

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Sex and age determination in European lobsters using AI-Enhanced bioacoustics

Feliciano Pedro Francisco Domingos [a,*], Isibor Kennedy Ihianle [a], Omprakash Kaiwartya [a], Ahmad Lotfi [a], Nicola Khan [c], Nicholas Beaudreau [d], Amaya Albalat [b], Pedro Machado [a]

[a] Nottingham Trent University, Department of Computer Science, Erasmus Darwin Building, Clifton, NG11 8NS, Nottingham, United Kingdom
[b] University of Stirling, Pathfoot Building, Stirling, FK9 4LA, Scotland, United Kingdom
[c] Ace Aquatec, 1 Water's Edge, Camperdown Street, Dundee, DD1 3HY, Scotland, United Kingdom
[d] MERINOV, Quebec Centre for Innovation in Aquaculture and Fisheries, 96 Montée de Sandy Beach, Suite 2.05, Gaspé, Quebec City, G4X 2V6, Quebec, Canada

ARTICLE INFO

Dataset link: https://doi.org/10.17632/WSM4VFXVSF, https://data.mendeley.com/datasets/wsm4vfxvsf, https://github.com/N0736086/lobster_soundscodes, https://github.com/N0736086/stacking-ensemble-pipeline/blob/main/stack_ensemble_pipelie.py

Keywords:
Lobster bioacoustics
Artificial intelligence
Machine learning
Deep learning
Age and sex classification
Aquaculture management
Passive acoustic monitoring

ABSTRACT

Monitoring aquatic species presents considerable challenges due to their elusive nature and complex habitats. Consequently, the development and application of innovative, non-invasive approaches, such as Passive Acoustic Monitoring (PAM), are paramount for effective ecological assessment and management. The present study addresses these challenges by focusing on the acoustic emissions of Homarus gammarus (European lobster), a key representative species of rocky benthic environments that underpins valuable local fisheries and aquaculture ventures. A comprehensive understanding of lobster habitats, welfare, reproduction, sex, and age is critical for robust aquaculture management, ecological research, conservation strategies, and sustainable fisheries. While bioacoustic emissions have been successfully employed to classify various aquatic species using Artificial Intelligence (AI) models, such as fish, the present research specifically leverages lobster bioacoustics to classify European lobster by age group (juvenile and adult) and sex (male and female). Despite lacking vocal cords, different lobster species produce characteristic sounds, including stridulation (European spiny lobster, Panulirus elephas; Caribbean spiny lobster, Panulirus argus), buzzing or carapace vibrations (European lobster, Homarus gammarus; American lobster, Homarus americanus), rattling (tropical spiny lobster, Panulirus ornatus), and clicking or snapping sounds. These acoustic signals are amenable to classification using advanced computational AI models. The dataset was collected at Johnshaven in Scotland, at a local lobster facility operated by Murray McBay and Company. Hydrophones were installed underwater in concrete tanks to record lobster sounds. We evaluate the performance of Deep Learning (DL) models, specifically One-Dimensional Convolutional Neural Networks (1D-CNN) and One-Dimensional Deep Convolutional Neural Networks (1D-DCNN), and six commonly used Machine Learning (ML) models (Support Vector Machine, k-Nearest Neighbours, Naive Bayes, Random Forest, Extreme Gradient Boosting, and Multi-Layer Perceptron) for age and sex classification. Mel-Frequency Cepstral Coefficients (MFCCs) were used as baseline features for all models, while a Multi-Feature Fusion (MFF) approach was employed to confirm the consistency of classification performance. MFCCs are well established for robust audio feature extraction, and both MFCC and MFF yielded consistent classification results. Most models achieved classification accuracies exceeding 97% for adult versus juvenile differentiation, with the exception of Naive Bayes (91.31%). For sex classification, all models except Naive Bayes exceeded 93.23% accuracy with MFCC features. These results highlight the strong potential of supervised ML and DL approaches to extract age- and sex-related information from lobster sounds. Overall, this research demonstrates a promising non-invasive approach for lobster monitoring, conservation, and management in aquaculture and fisheries, supporting the development of real-world edge-computing applications for PAM of underwater species.

* Corresponding author.
E-mail address: feliciano.domingos@ntu.ac.uk (F.P.F. Domingos).

https://doi.org/10.1016/j.ecoinf.2026.103910
Received 27 July 2025; Received in revised form 29 June 2026; Accepted 30 June 2026
Available online 13 July 2026
1574-9541/

1. Introduction

Effective lobster aquaculture and fisheries management, ecological research, and conservation strategies require a detailed understanding of habitat use, welfare, reproductive conditions, and the accurate identification of sex and age (Barroso et al., 2023; Looby et al., 2023). As highlighted by Hinchcliffe et al. (2022), the European lobster (Homarus gammarus) industry is currently transitioning from traditional fisheries to land-based aquaculture and sea-ranching to meet high market demand. However, this transition is hindered by high operational costs and the need for intensive monitoring of individual welfare and growth. Automated, non-invasive systems such as the Artificial Intelligence (AI)-enhanced bioacoustic platform proposed in this study are essential to overcome these technical bottlenecks, providing the real-time data required for sustainable large-scale production. Bioacoustics has emerged as a powerful tool to detect and monitor behavioural and ecological traits, with demonstrated success across avian and aquatic taxa, including fish and invertebrates (Bardeli et al., 2010; Noda et al., 2016; Malfante et al., 2018). Lobsters produce diverse acoustic signals depending on species, including stridulation (European spiny lobster (Panulirus elephas)), buzzing or carapace vibrations (American lobster (Homarus Americanus), European lobster (Homarus Gammarus)), rattling (Tropical spiny lobster (Panulirus ornatus)), and clicking or snapping sounds (Patek, 2002). These acoustic emissions can be systematically classified using AI models (Briones-Fourzán et al., 2014). The present study focuses on classifying buzzing sounds (carapace vibrations) produced by Homarus Gammarus to distinguish individuals by age class (adult vs. juvenile) and sex (male vs. female).

Sound production in lobsters arises from mechanical actions: Panulirus elephas and Panulirus ornatus generate stridulation by rubbing or contracting body parts, whereas Homarus Gammarus emit vibrations of the carapace in the 80–250 Hz range (Patek, 2002; Staaterman et al., 2010; Kerkhofs, 2024). Such bioacoustic emissions are vital, serving functions including predatory avoidance, defence, aggression, courtship, and mating (Domingos et al., 2024; Looby et al., 2023; Barroso et al., 2023; Weilgart, 2018). More than 900 species of fish and invertebrates are known to produce active sounds, reflecting the centrality of acoustic communication in marine ecosystems (Looby et al., 2023; Jézéquel et al., 2020b). For lobster aquaculture and fisheries, reliable and efficient methods for sex and age determination are critical to optimise yields, welfare, and sustainability (McGarvey et al., 2015). Traditional manual methods are time-consuming, labour intensive, and prone to human error, motivating exploration of DL approaches such as 1D-CNN, 1D-DCNN, alongside established Machine Learning (ML) models, for classification of Homarus Gammarus acoustic signals.

To situate our study within broader management contexts, it is essential to note that effective lobster management extends beyond sex and age identification. Population dynamics are shaped by distribution patterns, life-history traits, spawning behaviours, and environmental pressures, including climate change (Araújo and Guilhaumon, 2018; Phillips et al., 2015; Radhakrishnan et al., 2020). Emerging threats, such as contamination by microplastics and heavy metals, further impact lobster health and seafood safety (Hwang et al., 2019; Yadav et al., 2024). Fig. 1 illustrates how bioacoustics-based sex and age classification integrates within this wider ecological and policy framework. In addition, regulatory frameworks such as European Union (EU) fisheries policies are central to stock sustainability, though they face ongoing scrutiny regarding ecological effectiveness and socio-economic trade-offs (European Commission, Directorate-General for Maritime Affairs and Fisheries, 2025; Council of the European Union, 2025). Thus, management decisions are embedded in governance as much as in biology.

Our contribution addresses this interdisciplinary challenge by applying AI models to classify lobster bioacoustics. Recent advances in ML and DL have demonstrated strong potential for detecting and classifying acoustic signatures of fish and crustaceans, enabling improved resource monitoring (Niu et al., 2023; Bharath et al., 2025; Rountree et al., 2023). We argue that AI-enhanced bioacoustic monitoring offers a scalable, non-invasive method for lobster aquaculture and conservation, complementing ecological, socio-economic, and policy dimensions.

The main contributions of this article are as follows: (i) classification of lobster bioacoustics to determine age and sex; (ii) comparative evaluation of traditional ML and modern DL models, including four 1D-CNN and four 1D-DCNN architectures; (iii) analysis of model performance metrics to identify the most effective approaches; and (iv) a critical synthesis of related works on lobster acoustics and AI classification.

2. Related works

While research into Homarus Gammarus aquaculture has spanned 150 years, recent technical developments have focused on optimising environmental parameters and dietary requirements (Hinchcliffe et al., 2022). Despite these advances, the 'biological requirement' for low-stress monitoring remains a significant challenge. Bioacoustic analysis offers a solution to the limitations of physical handling, which, as noted by Hinchcliffe et al. can lead to increased mortality and behavioural stress in high-density tank environments.

2.1. Sound production and detection in lobsters

The understanding of how marine invertebrates produce and perceive sound remains limited (Taylor and Patek, 2010; Jézéquel et al., 2020b). In both Homarus Americanus and Homarus Gammarus, vibrations and sound are generated by opposing remotor and promotor muscle contractions beneath the second antenna (Kerkhofs, 2024). Using auditory evoked potential (AEP) methods, Jézéquel et al. (2021) confirmed that Homarus Americanus detect low-frequency sounds between 80 and 250 Hz, with peak neurophysiological sensitivity occurring in the 80–120 Hz range. The abdominal vibrations of Homarus Americanus typically exhibit a base frequency of approximately 180 Hz, consistent with earlier reports (Kerkhofs, 2024). For sound detection, external cuticular hairs appear to play a key role, while evidence suggests that crustacean statocysts are not primary auditory organs (Jézéquel et al., 2021). Spiny lobsters (*Panulirus* spp.) also produce sounds through stick–slip mechanisms, where antennal bases rub against the antennular plate (Patek, 2002; Staaterman et al., 2010).

2.2. Behavioural context of acoustic emissions

Lobster acoustic emissions are closely tied to behavioural states (Jézéquel et al., 2018; Looby et al., 2023; Domingos et al., 2024). Stressed Homarus Gammarus individuals vibrate their carapace, producing low-frequency “buzzing” sounds similar to Homarus Americanus (Jézéquel et al., 2018, 2021, 2020b). During feeding, Homarus Gammarus emit broadband, transient “rattles”, comparable to those of tropical spiny lobsters (*Palinurus longipes*, *P. argus*) (Jézéquel et al., 2018). Spiny lobsters also stridulate in anti-predatory contexts (Bouwma and Herrnkind, 2009; Kikuchi et al., 2015). Male Homarus Gammarus produce low-frequency buzzing during agonistic encounters, though tank-based hydrophones capture only ∼15% of such signals due to attenuation (Jézéquel et al., 2020b). Accelerometer-based detection has been employed to overcome the challenges of low signal capture.

2.3. Propagation and environmental considerations

Under natural low-noise conditions, spiny lobster sounds propagate over several kilometres, with the largest individuals exceeding 3 km (Jézéquel et al., 2020a). However, tank environments introduce reverberation and attenuation, which complicates analysis. Artificial sound source experiments demonstrated the importance of characterising tank acoustics before interpreting lobster behaviour from such recordings (Jézéquel et al., 2018). These environmental factors highlight both the potential and limitations of laboratory-based bioacoustic studies.

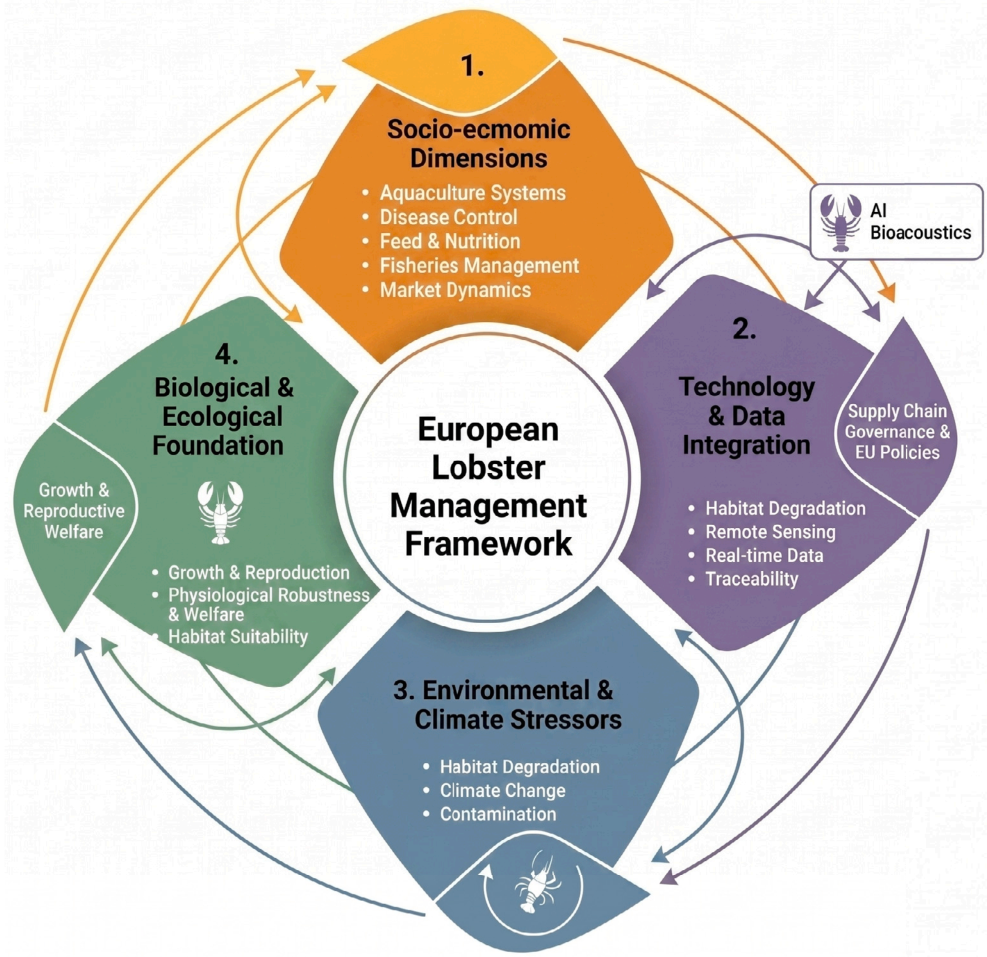


Fig. 1. Conceptual diagram of lobster management context showing how AI-enhanced bioacoustics for sex and age determination fits within broader ecological and policy considerations.

2.4. Applications of AI and signal classification

Beyond ecological studies, ML and DL approaches are increasingly used for acoustic classification to enhance research effectiveness (Malfante et al., 2018). Discriminative ML models, such as RF and Support Vector Machine (SVM), have achieved high accuracy in fish sound classification (up to 96.9%) (Malfante et al., 2018). ML techniques are also being explored for undersea communication and environmental monitoring, offering noise reduction and robust data handling (Domingos et al., 2024; Islam et al., 2025).

The selection of 1D-CNN architectures and classical ML methods for the present study is further supported by their demonstrated effectiveness in diverse environmental and remote sensing applications. As evidenced in recent literature, such as the comparative study by Günen (2022), these models provide a reliable framework for high-dimensional data classification, offering a robust balance between automated feature extraction and generalisability across varying datasets.

DL architectures, including Convolutional Neural Network (CNN)s and Deep Belief Network (DBN)s, have shown superior performance in underwater acoustic target classification, achieving accuracies above 94% (Yue et al., 2017). While MFCCs remain standard features for ML models (Davis and Mermelstein, 1980a), CNNs can directly process raw signals or Fast Fourier Transform (FFT)-based features, reducing reliance on handcrafted features. For lobster acoustics specifically, 1D-CNN and dilated CNNs (1D-DCNN) offer the ability to capture temporal dynamics directly from raw or minimally processed signals (Huang et al., 2019).

2.5. Towards predictive modelling in lobster bioacoustics

Recent advances emphasise predictive performance over inference in ecological modelling (Breiman, 2001b). Ensemble methods such as bagging and boosting improve robustness, with RF reducing variance through bootstrapped sampling (Breiman, 2001a, 1996a) and XGBoost offering efficient gradient boosting with regularisation (Chen and Guestrin, 2016). Stacking methods further enhance accuracy by combining diverse learners (e.g., SVM, RF, XGBoost, CNNs) and training a meta-learner on out-of-fold predictions (Wolpert, 1992; Breiman, 1996b; Hastie et al., 2001). These approaches hold strong potential for automated lobster classification, with applications in fisheries management, aquaculture monitoring, and conservation planning.

3. Materials and methodology

The study employed a structured methodology encompassing acoustic acquisition, preprocessing, feature engineering, and a comparative evaluation of machine learning (ML) and deep learning (DL) architectures. The workflow is summarised in the research methodology flowchart depicted in Fig. 2.

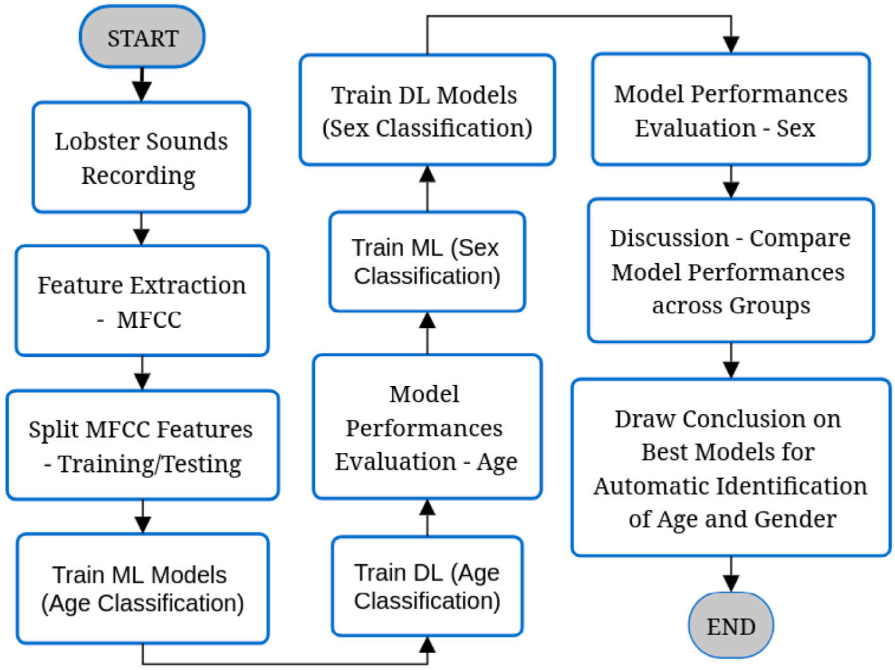


Fig. 2. Research methodology flowchart depicting the stages undertaken in this study.

Table 1
Summary of lobster acoustic dataset by sex and age categories.

Category	Number of individuals	Recording duration (s)
Juvenile female lobsters	6	1200
Juvenile male lobsters	6	1200
Adult male lobsters	6	3207
Adult female lobsters	6	1700
Total	**24**	**7307**

3.1. Material specifications and experimental setup

Underwater recordings were acquired at the Johnshaven Lobster Shop (Scotland) between 16–18 June 2024. The study species, Homarus Gammarus (European lobster), was housed in concrete tanks supplied with continuously circulating seawater and aeration. Individuals were restrained with rubber bands to prevent aggression as recommended in Shabani et al. (2009). Sex and age were visually identified by experts based on gonopore position and pleopod morphology (Mente, 2008).

The hardware configuration consisted of an Aquarian Audio H2d hydrophone connected to a Raspberry Pi 3 placed underwater, as depicted in Fig. 3. Real-time data acquisition was managed via the `PyAudio` library. Subsequent data processing and model training were performed on a Lenovo ThinkPad X1 Carbon Gen 13 which communicated with the Raspberry Pi via a router to which the raspberry is connected. Detailed hardware specifications are provided in Appendix 9.

3.2. Dataset summary and preprocessing

A total of 7307 s of recordings were collected from 24 individuals, stratified into four demographic groups as summarised in Table 1. Continuous recordings were segmented into 1-second intervals to serve as the primary units for classification.

Acoustic acquisition was performed at 22.05 kHz. Preprocessing involved a band-pass filter (50–8000 Hz) to mitigate low-frequency flow noise and high-frequency electronic artefacts. The full preprocessing pipeline is detailed in Table 2.

Table 2
Audio preprocessing and feature extraction pipeline.

Step	Description
High-pass filtering	Removal of DC offset using a cut-off frequency between 20 and 50 Hz
Band-pass filtering	Retention of frequencies between 50 and 8000 Hz to capture lobster sound energy
Segmentation	Signal divided into fixed windows of 1 s
SNR screening	Low-energy segments discarded based on signal-to-noise ratio thresholds
Feature extraction	Multi-Feature Fusion (MFF), MFCC, Zero Crossing Rate(ZCR), etc.) extracted from each 1s segment
Normalisation	Z-score standardisation applied prior to model training

3.3. Experimental design: Baseline vs. Refined fusion

The research was conducted in two stages. First, a baseline evaluation was performed using only MFCCs. Subsequently, a refined Multi-Feature Fusion (MFF) approach was developed to better characterise the time–frequency structure of the target *H. gammarus* buzzing/carapace-vibration signals. The refined feature vector $\mathcal{F}$ is a 62-dimensional concatenation comprising:

- **Spectral Envelope:** 40 MFCCs (baseline representation).
- **Spectral Texture:** Spectral Centroid, Spectral Rolloff, and Spectral Contrast.
- **Temporal Dynamics:** Zero Crossing Rate (ZCR) to capture changes in waveform complexity associated with repeated chirp/buzz-like elements.
- **Tonal Content:** Chroma STFT to summarise relative energy distribution across frequency bands.

3.4. Classification framework

The dataset was partitioned using an 80:20 split, stratified by individual to prevent identity leakage. Features were standardised using z-score normalisation ($z = (x - \mu)/\sigma$).

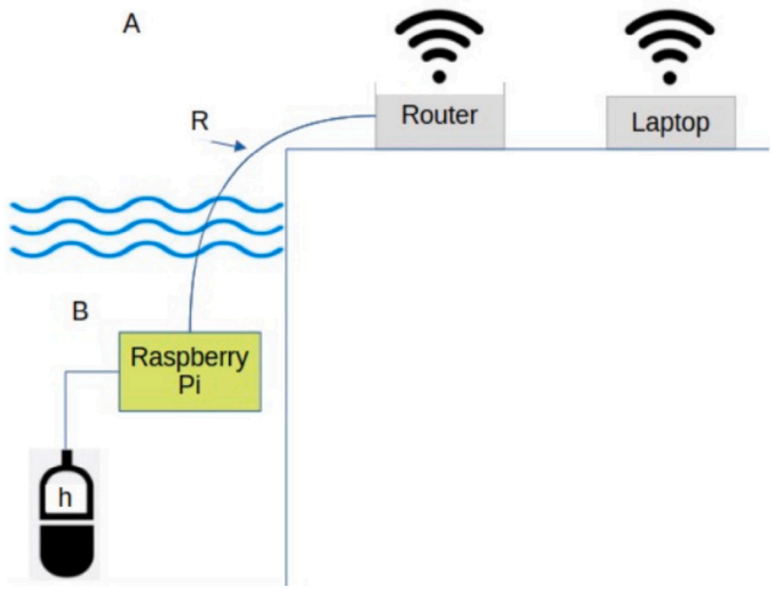


Fig. 3. Schematic of the acoustic recording system. Area A: airborne environment; Area B: underwater environment; arrow R points to the RJ45 cable for SSH connection.

Table 3
Architecture specifications of the 1D-CNN utilised in the classification pipeline.

Component	Description
Model type	1D-CNN
Convolution layer	Conv1D (64 filters, kernel size 3, ReLU activation)
Pooling layer	MaxPooling1D (pool size 2)
Flatten layer	Transition from 3D feature maps to One-Dimensional (1D) vector
Hidden layer	Fully connected dense layer (128 units, ReLU activation)
Output layer	Dense layer (1 unit, Sigmoid activation)

Table 4
Mean acoustic features across lobster groups.

Group	Peak frequency (Hz)	RMS energy	Spectral centroid (Hz)
Juvenile	97.64	0.397	1110.35
Adult	151.79	0.206	1141.13
Female	107.85	0.329	1034.72
Male	151.67	0.229	1194.44

Table 5
Analysis of variance (ANOVA) results for acoustic feature comparison across lobster groups.

Feature	F-value	p-value	Significance
Peak frequency (Hz)	74.52	< 0.001	Significant
RMS energy	752.88	< 0.001	Significant
Spectral centroid (Hz)	329.92	< 0.001	Significant

3.5. Conventional machine learning deep learning architectures

Six supervised classifiers were evaluated: RF (300 decision trees), XGBOOST, Multi Layer Perceptron (MLP) (128 hidden neurons), NB, SVM, and KNN ($k = 5$). We compared standard and dilated 1D-CNN architectures (1–4 layers). The generalised CNN architecture utilised for these experiments is summarised in Table 3.

3.6. Performance evaluation and uncertainty quantification

Model performance was quantified using accuracy and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) while computing confidence intervals (CI). To ensure statistical rigour, 95% confidence intervals (CI) were estimated using non-parametric bootstrap resampling ($n = 1000$). The 2.5th and 97.5th percentiles of the empirical distribution were used as bounds ($CI_{95\%} = [P_{2.5}, P_{97.5}]$). Besides these performance during general classification, the following additional metrics were used: precision, F1-score, and inference time, as shown in Tables 9 and 11.

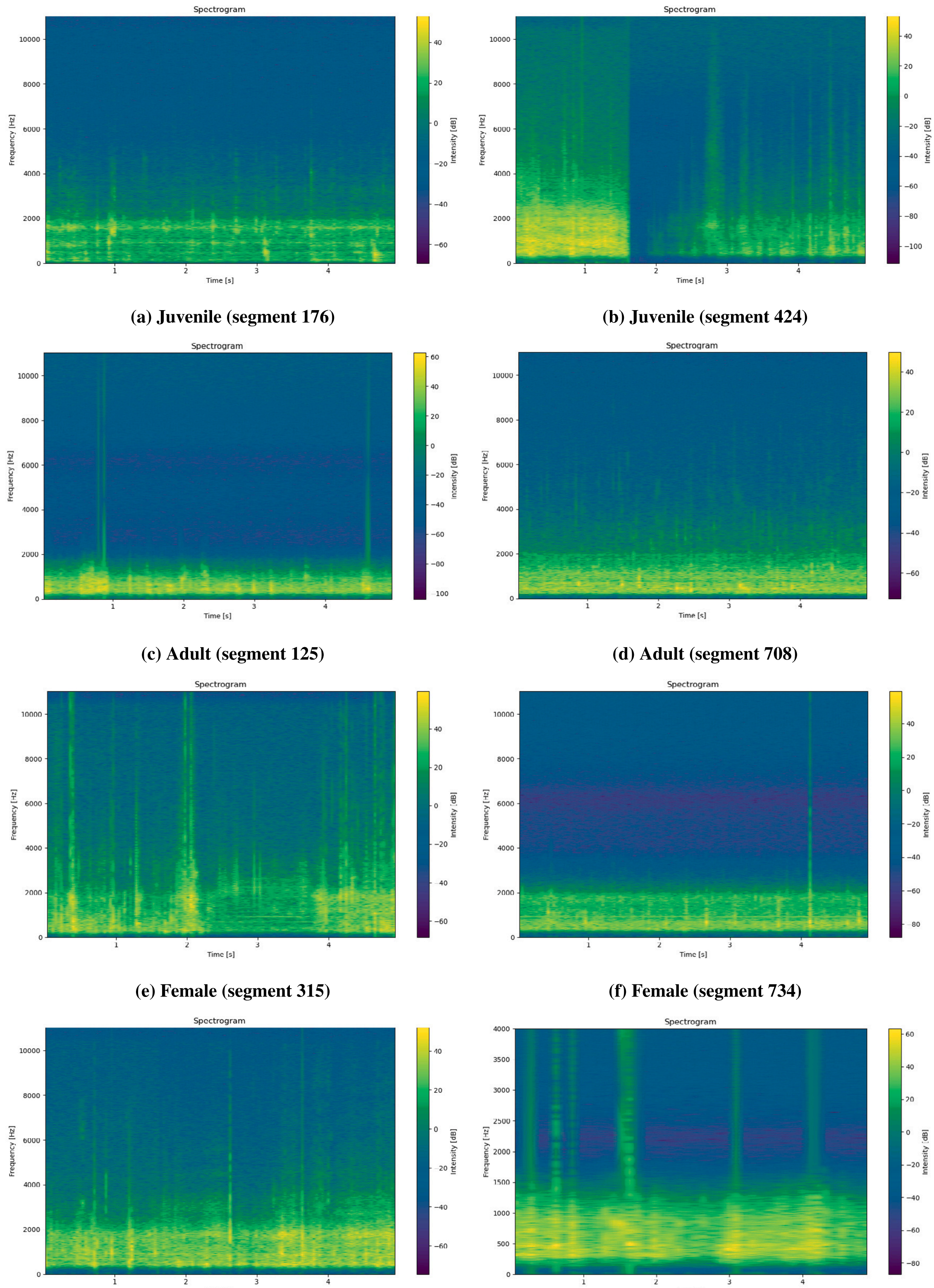
4. Experiments and results

The statistical analysis using one-way ANOVA reveals highly significant differences across all examined acoustic features. Peak frequency ($F = 74.52$, $p < 0.001$), RMS energy ($F = 752.88$, $p < 0.001$), and spectral centroid ($F = 329.92$, $p < 0.001$) demonstrate strong separation between lobster groups. The extremely low p-values indicate that the probability of these differences arising from random variation is negligible. These findings confirm that the observed acoustic differences are consistent and biologically meaningful, reflecting structural variation across age and sex categories (see Table 4).

Before evaluating model performance, representative acoustic signatures were examined to visualise the time–frequency structure of the analysed signals. Fig. 4 presents two representative spectrogram examples per biological category (juvenile, adult, female, and male), selected from distinct recordings after quality screening to ensure the characteristic low-frequency "buzzing/chirp" pattern reported for *H. gammarus*.

While all groups share a broadly similar low-frequency-dominant signature, group-related differences are quantified in the frequency-domain profiles and statistical comparisons (Fig. 5(a) and Table 5), rather than relying on visual inspection of single spectrogram examples alone. Although some spectrogram panels show additional broadband components and/or harmonics extending above 250 Hz (due to tank acoustics and non-stationary events), the dominant energy and the fundamental buzzing activity remain concentrated within the biologically relevant 80–250 Hz band (Fig. 5).

The study evaluates whether key biological attributes (age and sex) can be inferred from European lobster bioacoustic recordings using a range of ML and DL model types. Two classification experiments were conducted: (i) age determination (Adult vs. Juvenile) and (ii)



(a) Juvenile (segment 176) **(b) Juvenile (segment 424)**

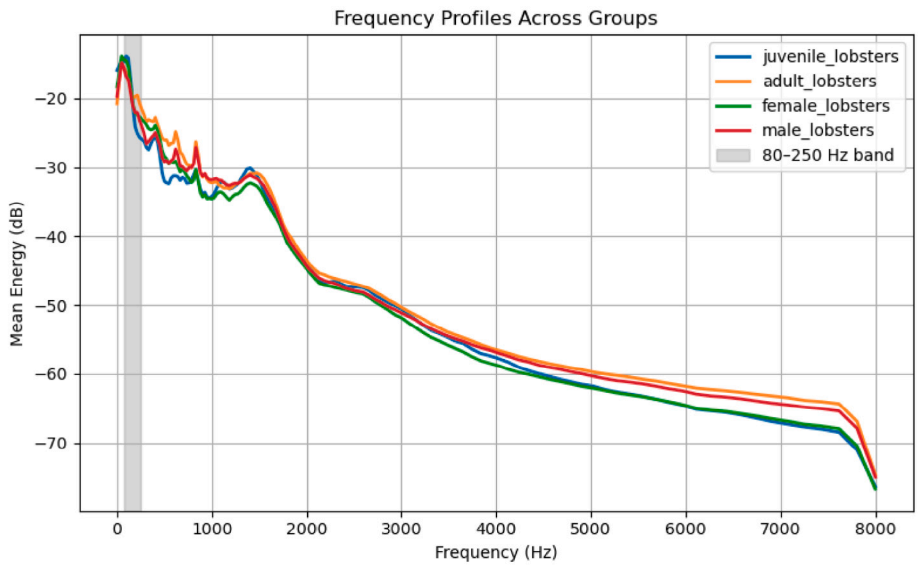
(c) Adult (segment 125) **(d) Adult (segment 708)**

(e) Female (segment 315) **(f) Female (segment 734)**

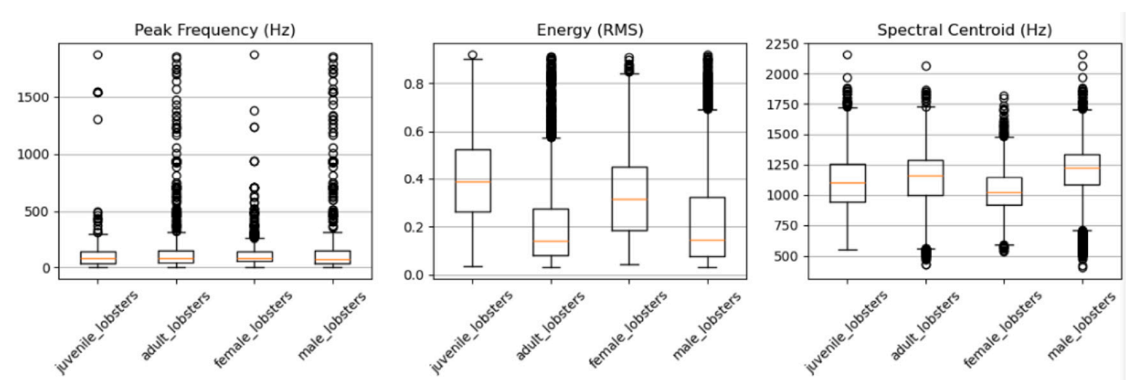
(g) Male (segment 286) **(h) Male (segment 621)**

Fig. 4. Representative spectrograms of the analysed lobster acoustic signals (two examples per biological group). (a–b) juvenile, (c–d) adult, (e–f) female, and (g–h) male. All panels are drawn from distinct analysed segments (segment IDs indicated) and are not duplicated. The examples show energy concentrated at low frequencies with repeated chirp/buzz-like elements, consistent with published descriptions of *H. gammarus* buzzing sounds.

sex determination (Male vs. Female). A summary of the experimental design and evaluation protocol is provided in Table 6. Tables 7 and 8 report the classification performance of all evaluated machine-learning and deep-learning models for the age determination task. In addition to point estimates, 95% bootstrap confidence intervals are reported for classical machine-learning models, providing an estimate of uncertainty



(a) Frequency Profiles



(b) Statistical Distributions

Fig. 5. (a) Frequency-domain profiles computed from the full set of analysed segments (mean spectral energy distribution per group), showing energy distribution across frequencies for each lobster group. The shaded 80–250 Hz band corresponds to the biologically relevant range of lobster carapace vibrations, confirming the presence of expected low-frequency acoustic activity. (b) Boxplots of key acoustic features (peak frequency, RMS energy, and spectral centroid) across groups. Clear separation between distributions is observed, with statistically significant differences confirmed by ANOVA ($p < 0.001$ for all features).

around accuracy and Area Under the Curve (AUC). Confidence intervals were not computed for deep-learning models due to the substantially higher computational cost associated with repeated network training during bootstrapping.

4.1. Experiment 1: Age-based bioacoustics classification

We framed age classification as a binary problem (Adult vs. Juvenile) and evaluated six classical ML models (RF, XGBOOST, MLP, NB, SVM, KNN) and a family of DL models (1D-CNN and 1D-DCNN with 1–4 hidden layers). Tables 9 and 11 summarise results across MFCC dimensionalities (40, 50, 60). Figs. 6 and 7 illustrate metric trends for KNN and SVM as MFCC number varies; confusion matrices for leading models are presented in Figs. 8–13. Key observations include high discrimination overall. With the exception of NB, nearly all models achieve accuracy ≳97% on the Adult vs. Juvenile task. ML such as SVM and MLP are among the top performers in several metrics; the highest single accuracy per the provided point estimates is **98.50%** (SVM, MFCC=50). MLP achieves the best AUC-ROC (**99.86%**, MFCC=60). In terms of inference time trade-offs, classical ML methods such as KNN and NB yield very low IT, 0.87 ms and 2.10 ms, respectively; whereas DL models require seconds per sample. The IT difference between model types (ML and DL) demonstrates an important operational consideration in computing requirements. Effect of MFCC dimensionality is noted; there is no universal, monotonic improvement with larger MFCC

Table 6
Summary of experimental design and evaluation protocol.

Aspect	Description
Biological attributes	Age (Adult vs. Juvenile) and Sex (Male vs. Female)
Experiments	Experiment 1: Age-based classification Experiment 2: Sex-based classification
Models evaluated	Six classical machine-learning algorithms and eight deep-learning variants (1D-CNN and 1D-DCNN)
Input features	MFCCs computed using 40, 50, and 60 coefficients
Feature processing	Optional PCA for feature decorrelation
Evaluation metrics	Accuracy, Precision, Recall, F1-score, and AUC-ROC
Efficiency metric	Per-sample inference time (IT)
Evaluation protocol	Performance estimated on held-out test partitions using point estimates

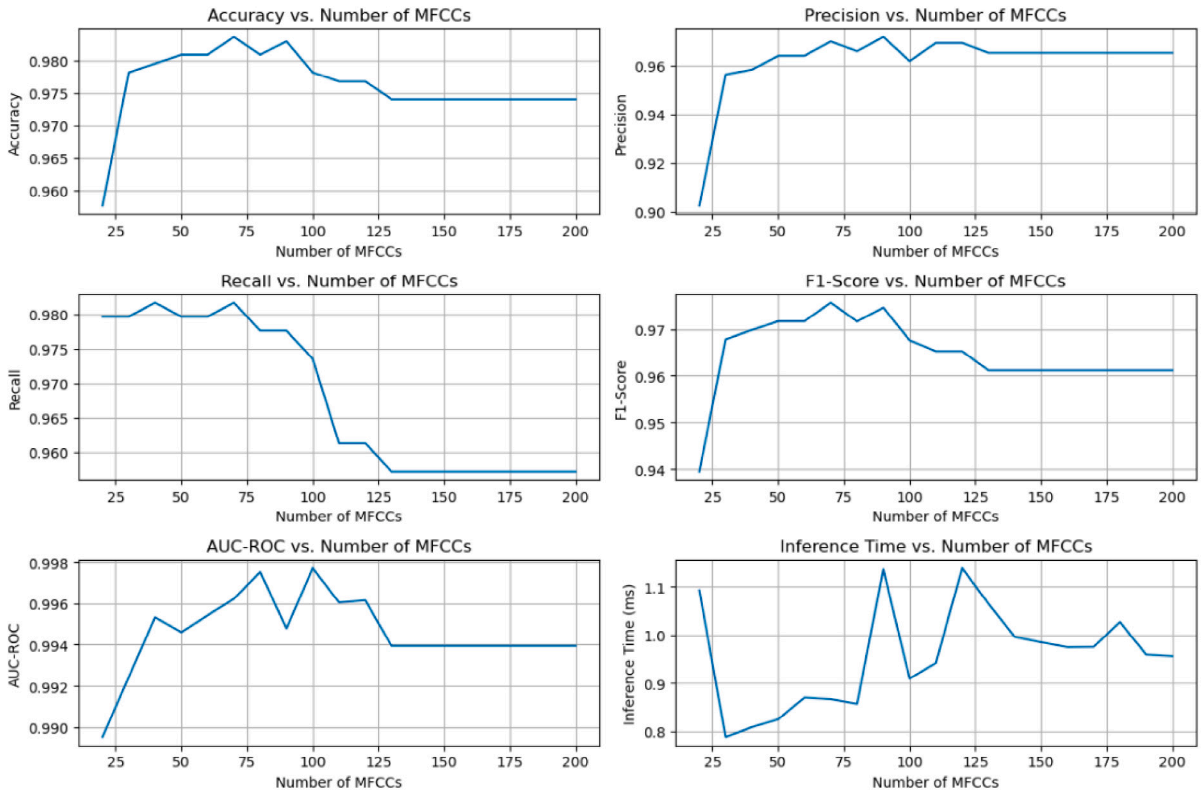


Fig. 6. Performance metrics for KNN as a function of MFCC dimensionality (Adult vs. Juvenile).

Table 7
Performance comparison of machine-learning models with reported 95% confidence intervals for age determination using MFCC features.

Model	Accuracy (%)	95% CI	AUC (%)	95% CI
Random forest	97.13	96.17–97.95	99.43	99.07–99.73
XGBoost	98.08	97.40–98.77	99.63	99.39–99.84
MLP	98.02	97.26–98.70	99.77	99.61–99.89
Naive Bayes	90.90	89.46–92.41	97.02	96.17–97.78
SVM	98.29	97.61–98.91	99.61	99.24–99.88
KNN	97.74	96.99–98.43	99.27	98.75–99.71

Table 8
Performance comparison of machine-learning models with reported 95% confidence intervals for sex determination using MFCC features.

Model	Accuracy (%)	95% CI	AUC (%)	95% CI
Random Forest	93.02	91.72–94.32	97.83	97.22–98.40
XGBoost	93.50	92.27–94.73	98.61	98.16–99.01
MLP	93.02	91.66–94.25	98.56	98.15–98.96
Naive Bayes	80.23	78.18–82.22	88.89	87.19–90.54
SVM	95.08	93.98–96.17	98.79	98.37–99.16
KNN	95.14	94.05–96.24	98.87	98.42–99.26

counts: model-dependent behaviour is observed; some models improve slightly with increasing MFCC, where as others remain unchanged or vary non-monotonically.

From Table 10 it is conclusive that MLP and SVM tie as best by average rank (MLP best by small margin), KNN and XGBOOST are competitive mid-pack, RF lags mainly due to inference time, and NB ranks last due to poor discrimination despite fast inference.

Based on the results in Table 12, 1D-CNN (1 layer) and 1D-DCNN (1–2 layers) yield the best average rank: good discrimination with comparatively lower inference costs among DLs. Very deep variants (3–4 layers) rarely yield substantial gains but increase inference latency. The results for age group determination indicate that the classification of age groups (Adult vs. Juvenile) yielded even higher performance metrics, as shown in Table 13. The Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) shared high accuracies of 98.43%, while KNN reached a peak accuracy of 98.56%.

4.2. Experiment 1: Comparative analysis

Overall, classical ML and shallow DL models both achieved strong performance on the adult vs. juvenile task. SVM and MLP are among the most consistent classifiers with top-tier accuracy and F1; SVM reaches the top accuracy point while MLP yields the top AUC. KNN has excellent recall and the fastest inference time, making it suitable for low-latency systems, while NB computationally cheap though substantially weaker in accuracy. DL models can match or slightly exceed ML performance, notably some 1D-DCNN variants; however, generally at a significantly higher inference cost. As already pointed out, MFCC dimensionality matters but is model dependent. We cannot recommend a single MFCC count for all models; nested CV and broader sweeps (e.g., 20–80) are advised (see Fig. 14).

4.3. Experiment 2: Sex-based bioacoustics classification

We evaluated the capacity of ML and DL models to discriminate sex (male vs. female) from lobster bioacoustic recordings. Models tested include RF, XGBOOST, MLP, NB, SVM, KNN and a family of 1D-CNN (1D-CNN and 1D-DCNN) with varying depths. All inputs were derived from MFCC representations (40, 50 and 60 coefficients) and preprocessed with PCA where indicated. Performance is summarised in Table 14 (ML classifiers) and Table 16 (DL classifiers). Fig. 15 illustrates metric trends for the 1D-CNN (1 hidden layer) across MFCC dimensions.

The ranking summary in Table 15 indicates that for the classification of Male vs Female SVM and MLP are the strongest performers by average rank (SVM marginally best on accuracy and multiple metrics; MLP achieves top AUC), while KNN is notable for best precision and

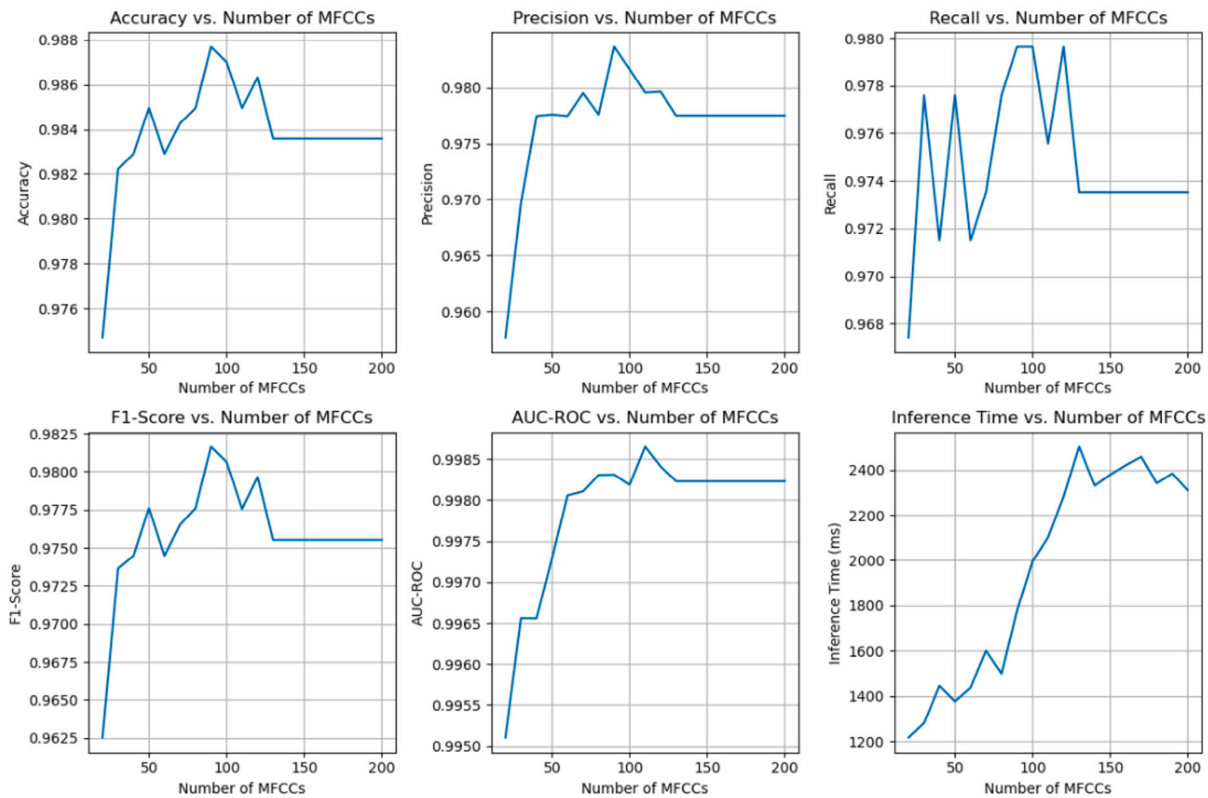


Fig. 7. Performance metrics for SVM as a function of MFCC dimensionality (Adult vs. Juvenile).

Table 9
Performance metrics of ML models for Adult vs. Juvenile classification (MFCC = 40, 50, 60). Best value per column is **bolded** (for IT lower is better).

Model	MFCC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC-ROC (%)	Inference time (ms)
Random forest	40	97.20	97.27	94.30	95.76	99.50	2300.04
	50	97.26	97.47	94.30	95.86	99.60	2511.00
	60	97.20	97.67	93.89	95.74	99.62	2535.04
XGBoost	40	97.47	96.71	95.72	96.21	99.65	120.47
	50	97.61	96.53	96.33	96.43	99.74	154.90
	60	97.67	96.54	96.54	96.54	99.75	148.32
MLP	40	98.43	**98.35**	96.95	97.64	99.76	1271.72
	50	97.88	96.75	96.95	96.85	99.82	1452.11
	60	98.43	97.37	97.96	97.66	**99.86**	2015.05
Naive Bayes	40	89.12	80.29	89.61	84.70	97.09	2.10
	50	91.31	82.04	94.91	88.01	97.73	2.31
	60	90.83	80.94	95.11	87.45	97.80	2.38
SVM	40	98.29	97.75	97.15	97.45	99.66	1381.98
	50	**98.50**	97.76	**97.76**	**97.76**	99.73	1437.78
	60	98.29	97.75	97.15	97.45	99.81	1507.90
KNN	40	97.95	95.83	**98.17**	96.98	99.53	**0.87**
	50	98.08	96.39	97.96	97.17	99.46	0.91
	60	98.08	96.39	97.96	97.17	99.54	1.81

Notes: Best values per column are bolded.

Table 10
Compact ranking summary (Adult vs Juvenile) — ML models. Lower average rank = better. Ranks computed from point estimates (Accuracy, Precision, Recall, F1, AUC) and Inference Time (IT, lower is better).

Model	MFCC	Acc_rank	Prec_rank	Rec_rank	F1_rank	AUC_rank	IT_rank	AvgRank
MLP	40	2	1	3	2	1	4	**2.17**
SVM	50	1	2	2	1	3	5	**2.33**
KNN	60	3	5	1	3	5	1	3.00
XGBoost	60	4	4	4	4	2	3	3.50
RF	50	5	3	6	5	4	6	4.83
Naive Bayes	50	6	6	5	6	6	2	5.17

Table 11
Performance metrics of DL models (1D-CNN and 1D-DCNN families) for Adult vs. Juvenile classification. Best value per column is **bolded** (for IT lower is better).

Model (layers)	MFCC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC-ROC (%)	Inference time (ms)
1D-CNN (1 L)	40	97.88	96.00	97.76	96.87	99.80	7537.79
	50	**98.15**	97.15	97.35	**97.25**	**99.83**	16 249.91
	60	97.67	98.11	94.91	96.48	99.79	6463.39
1D-CNN (2 L)	40	97.26	94.65	97.35	95.98	99.72	10 580.45
	50	97.67	96.35	96.74	96.54	99.72	27 973.32
	60	97.40	95.03	97.35	96.18	99.72	9178.96
1D-CNN (3 L)	40	97.47	95.95	96.54	96.24	99.66	13 254.90
	50	97.40	98.09	94.09	96.05	99.59	29 017.33
	60	97.95	**98.53**	95.32	96.89	99.73	12 854.35
1D-CNN (4 L)	50	95.90	98.00	89.61	93.62	99.58	**3318.51**
	60	97.67	95.07	**98.17**	96.59	99.56	13 406.18
1D-DCNN (1 L)	40	98.02	96.39	97.79	97.07	99.81	7478.64
	50	98.02	96.57	97.56	97.06	99.82	8040.94
	60	97.74	96.54	96.74	96.64	**99.83**	6583.42
1D-DCNN (2 L)	40	98.08	96.77	97.56	97.16	99.70	10 615.49
	50	97.74	95.44	97.96	96.68	99.68	12 299.25
	60	97.47	98.30	94.09	96.15	99.58	9735.64
1D-DCNN (3 L)	40	96.24	98.23	90.43	94.17	99.54	12 310.96
	50	97.67	96.73	96.33	96.53	99.64	14 671.59
	60	96.99	93.40	97.96	95.63	99.63	12 111.80
1D-DCNN (4 L)	50	95.42	91.73	94.91	93.29	99.14	15 311.84
	60	97.13	97.06	94.30	95.66	99.69	12 630.36

Notes: Best values per column are bolded; ties are bolded for all tied entries. Inference time is per-sample wall-clock time measured on the tuning hardware.

Table 12
Compact ranking summary (Adult vs Juvenile), DL models (1D-CNN/1D-DCNN families). Lower average rank = better.

Model (layers)	MFCC	Acc_rank	Prec_rank	Rec_rank	F1_rank	AUC_rank	IT_rank	AvgRank
1D-CNN (1 L)	50	1	2	4	1	1	7	**2.67**
1D-DCNN (1 L)	40	3	6	2	3	2	1	**2.83**
1D-DCNN (2 L)	40	2	4	3	2	5	2	3.00
1D-CNN (3 L)	60	4	1	7	4	3	4	3.83
1D-CNN (4 L)	60	6	8	1	5	8	5	5.50
1D-CNN (2 L)	50	6	7	5	6	4	8	6.00
1D-DCNN (4 L)	60	8	3	8	8	6	3	6.00
1D-DCNN (3 L)	50	6	5	6	7	7	6	6.17

Table 13
Comprehensive classification performance for **age group determination** using multi-feature fusion.

Model	Acc (%)	95% CI	AUC (%)	95% CI
Random Forest	97.40	96.65–98.15	99.68	99.50–99.83
XGBoost	97.81	97.06–98.50	99.78	99.66–99.88
MLP	98.43	97.74–99.04	99.89	99.81–99.95
Naive Bayes	91.04	89.60–92.48	97.66	96.98–98.28
SVM	98.43	97.81–98.97	99.88	99.79–99.95
KNN	**98.56**	**97.95–99.11**	**99.43**	**98.98–99.81**
1D-CNN (1L)	98.08	97.40–98.77	99.89	99.82–99.95
1D-DCNN (1L)	98.29	97.61–98.97	99.89	99.81–99.95
1D-CNN (2L)	97.67	96.85–98.43	99.81	99.69–99.91
1D-DCNN (2L)	97.81	97.06–98.50	99.78	99.63–99.91
1D-CNN (3L)	97.81	97.06–98.50	99.79	99.66–99.89
1D-DCNN (3L)	97.20	96.30–98.02	99.62	99.40–99.80
1D-CNN (4L)	98.15	97.54–98.77	99.81	99.70–99.91
1D-DCNN (4L)	97.40	96.51–98.22	99.56	99.24–99.80

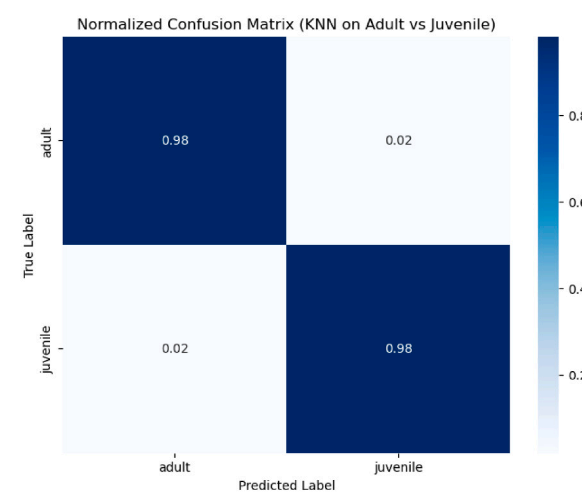


Fig. 8. Normalised confusion matrix for KNN (Adult vs. Juvenile).

best (lowest) inference time. XGBOOST and RF occupy the mid-range; NB ranks lowest due to substantially weaker discrimination despite very low IT. The results for sex determination show that the performance for sex classification using the fused feature set is consistent, as summarised in Table 18. Most models achieved AUC scores above 97%, with the KNN algorithm providing the highest accuracy at 95.42% (AUC: 99.06%).

4.4. Experiment 2: Comparative analysis

The results in Table 17 summarises that among DLs, shallow architectures (1D-CNN 1-layer and 1D-DCNN 1-layer) obtain the best average ranks, balancing strong discrimination and comparatively lower inference costs. Some deeper architectures have high precision but suffer from larger inference time; deeper does not uniformly improve

Table 14
Performance metrics of ML models for Male vs. Female classification (MFCC = 40, 50, 60). Best value per column is **bolded**.

Model	MFCC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC-ROC (%)	Inference time (ms)
Random forest	40	93.23	94.58	94.26	94.42	98.08	2559.0391
	50	92.54	93.23	94.59	93.91	98.16	2986.6513
	60	92.54	93.33	94.48	93.90	98.18	2968.7502
XGBoost	40	94.12	94.75	95.61	95.18	98.73	334.6043
	50	95.21	96.16	95.95	96.05	98.81	231.1075
	60	95.28	96.17	96.06	96.11	98.97	221.2673
MLP	40	95.35	96.80	95.50	96.15	99.03	3324.9233
	50	95.14	96.58	95.38	95.98	**99.19**	2794.5994
	60	95.90	96.83	96.40	96.61	99.16	2260.6717
Naive Bayes	40	81.60	87.52	81.31	84.30	91.04	2.3317
	50	82.56	88.36	82.09	85.11	90.85	2.0922
	60	82.01	88.82	80.52	84.47	89.83	2.3787
SVM	40	95.42	96.07	96.40	96.23	98.86	2586.4148
	50	**96.17**	96.95	**96.73**	**96.84**	99.03	2609.8358
	60	95.96	96.94	96.40	96.67	99.11	2811.5256
KNN	40	94.87	95.93	95.61	95.77	98.85	0.8528
	50	95.49	97.03	95.50	96.25	98.97	**0.6002**
	60	95.62	**97.14**	95.61	96.37	99.00	0.6658

Notes: Best per-column values are bolded (for inference time lower is better).

Table 15
Compact ranking summary (Male vs Female) — ML models. Lower average rank = better. Cells shaded indicate rank ≤ 2 (top-2).

Model	MFCC	Acc_rank	Prec_rank	Rec_rank	F1_rank	AUC_rank	IT_rank	AvgRank
KNN	60	3	1	4	3	3	1	2.50
MLP	60	2	3	2	2	1	4	**2.33**
Naive Bayes	50	6	6	6	6	6	2	5.33
Random forest	40	5	5	5	5	5	5	5.00
SVM	50	1	2	1	1	2	6	**2.17**
XGBoost	60	4	4	3	4	4	3	3.67

Table 16
Performance metrics of DL models (1D-CNN and 1D-DCNN families) for Male vs. Female classification. Best value per column is **bolded**.

Model (layers)	MFCC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC-ROC (%)	Inference time (ms)
1D-CNN (1 L)	40	94.60	94.50	96.73	95.60	98.75	8253.8011
	50	94.75	95.40	96.03	95.69	99.04	8069.7830
	60	**95.35**	95.61	96.82	**96.20**	99.13	**6433.0564**
1D-CNN (2 L)	40	94.25	95.17	95.38	95.28	98.67	10 445.2045
	50	95.04	96.09	**97.78**	95.92	98.95	12 180.0492
	60	94.94	95.83	95.86	95.83	99.00	8758.4734
1D-CNN (3 L)	40	92.13	90.64	97.07	93.75	98.10	13 290.2761
	50	93.61	94.98	94.51	94.72	98.64	15 455.8352
	60	94.63	96.05	95.08	95.56	98.78	12 078.9611
1D-CNN (4 L)	50	93.66	93.45	96.31	94.86	98.49	15 676.7719
	60	93.93	94.69	95.39	95.02	98.62	13 249.9224
1D-DCNN (1 L)	40	94.19	94.66	95.83	95.24	98.79	7332.1474
	50	**95.35**	95.44	96.99	**96.20**	99.19	7940.4736
	60	95.11	95.70	96.37	96.00	**99.21**	6515.3026
1D-DCNN (2 L)	40	93.91	96.26	93.58	94.92	98.47	10 611.2289
	50	94.87	95.75	95.83	95.76	98.97	12 129.8255
	60	95.03	96.01	95.83	95.85	99.05	8682.1848
1D-DCNN (3 L)	40	92.95	93.47	95.05	94.25	98.27	12 144.5677
	50	93.98	94.45	95.72	95.08	98.58	14 186.0330
	60	94.27	95.35	95.22	95.28	98.63	11 781.5477
1D-DCNN (4 L)	50	93.23	**98.80**	93.92	94.40	98.46	15 143.1859
	60	92.89	96.34	91.78	94.00	98.37	12 650.9302

Notes: Best per-column values are bolded (for inference time lower is better). Ties are bolded for all tied entries.

the overall rank. To summarise the main findings, across the classical ML classifiers, SVM achieved the highest point accuracy (96.17%, MFCC=50) and high values on other discrimination metrics (precision, recall, F1) (see Table 14). XGBOOST and MLP are close competitors;XG-BOOST attains accuracies above 95% for MFCC=50 and MFCC=60, while the MLP likewise attains its best accuracy at MFCC=60 (95.90%). NB performs substantially worse than the other ML methods (≈82%–83% accuracy), illustrating that strong parametric assumptions are less appropriate for this task. For the DL family, shallow architectures (1–2 hidden layers) generally match or slightly outperform the best ML models on overall accuracy and AUC-ROC, with the 1D-CNN and 1D-DCNN achieving peak accuracies of 95.35% under different MFCC settings (Table 16). AUC-ROC values are uniformly high across DL models (> 98%), indicating strong separability between male and female classes. IT differs substantially between paradigms: ML classifiers (except RF which uses many trees) generally yield low inference latency relative to DL models; DL ITs are several seconds per sample (Tables 14, 16), which may constrain real-time deployments.

Table 17
Compact ranking summary (Male vs Female) — DL models (1D-CNN/1D-DCNN). Lower average rank = better. Cells shaded indicate rank ≤ 2 (top-2).

Model (layers)	MFCC	Acc_rank	Prec_rank	Rec_rank	F1_rank	AUC_rank	IT_rank	AvgRank
1D-CNN (1 L)	60	1	5	3	1	2	1	**2.17**
1D-CNN (2 L)	50	3	2	1	3	4	6	3.17
1D-CNN (3 L)	60	5	3	7	5	5	5	5.00
1D-CNN (4 L)	60	7	8	5	7	7	7	6.83
1D-DCNN (1 L)	50	1	6	2	1	1	2	**2.17**
1D-DCNN (2 L)	60	4	4	4	4	3	3	3.67
1D-DCNN (3 L)	60	6	7	6	6	6	4	5.83
1D-DCNN (4 L)	50	8	1	8	8	8	8	6.83

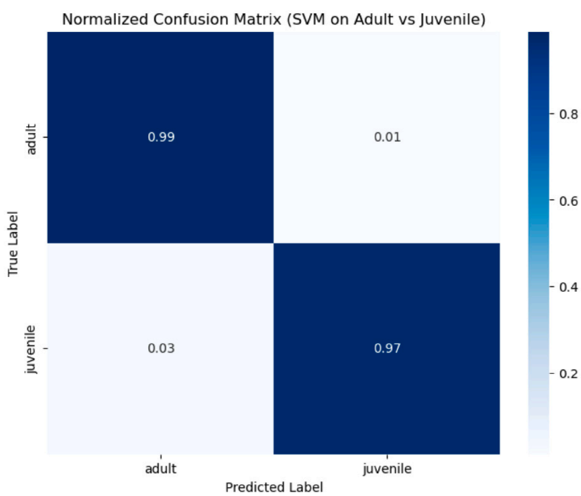


Fig. 9. Normalised confusion matrix for SVM (Adult vs. Juvenile).

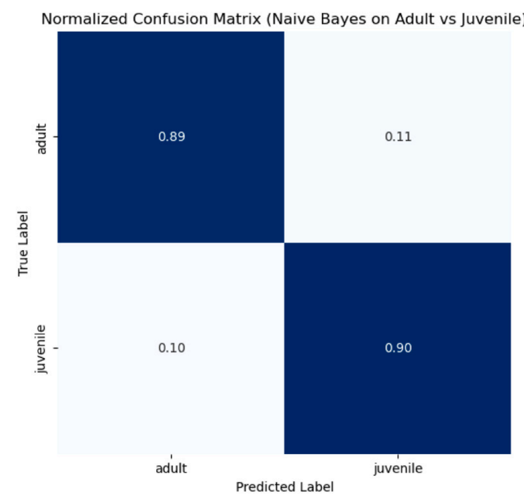


Fig. 11. Normalised confusion matrix for NB (Adult vs. Juvenile).

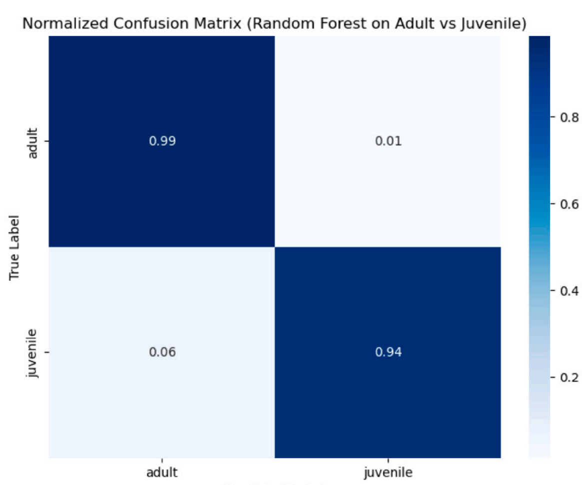


Fig. 10. Normalised confusion matrix for RF (Adult vs. Juvenile).

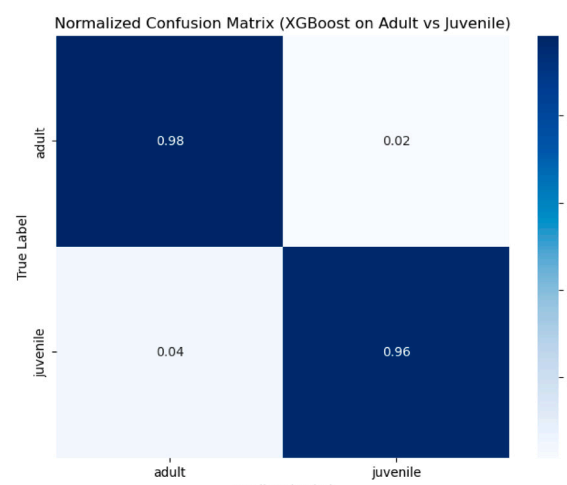


Fig. 12. Normalised confusion matrix for XGBoost (Adult vs. Juvenile).

The results lead to precise interpretation and practical implications as they indicate three practical points: (i) Top accuracy vs. operational cost, (ii) Model-dependent sensitivity to input dimensionality, and (iii) Shallow architectures can suffice. SVM and XGBOOST produce the best discrimination with modest inference cost; SVM is competitive albeit with higher IT than lightweight ML methods. Deep models yield comparable discrimination but incur substantially higher ITs; this trade-off should weigh in deployment decisions where latency matters. Pertaining to input dimensionality, increasing MFCC count does not universally improve performance. For some models, such as (XGBOOST and MLP, modest gains occur when moving from 40 to 50/60 MFCCs; for RF and some CNN variants the effect is neutral or inconsistent. Thus, optimal MFCC dimensionality appears model-dependent and should be identified per algorithm (preferably by nested CV). With regards to architectural depth, for both 1D-CNN and 1D-DCNN families the best performance is often achieved with one or two hidden layers; deeper architectures yield diminishing returns and substantially larger inference times, suggesting over-parametrisation for the implemented dataset.

On inference time and deployment, DL models exhibit inference times in the multi-second range per sample, as shown in Tables 16; whereas some ML models, such as KNN and NB, have very low latency. When deployment latency is critical, like in-situ monitoring, edge devices, we recommend profiling models under representative hardware

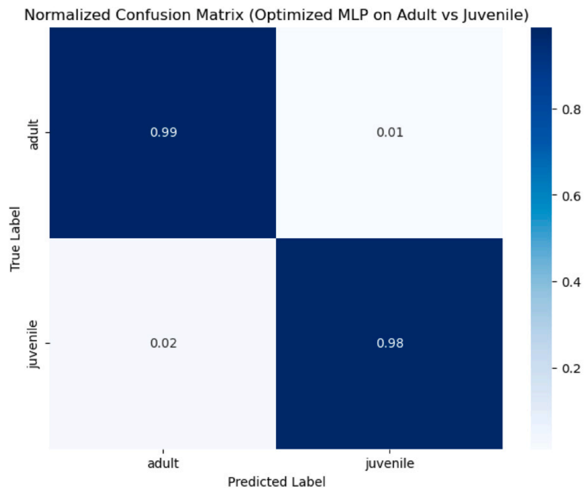


Fig. 13. Normalised confusion matrix for MLP (Adult vs. Juvenile).

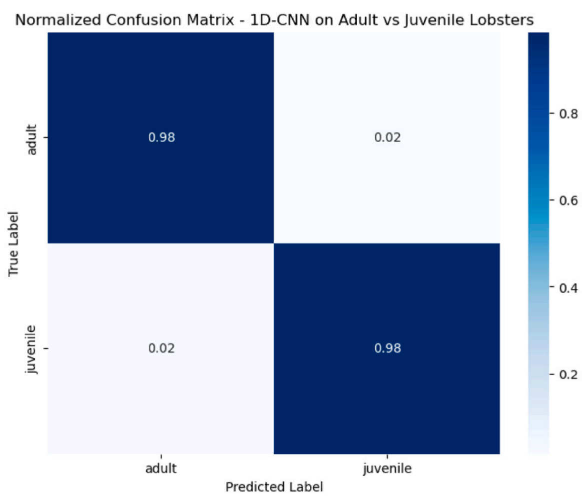


Fig. 14. Normalised confusion matrix for 1D-CNN (1 layer) (Adult vs. Juvenile).

Table 18
Comprehensive classification performance for **sex determination** using multi-feature fusion.

Model	Acc (%)	95% CI	AUC (%)	95% CI
Random forest	92.07	90.63–93.43	97.50	96.78–98.13
XGBoost	94.05	92.82–95.28	98.59	98.12–99.00
MLP	94.80	93.71–95.90	98.87	98.48–99.21
Naive Bayes	77.50	75.44–79.62	87.52	85.76–89.23
SVM	94.87	93.71–95.96	98.82	98.39–99.19
KNN	**95.42**	**94.32–96.58**	**99.06**	**98.66–99.41**
1D-CNN (1L)	94.60	93.50–95.83	98.71	98.23–99.12
1D-DCNN (1L)	94.46	93.30–95.62	98.88	98.46–99.26
1D-CNN (2L)	95.35	94.25–96.44	98.75	98.31–99.17
1D-DCNN (2L)	94.46	93.23–95.69	98.79	98.36–99.16
1D-CNN (3L)	94.66	93.43–95.76	98.71	98.25–99.15
1D-DCNN (3L)	93.91	92.68–95.08	98.88	98.50–99.23
1D-CNN (4L)	93.50	92.27–94.80	98.25	97.67–98.75
1D-DCNN (4L)	92.68	91.31–94.05	98.23	97.71–98.70

and considering model compression (quantisation) for CNNs models, pruning or smaller ensembles for tree/boosting methods, lightweight feature sets selection via feature-importance (SHAP) or sequential feature selection.

4.5. Acoustic signal representation across lobster groups

Fig. 4 presents representative spectrograms from specific analysed segments (two examples per group) for juvenile, adult, female, and male lobsters. Across categories, the examples show energy concentrated at low frequencies with repeated chirp/buzz-like elements, consistent with published descriptions of *H. gammarus* buzzing.

Although representative spectrograms can look broadly similar across groups, the group-related differences are supported quantitatively by the frequency-domain profiles and statistical comparisons (Fig. 5 and Table 5).

4.6. Frequency-domain analysis

Fig. 5(a) shows the mean frequency profiles for each lobster group. Energy is consistently concentrated within the low-frequency region, with a prominent band between 80–250 Hz corresponding to the biologically relevant vibration range. Systematic shifts in peak frequency are observed, with juvenile lobsters exhibiting lower peak frequencies (approximately 100 Hz), while adult and male lobsters show higher peak frequencies (approximately 150 Hz).

These patterns are consistent with expected size-dependent biomechanical properties and support the biological origin of the recorded signals.

4.7. Statistical feature analysis

Fig. 5(b) presents boxplots of key acoustic features across lobster groups. Clear separation between distributions is observed for peak frequency, RMS energy, and spectral centroid.

Statistical analysis using one-way ANOVA reveals highly significant differences across all examined acoustic features. Peak frequency (F = 74.52, $p < 0.001$), RMS energy ($F = 752.88$, $p < 0.001$), and spectral centroid ($F = 329.92$, $p < 0.001$) demonstrate strong separation between groups. The extremely low p-values indicate that the probability of these differences arising from random variation is negligible.

Overall, the results confirm that acoustic emissions differ systematically across lobster groups and reflect consistent, biologically meaningful patterns associated with age and sex categories.

5. Stacking ensemble: proposed pipeline and rationale

Ensemble learning synthesises multiple base learners to produce a single, often more accurate and robust predictor than any individual model (Sill, 2009; Zhou, 2012; Breiman, 1996b; Wolpert, 1992; Dietterich, 2000; Hastie et al., 2001). Following this principle, we propose a stacking ensemble that explicitly leverages complementary model inductive biases to improve discrimination and calibration on bioacoustic classification tasks.

5.1. Design rationale

We assemble a diverse set of base learners, namely RF, an XGBOOST, SVM and 1D-CNN, because (i) tree ensembles like RF and XGBOOST provide complementary variance and bias reduction mechanisms (Breiman, 2001a; Chen and Guestrin, 2016), (ii) SVMs often perform well on moderate dimensional engineered feature spaces, and (iii) 1D-CNNs can extract local temporal/spectral patterns directly from short waveform or spectrogram segments. Combining these heterogeneous learners increases the likelihood that different aspects of the signal are captured, which stacking can exploit (Wolpert, 1992; Breiman, 1996b).

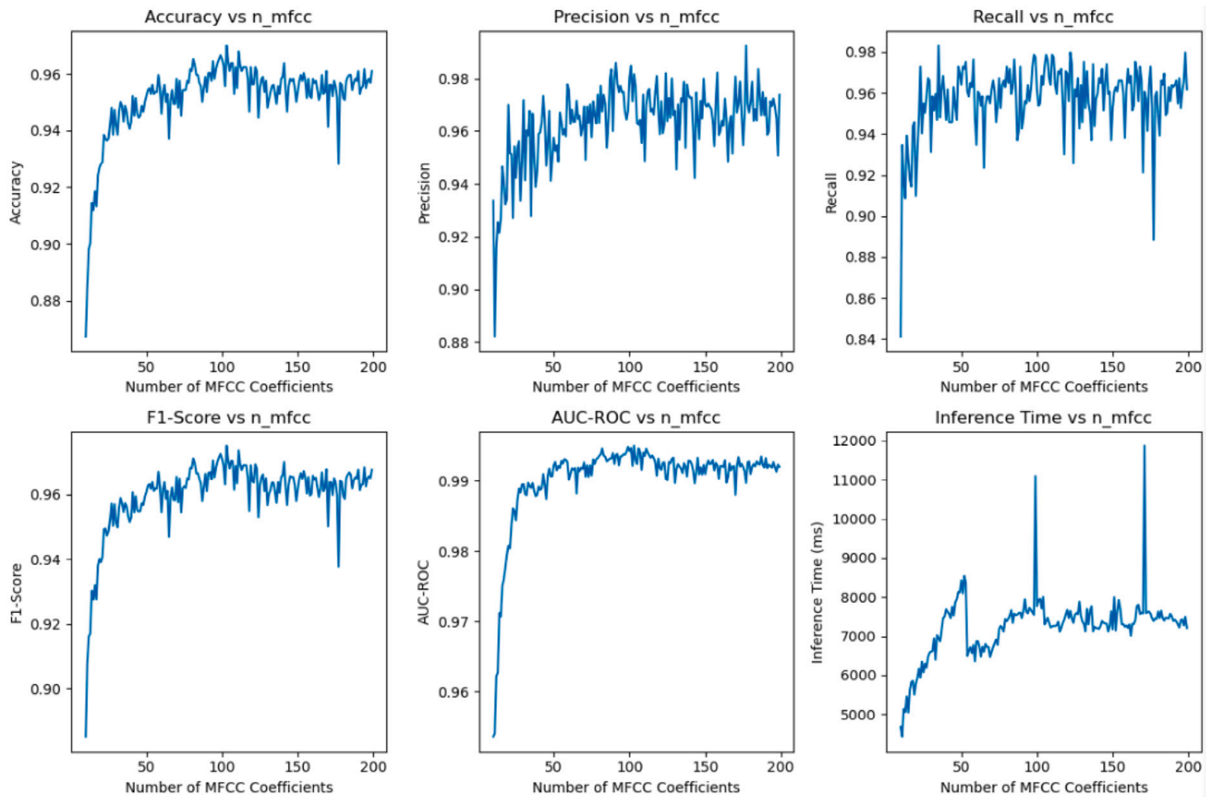


Fig. 15. Performance of 1D-CNN (1 hidden layer) as a function of MFCC dimensionality for male vs. female classification. Points show mean performance per metric (accuracy, precision, recall, F1, AUC-ROC, inference time) computed across cross-validation folds.

5.2. Data streams and feature engineering

The pipeline maintains two parallel feature streams: (1) engineered tabular features (spectral summaries, statistical descriptors and other domain-specific features), used by RF, XGBOOST and SVM; and (2) raw or minimally preprocessed short waveform/spectrogram segments used to train the 1D-CNN. Prior to modelling, tabular features are standardised and, where appropriate, decorrelated with PCA; PCA is only applied within training folds to avoid leakage.

5.3. Training protocol and leakage prevention

To avoid target leakage and to produce out-of-sample meta-training data, we use stratified K-fold cross-validation on the training set (here $K = 5$ unless otherwise stated) to generate OOF predicted probabilities for each base learner. For each fold:

1. Train each base learner on the $K - 1$ folds (using the same preprocessing pipeline fit only on those folds).
2. Predict class probabilities on the held-out fold and store these OOF probabilities.

After completing K folds we obtain an OOF prediction matrix of shape $(N_{\text{train}}, B{\times}C)$ where B is the number of base learners and C the number of classes. These OOF features are then used to train a low-capacity meta-learner (logistic regression with L_2 regularisation by default) that maps the base learners' probability outputs to final predictions. Finally, for deployment each base learner is retrained on the full training set (with preprocessing pipelines refit on full training data) and the meta-learner is used to combine their outputs at test time. This procedure prevents information flow from test folds into the meta-learner during training and is standard in stacking implementations (Wolpert, 1992; Breiman, 1996b).

With regards to meta-learner and calibration, we recommend a low-capacity, regularised meta-learner (logistic regression or a small gradient-boosting machine) to avoid overfitting the OOF matrix. Because stacked outputs can be miscalibrated, we emphasise post-hoc calibration diagnostics and correction (e.g., reliability diagrams, Brier score, Platt scaling or isotonic regression) as part of the validation pipeline (Breiman, 2001a). Where probabilistic interpretation is required, calibration should be reported alongside discrimination metrics. For practical considerations, we recommend hyperparameter optimisation for base learners inside the inner cross-validation loop (or via nested CV) to avoid optimistic bias, parallelised training of base learners and careful resource reporting (Central Processing Unit (CPU)/Graphics Processing Unit (GPU), wall-clock time) to document computational cost, storing OOF predictions, fitted preprocessing objects, random seeds, and software environment (package versions or containers) to ensure reproducibility according to Findable, Accessible, Interoperable, and Reusable (FAIR) principles, interpreting ensemble behaviour using feature-importance measures (e.g. SHAP for tree-based learners), and inspecting meta-learner coefficients to understand how base learners are weighted.

Concerning implementation, a schematic of the proposed stacking pipeline is shown in Fig. 16. The implementation (data preprocessing pipelines, model definitions, and OOF-generation scripts) is provided in the code repository cited in the manuscript (Domingos, 2025). We also provide recommended default meta-learner settings (logistic regression with L_2 penalty, regularisation strength selected by CV) and include a brief ablation table in Appendix A (Table A.28) comparing averaging, majority-vote, and stacked meta-learner performance.

6. Model architectures and parametrisation

A clear, concise description of model architectures and their parametrisation is essential for scientific reproducibility and interpretability. For

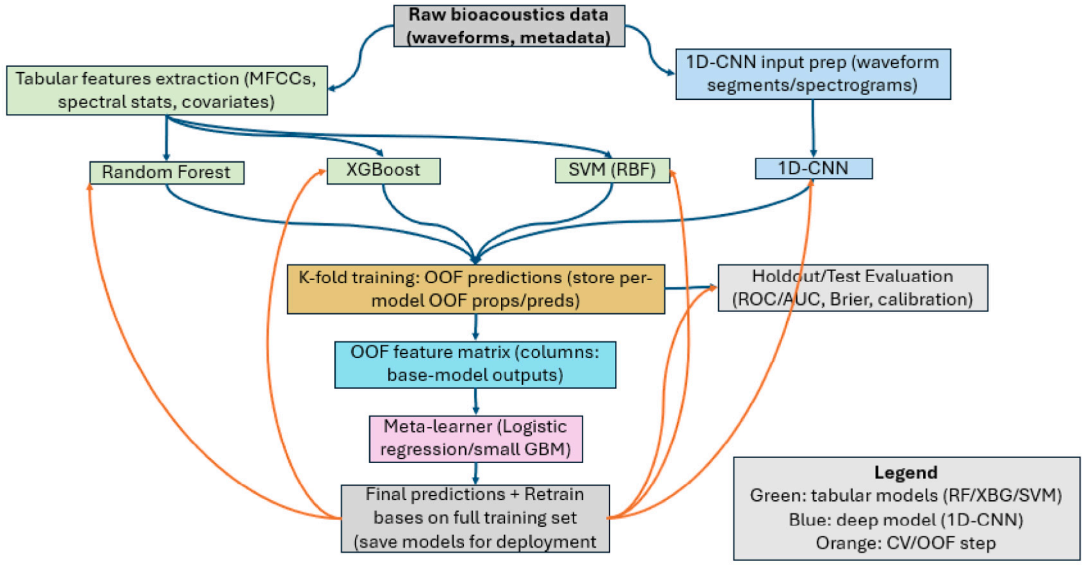


Fig. 16. Schematic of the proposed stacking ensemble pipeline. Two parallel feature streams (engineered tabular features and 1D-CNN-ready waveform/spectrogram segments) are processed by diverse base learners. OOF probability outputs from each base learner form the meta-feature matrix used to train a low-capacity meta-learner; final predictions are produced by retraining base learners on the full training set and combining their outputs through the fitted meta-learner.

Table 19
KNN grid-search and PCA settings. PCA was applied to decorrelate MFCC features; TEV is the cumulative total explained variance by the retained components.

Model	# MFCC	Original shape	Shape after PCA	TEV	Algorithm	# Neighbours	p	Weight	Best CV Acc. (%)	Grid time (s)
KNN	40	(7307, 40)	(7307, 40)	1.00	auto	5	1	uniform	97.72	10.93
KNN	50	(7307, 50)	(7307, 40)	0.98	auto	7	2	uniform	97.79	12.10
KNN	60	(7307, 60)	(7307, 40)	0.96	auto	9	1	uniform	97.72	12.21

Notes: 'Algorithm' is the neighbour-search method passed to `scikit-learn`'s `KNeighborsClassifier`; 'p' is the Minkowski exponent (1 = Manhattan, 2 = Euclidean). Best CV accuracy is the mean over cross-validation folds (reported as percent). Grid time is wall-clock tuning time on the reported machine.

each method studied we document the architecture (where applicable), the search space explored during hyperparameter tuning, and the final selected settings. We also report computational costs for hyperparameter optimisation to provide practical guidance for reproducibility and scale-up. The following subsections summarise the optimisation protocol and the best configurations found for each algorithm applied to MFCC-derived features (preprocessed with principal component analysis where indicated).

6.1. KNN parameter optimisation

We optimised the KNN classifier via grid search across plausible neighbourhood sizes and distance metrics. Input features were derived from MFCC coefficients; principal component analysis (PCA) was applied primarily for decorrelation and to stabilise optimisation; the retained component count and cumulative Total Explained Variance (TEV) are reported in Table 19, which summarises the results for three initial MFCC configurations. For each case we report TEV after PCA, the neighbour-search algorithm, the selected number of neighbours (k), the Minkowski exponent (p), the weight function used for voting, the mean cross-validated test accuracy (expressed as %), and the wall-clock time required for grid search. The results indicate a modest improvement when increasing the input MFCCs from 40 to 50 (97.72% → 97.79%), with negligible additional computational cost; further increasing to 60 MFCCs did not yield further gains. These findings support the use of PCA for compact, decorrelated representations and suggest that modest increases in MFCC dimensionality can improve KNN performance for this dataset.

6.2. SVM parameter optimisation

We performed a grid search to identify optimal SVM hyperparameters for classification using MFCC-derived features preprocessed with principal component analysis (PCA). PCA was applied to decorrelate features and reduced each input set to 30 components while preserving a high cumulative TEV. Table 20 summarises the best validation configurations for each initial number of MFCC coefficients. For each case we report TEV after PCA and the hyperparameters that yielded the best cross-validated performance (regularisation parameter C, kernel scale 'gamma', and kernel type). Reported grid-search wall-clock times provide an empirical indication of tuning cost. These results indicate the SVM is robust across input dimensionalities, with only modest adjustments to regularisation and kernel scaling required.

6.3. Random forest parameter optimisation

We performed a grid search to identify robust RF hyperparameters using MFCC-derived inputs preprocessed by PCA. PCA truncated each input set to 30 components while retaining the majority of variance (high TEV). Table 21 summarises the best hyperparameters discovered for each initial MFCC count.

For each configuration we report TEV after PCA and the RF hyperparameters that maximised validation performance: tree depth ('max_depth'), minimum samples per leaf and split, number of estimators, and the wall-clock grid-search time. These results reveal the degree of tree complexity and minimal leaf-size regularisation required under the different input dimensionalities.

6.4. Naive Bayes parameter optimisation

We performed a focused grid search to determine the Naive Bayes (NB) configuration that yields robust classification performance on MFCC-derived feature sets preprocessed with principal component analysis (PCA). For these experiments PCA was applied to decorrelate

Table 20
SVM grid-search and PCA settings. PCA reduced feature dimensionality to 30 components while retaining high TEV.

Model	# MFCC	Original shape	Shape after PCA	TEV	C	Gamma	Kernel	Grid time (s)
SVM	40	(5845, 40)	(5845, 30)	0.97	10	auto	rbf	24.58
SVM	50	(5845, 50)	(5845, 30)	0.94	10	scale	rbf	23.60
SVM	60	(5845, 60)	(5845, 30)	0.91	1	auto	rbf	26.63

Notes: TEV = total explained variance after PCA. ‘gamma’ values such as ‘auto’/‘scale’ are shown verbatim; numeric values (if used) will align via ‘siunitx’. Grid times are wall-clock seconds measured on the tuning workstation.

Table 21
RF grid-search and PCA settings. PCA reduced features to 30 components while preserving most variance.

Model	# MFCC	Original shape	Shape after PCA	TEV	max_depth	min_samples_leaf	min_samples_split	n_estimators	Grid time (s)
RF	40	(5845, 40)	(5845, 30)	0.97	None	1	2	200	61.36
RF	50	(5845, 50)	(5845, 30)	0.94	20	3	2	200	61.78
RF	60	(5845, 60)	(5845, 30)	0.91	None	1	10	200	62.56

Notes: TEV = total explained variance after PCA. ‘None‘ indicates unrestricted tree depth. ‘n_estimators’ denotes the number of trees in the forest. Grid times are wall-clock seconds on the tuning machine.

Table 22
NB grid-search and PCA settings. PCA was used to decorrelate features and a fixed 30 components were retained for subsequent classification.

Model	# MFCC	Original shape	Shape after PCA	TEV	# PCA comps	Mean test score	Std test score	Rank	Grid time (s)
Naive Bayes	40	(5845, 40)	(5845, 30)	0.97	30	0.929600	0.011571	1	0.28
Naive Bayes	50	(5845, 50)	(5845, 30)	0.94	30	0.935672	0.010552	1	0.81
Naive Bayes	60	(5845, 60)	(5845, 30)	0.91	30	0.933105	0.005837	1	0.56

Notes: TEV = total explained variance after PCA. Mean and standard deviation refer to the cross-validation test score (reported here to four decimal places). Grid-search times are wall-clock seconds measured on the tuning workstation.

Table 23
XGBOOST grid-search and PCA settings. PCA was applied to decorrelate features while preserving original dimensionality when TEV remained high.

Model	# MFCC	Original shape	Shape after PCA	TEV	colsample_bytree	learning_rate	max_depth	n_estimators	Subsample	Grid time (s)
XGBoost	40	(5845, 40)	(5845, 30)	0.97	0.70	0.10	5	300	0.70	28.16
XGBoost	50	(5845, 50)	(5845, 30)	0.94	0.90	0.20	4	300	0.80	27.79
XGBoost	60	(5845, 60)	(5845, 30)	0.91	0.90	0.10	4	300	0.90	24.73

Notes: TEV = total explained variance after PCA. ‘colsample_bytree’ and ‘subsample’ control feature- and row-level regularisation respectively; ‘n_estimators’ denotes the number of boosted trees. Grid times are wall-clock seconds measured on the tuning machine.

the features and then truncated to a fixed 30 components — this choice reflects consistently high total explained variance (TEV) across the tested initial feature counts.

Table 22 summarises the optimisation results for the different initial MFCC dimensions. For each configuration we report TEV after PCA, the number of PCA components retained, mean and standard deviation of the test score (cross-validation), the ranking of the best configuration, and wall-clock grid-search time. The results indicate stable and high predictive performance across input dimensionalities, with minimal tuning cost and low variance in validation performance, supporting the choice of 30 PCA components as a parsimonious and effective representation for NB in this task.

6.5. XGBoost parameter optimisation

We conducted a grid search to identify the optimal XGBOOST hyperparameters for classification using MFCC features preprocessed with principal component analysis (PCA). PCA was used primarily to decorrelate features while preserving the original dimensionality when TEV remained high.

Table 23 summarises the best-performing hyperparameter configurations for different initial MFCC counts. For each configuration we report TEV after PCA and the hyperparameters that produced the best validation performance: column subsampling (‘colsample_bytree‘) and row subsampling (‘subsample‘) (regularisation), learning rate, maximum tree depth (‘max_depth‘) (model complexity), and the number of estimators. We also report wall-clock grid-search time to give an empirical sense of tuning cost. These results show modest shifts in regularisation and learning rate settings as the input dimensionality changes, while the number of estimators remained consistent across experiments.

6.6. MLP parameter optimisation

We performed a systematic grid search to optimise the MLP architecture for classification using MFCC features. Principal component analysis (PCA) was applied primarily to decorrelate features; in this study we retained the original feature dimensionality (i.e., PCA served as a rotation rather than a dimensionality reduction) because the cumulative total explained variance (TEV) remained high.

Table 24 summarises the grid-search results for different initial MFCC feature counts. For each configuration we report the total explained variance after PCA, the hyperparameters that produced the best validation performance (activation, hidden-layer size, learning-rate policy, solver, and regularisation strength α), and the wall-clock time required for the grid search. The selected hyperparameters reflect the trade-offs between representational capacity and regularisation required for our dataset; runtimes provide an indication of the computational cost of tuning.

6.7. 1D-CNN parameter optimisation with one hidden layer

Table 25 reports the best hyperparameter configurations found for a 1D-CNN with a single hidden layer, when the original MFCC feature sets (40, 50, 60 coefficients) are reduced to 30 components by PCA. For each setting we report the post-PCA feature shape, the total explained variance (TEV), the best-performing hyperparameters discovered by grid search (batch size, epochs, dense units, filters, kernel size, pooling) and elapsed grid-search time.

Across the explored configurations the optimisation consistently selected the same batch size, epoch count and dense-layer width, indicating a robust, compact architecture for this task. Kernel size and

Table 24
PCA grid-search and PCA settings (PCA used for decorrelation while retaining original dimensionality).

Model	# MFCC	Original shape	Shape after PCA	TEV	Activation	Hidden	LR policy	Solver	alpha	Grid time (s)
MLP	40	(5845,40)	(5845,40)	0.97	tanh	128	constant	adam	0.0010	110.14
MLP	50	(5845,50)	(5845,50)	0.98	relu	64	constant	adam	0.0100	90.02
MLP	60	(5845,60)	(5845,60)	0.99	relu	64	constant	adam	0.0001	80.06

Notes: Shapes shown as (samples, features). TEV = total explained variance after PCA.

Table 25
1D-CNN (one hidden layer): PCA + grid-search optimisation results (vertical-rule style).

Model	#MFCCs	Orig. shape	PCA shape	TEV	Batch	Epochs	Dense	Filters	Kernel	Pool	Optimiser	Grid time (s)
1D-CNN (1 HL)	40	(5845, 40)	(5845, 30)	0.97	32	10	128	64	5	2	adam	35.51
1D-CNN (1 HL)	50	(5845, 50)	(5845, 30)	0.94	32	10	128	64	3	2	rmsprop	35.65
1D-CNN (1 HL)	60	(5845, 60)	(5845, 30)	0.91	32	10	128	64	5	2	rmsprop	30.12

Table 26
1D-CNN (two hidden layers): PCA + grid-search optimisation results.

Model	Feature shape (post-PCA)	Total Expl. Var.	Batch	Epochs	Dense	Filters (L1,L2)	Kernel (L1,L2)	Pool (L1,L2)	Grid time (s)
1D-CNN (2 HL)	(5845, 30)	0.97	32	20	64	64, 128	5, 5	2, 2	11 114.43
1D-CNN (2 HL)	(5845, 30)	0.94	64	20	128	128, 256	3, 5	2, 2	11 112.93
1D-CNN (2 HL)	(5845, 30)	0.91	64	20	256	64, 128	5, 5	2, 2	11 223.71

optimiser choices exhibited minor variation between MFCC settings, reflecting subtle dependencies of receptive-field design and optimisation dynamics on input dimensionality. Grid-search times (≈30–36 s) were short, indicating that single-hidden-layer configurations can be tuned efficiently; this makes them attractive for rapid prototyping and for deployment scenarios where compute budgets are limited.

Recommendations. For reproducible and efficient hyperparameter search we recommend:

- using randomised or Bayesian optimisers (e.g., Optuna, Hyperopt) to explore the space more efficiently than exhaustive grid search;
- adopting early stopping and short pilot runs to prune poor configurations quickly;
- employing multi-fidelity tuning (subsampled data, reduced epochs) prior to full training;
- logging experiments (MLflow, Weights & Biases or DVC) to ensure reproducibility and to enable warm-starting of subsequent searches.

These practices retain or improve performance while substantially cutting wall-clock time and compute cost, which is particularly important for lifecycle management and operational deployment.

6.8. Parameter optimisation for 1D-CNN with two hidden layers

Table 26 summarises the optimal architectures and training hyperparameters found for the 1D-CNN with two hidden layers when varying the number of MFCC features and applying PCA for dimensionality reduction. The table reports the post-PCA feature shape, total explained variance, best-performing hyperparameters discovered by grid search (batch size, epochs, dense units, filters, kernel sizes, pooling sizes and optimiser), and the elapsed grid-search time in seconds.

The results indicate consistent high performance across settings but also highlight that exhaustive grid search for DL architectures is computationally expensive: the reported grid-search times (≈11,000–11,200 s) reflect substantial compute. For practical model development we recommend complementary strategies to reduce computation while preserving search effectiveness, such as randomised or Bayesian hyperparameter search, early stopping, progressive resizing, lower-fidelity estimates (reduced epochs or subset of data), and transfer learning with fine-tuning. These approaches can dramatically reduce wall-clock time and cost while finding comparably good hyperparameters.

Given the heavy computational burden of exhaustive grid searches for 1D-CNN hyperparameters, recommended approaches entail (1) using randomised or Bayesian optimisation, such as Optuna, Hyperopt, instead of full-grid search to explore hyperparameter space more efficiently, (2) employing early stopping and reduced-epoch pilot runs to prune poor configurations quickly, (3) adopting multi-fidelity optimisation (shorter runs or smaller data subsets) before committing to full training, (4) leveraging transfer learning or pre-trained backbones where applicable, followed by light fine-tuning, (5) logging and versioning experiments like MLflow, Weights & Biases, or DVC, to ensure reproducibility and enable warm-starting of searches. These actions typically reduce search time substantially while retaining or improving model performance and reproducibility.

6.9. 1D-CNN architecture

A 1D-CNN is a neural network architecture tailored to one-dimensional sequential data such as time series or audio. A canonical design begins with a 1D input layer and proceeds through a sequence of convolutional blocks: small-kernel 1D convolutional layers (typically with ReLU activation) interleaved with pooling layers for progressive feature aggregation and dimensionality reduction. After the convolutional stack, a Flatten (or global-pooling) layer converts feature maps to a vector that feeds one or more fully connected (Dense) layers with optional Dropout for regularisation, finishing with a task-dependent output layer (e.g., sigmoid for binary classification or softmax for multi-class)(see Fig. 17).

In practice, design choices such as kernel size, number of filters per layer, pooling strategy, and the depth of the network determine the trade-off between representational capacity and inference cost. Small kernels (e.g., 3–7 samples) stacked across multiple layers allow the network to learn hierarchical temporal features, while pooling (max or average) controls temporal resolution and reduces compute. For resource-constrained deployments, shallow architectures with fewer filters and modest pooling often provide a favourable balance of accuracy and latency; for research experiments, modest increases in depth and filter count can improve performance but incur higher training and inference cost. Hyperparameters should be selected via cross-validation and evaluated in terms of both predictive metrics and operational constraints (latency, memory, power).

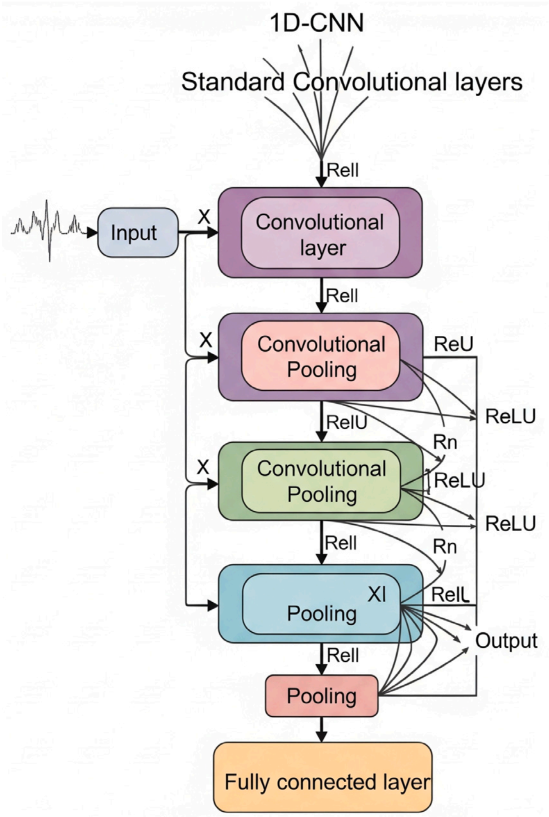


Fig. 17. Generic schematic of the 1D-CNN architecture used in this study.

6.10. 1D-DCNN architecture

A 1D-DCNN extends the 1D-CNN paradigm by using dilated convolutions to enlarge the receptive field without substantially increasing parameter count or reducing temporal resolution. The typical architecture begins with a 1D input layer for sequential data, followed by multiple blocks of dilated 1D convolutional layers (usually with ReLU activation), optionally interleaved with pooling. Dilated convolutions apply filters with a specified *dilation rate*, effectively inserting gaps between filter taps and enabling the network to model long-range dependencies while keeping networks relatively shallow. After the convolutional blocks, a Flatten (or global pooling) layer prepares features for fully connected (Dense) layers with optional Dropout for regularisation, concluding with a task-specific output layer (e.g., sigmoid for binary, softmax for multi-class)(see Fig. 18).

Compared with a standard 1D-CNN, the principal distinction is the convolution operator. Standard 1D convolutions process contiguous input segments, while dilated convolutions insert defined gaps (dilation) between kernel elements, permitting a much larger effective receptive field for the same kernel size and depth. Practically, this yields three main advantages for long sequential signals such as audio: (1) the ability to capture long-range context without deep stacking of layers, (2) improved parameter efficiency when modelling extended dependencies, and (3) preservation of temporal resolution because dilated convolutions can reduce the need for aggressive downsampling. These properties make 1D-DCNN particularly suitable for tasks where distant temporal structure is informative but excessive pooling would lose fine-grained cues.

There are trade-offs: dilation patterns must be chosen carefully (e.g., exponential vs linear schedules) to avoid blind spots in coverage, and dilated layers may be more sensitive to hyperparameter settings (kernel size, dilation rates, and regularisation). In practice, shallow dilated stacks (one or two dilated blocks) often provide a good balance of performance and inference cost for our lobster audio tasks; further gains are attainable through systematic tuning, judicious use of global pooling, and model-compression techniques for deployment on resource-constrained hardware.

Table 27
Mean acoustic characteristics extracted from 1-second representative segments across groups.

Group	Peak Freq (Hz)	Centroid (Hz)	Bandwidth (Hz)	RMS energy
Adult	179.60	1141.42	1545.52	0.217
Juvenile	102.00	1070.53	1475.33	0.440
Male	119.04	1158.67	1555.03	0.226
Female	119.76	1043.47	1465.47	0.335

7. Discussion

The extremely low p-values ($p \ll 0.001$) across all features indicate that group-specific differences are robust and reproducible. In particular, energy differences provide the strongest discriminative signal, suggesting that behavioural or physiological factors strongly influence acoustic emission intensity.

The classification results demonstrate that bioacoustic monitoring, enhanced by multi-feature fusion, is a highly effective tool for lobster population management. The high accuracies observed in Section 4 are underpinned by the distinct acoustic signatures of the different biological groups. As quantified in Table 27, there are significant physiological-acoustic correlations that the models leverage (see Fig. 19).

Notably, the Juvenile group produced nearly double the RMS energy (0.440) compared to Adults (0.217). This suggests that juvenile buzzing segments in our dataset tend to be higher-energy on average, while adult segments exhibit comparatively higher spectral centroids and broader spectral distributions. Similarly, the energy difference between Females (0.335) and Males (0.226) indicates that sex-based classification relies heavily on amplitude-related features and spectral brightness. To ensure these features were biological in origin rather than artefacts of the tank environment, control recordings of empty tanks were used to filter out pump and filtration noise, confirming that the models are detecting intrinsic morphological-acoustic differences.

7.1. Comparative analysis of sex and age classification

Binary classification of adult versus juvenile achieved the strongest results overall (Tables 9 and 11), with SVM and KNN reaching up to 98.56% accuracy. Sex classification proved slightly more challenging (peak 96.17%), suggesting that ontogenetic acoustic shifts are more salient than sex-specific cues. The MFF was critical in this regard; features such as ZCR and Spectral Contrast complemented MFCCs by capturing temporal/spectral modulation patterns within buzzing segments that can be attenuated by cepstral averaging alone.

7.2. Deployment trade-offs for edge computing

In evaluating the transition to field deployment, a critical trade-off exists between model complexity and power budgets. While DL architectures attained high accuracy, they did so at substantially higher inference times. Classical ML models — specifically KNN and SVM — offer a strategic advantage for low-power devices like the Raspberry Pi 3. Given that KNN provided the fastest inference time (0.87 ms) with competitive accuracy, it is the most viable candidate for long-term, battery-powered Passive Acoustic Monitoring (PAM) in remote fisheries environments.

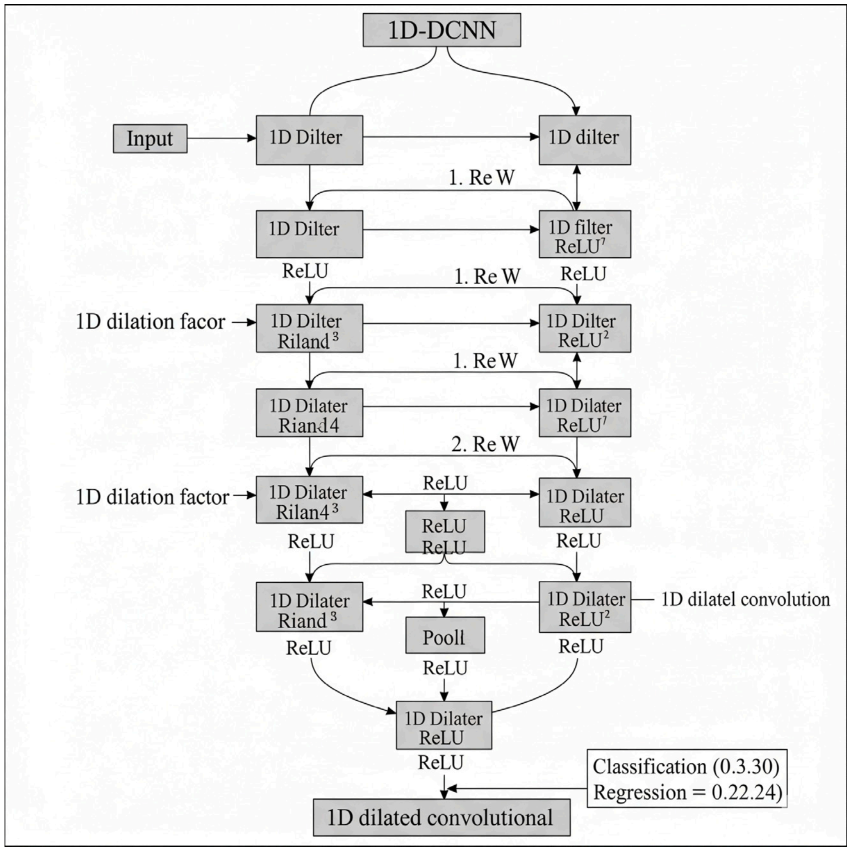


Fig. 18. Schematic of the 1D-DCNN architecture used in this study.

7.3. Policy and industry impact

Recent international legal frameworks, including the COP15 decisions and the BBNJ Agreement, have established a global mandate to protect at least 30% of jurisdictional waters by 2030. However, as noted in recent legal analyses of EU fisheries (Eklund, 2023), achieving these targets is complicated by the 'clash' between the Common Fisheries Policy (CFP) and broader environmental mandates. The AI-enhanced monitoring proposed in this study addresses this implementation gap by providing the high-resolution demographic data — such as sex and age distribution — necessary to align these competing legal regimes. By reducing scientific uncertainty, bioacoustic AI enables Member States to fulfil their environmental obligations while maintaining the collective resource-sharing principles of the CFP. Integrating automated sex and age determination into hatchery management aligns with the strategic 'opportunities' identified in recent European aquaculture reviews (Hinchcliffe et al., 2022). By streamlining the sorting process and improving the accuracy of stock assessments for sea-ranching programs, bioacoustic AI supports the EU's goal of strengthening regional food security and the 'blue economy' through technologically advanced crustacean farming.

7.4. Limitations and future work

While the dataset comprises a large number of 1-second acoustic segments, the biological replication is limited to 24 individual lobsters. As these segments are not independent biological samples, caution is required when generalising these results to wild populations or diverse environmental settings. Future studies should incorporate a larger number of individuals across multiple geographic sites to better capture intra-population acoustic variability.

Methodologically, this study focused on three specific MFCC dimensionalities (40, 50, and 60). While these were sufficient to demonstrate the effectiveness of multi-feature fusion, a broader parametric sweep would further characterise the relationship between feature dimensionality and model performance. Furthermore, while the reported 95% bootstrap confidence intervals (±0.2–0.6%) demonstrate high statistical stability across the large segment count, future work should evaluate performance in non-controlled, noisier natural habitats. To address the practical deployment concerns raised, subsequent research will prioritise the accuracy–latency trade-off, specifically targeting model compression for resource-constrained Underwater Internet of Things (UIOT) edge devices.

8. Conclusions

In conclusion, while DL models can deliver very high accuracy for lobster sound classification, their substantial training and inference costs, and the engineering required for edge deployment, often make traditional ML approaches (such as, SVM, KNN, XGBOOST) the more pragmatic choice for many real-world bioacoustic applications (Strubell et al., 2019; Sze et al., 2017; Chen and Guestrin, 2016). Efficient feature extractors such as MFCC remain powerful front-ends; when paired with compact classifiers they frequently achieve an attractive trade-off between predictive performance and computational cost (Davis and Mermelstein, 1980b; Han et al., 2015; Strubell et al., 2019).

Our results demonstrate that Homarus Gammarus buzzing/carapace-vibration signals are classifiable with AI models and that both conventional ML and DL approaches are viable for age and sex discrimination under controlled (tank) conditions as described in the current study. For operational use within PAM and underwater UIOT systems, model selection must balance discrimination against latency, power and memory constraints; model-compression and efficient-inference techniques can mitigate but not eliminate the overhead of large DL architectures (Han et al., 2015; Sze et al., 2017). Observed performance differences

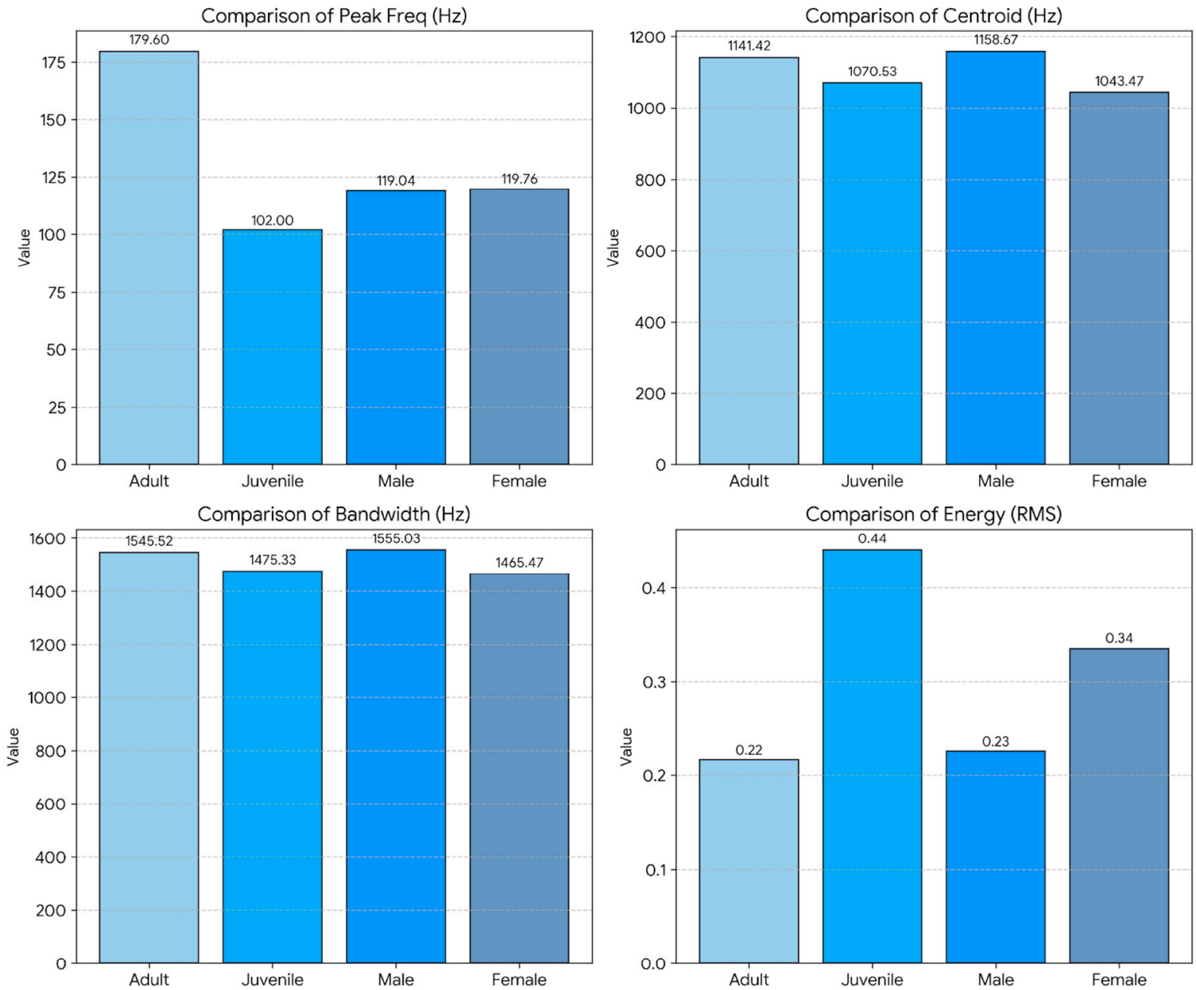


Fig. 19. Comparative analysis of extracted acoustic features across lobster demographics.

between adults vs juveniles and males vs females likely reflect dataset limitations (size, imbalance, recording quality) and weaker sex-specific acoustic cues; expanding representative sampling and improving in-situ annotation will be important next steps. Future work will explore sequence models and transformer-style architectures and perform rigorous cross-region validation to assess transferability (Gong et al., 2021; Verma and Berger, 2021).

Implementing the proposed stacking-ensemble pipeline would bring several concrete benefits. By combining diverse base learners (trees, kernel methods, and neural networks) into a stacked meta-learner, the pipeline typically improves discrimination and reduces variance relative to single models (Wolpert, 1992; Breiman, 1996b; Zhou, 2012). Boosting and bagging-based components (e.g., XGBOOST, RF) capture complementary signal structure in noisy bioacoustic data and tend to generalise more robustly across heterogeneous conditions (Chen and Guestrin, 2016; Dietterich, 2000). The OOF-based stacking design prevents information leakage and enhances reproducibility, while a regularised meta-learner often yields better-calibrated probabilistic outputs, which is a critical asset when model outputs feed management thresholds. These gains come with practical trade-offs: fold-wise training and CNN retraining increase compute and require disciplined evaluation, model registry and monitoring (drift detection, scheduled re-training). In our judgement, the modest operational overhead is justified by the improvements in accuracy, robustness and calibrated uncertainty that support management decisions.

The conservation and policy implications of this research are significant within the dual frameworks of the European Green Deal (EGD) and the Common Fisheries Policy (CFP). By providing a non-invasive, automated method for determining the sex and age structure of lobster populations, this AI-enhanced bioacoustic approach offers a scalable solution for monitoring Marine Protected Areas (MPAs) and achieving the '30 by 30' protection targets mandated by the Kunming-Montreal Global Biodiversity Framework (COP15) and the Biological Diversity of Areas Beyond National Jurisdiction (BBNJ) Agreement (Eklund, 2023). This high-precision demographic data supports a 'sustainable blue economy' by enabling the implementation of the Precautionary Approach (Melchiorre, 2024), ensuring that 'Technical Measures' under EU Regulation 2019/1241 — such as the protection of spawning biomass — are based on reliable, real-time observational data. Ultimately, this technology serves as a critical bridge between the digitalisation of fisheries and the adaptive governance required to fulfil the biodiversity objectives of the EU's International Ocean Governance agenda and the 2030 Biodiversity Strategy.

Finally, to enhance transparency and reproducibility we have published the data and code with ISO-compliant metadata and included ensemble baselines to strengthen predictive comparisons. Together, these measures advance the practical applicability of AI-enhanced bioacoustics for lobster management and conservation.

CRediT authorship contribution statement

Feliciano Pedro Francisco Domingos: Writing – original draft, Software, Investigation, Conceptualization. **Isibor Kennedy Ihianle:**

Writing – review & editing, Supervision. **Omprakash Kaiwartya:** Writing – review & editing, Supervision. **Ahmad Lotfi:** Writing – review & editing, Supervision. **Nicola Khan:** Writing – review & editing, Methodology. **Nicholas Beaudreau:** Review. **Amaya Albalat:** Writing – review & editing, Methodology. **Pedro Machado:** Writing – review & editing, Supervision, Resources, Project administration, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are profoundly grateful to Murray McBay and Co., particularly their Lobster Shop in Scotland, for their generous support in allowing us to record lobster sounds within their facility. Their invaluable assistance in sorting female, male, juvenile, and adult lobsters was crucial to this research. We extend our sincere thanks to Dr. Nicola Khan for her pivotal role in establishing the connection with Murray McBay and Co. The work reported in this article contributes to the EnviroBrain Impact Case Study.

Funding: This work was supported by a Nottingham Trent University PhD Studentship (2023). The funder had no role in the study design, data collection, analysis, interpretation of data, or preparation of the manuscript.

Appendix A. Supplementary material: stacking meta-learner defaults and ablation

A.1. Recommended default meta-learner settings

Unless otherwise stated, the stacked ensemble uses a logistic-regression meta-learner trained on the out-of-fold (OOF) probability matrix. The base learners operate on complementary feature representations: MFCC features are used for the classical machine learning models (random forest, XGBoost, and SVM), while the raw 1-second waveform (downsampled to 8 kHz) is used as input to a lightweight 1D-CNN. The meta-learner applies an L_2 penalty and selects the regularisation strength by cross-validation on the training set (i.e., using only the OOF meta-features). Concretely, we tune the inverse-regularisation parameter C over a log-spaced grid ($C \in \{10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$) and select the value that maximises the mean validation score (same metric as in the main experiments). Class weights are left unmodified unless class imbalance is severe, in which case `class_weight=balanced` is used.

A.2. Ablation: averaging vs majority-vote vs stacking

Table A.28 reports an ablation comparing three ensemble combination rules applied to the same set of base learners: (i) probability averaging, (ii) majority vote, and (iii) stacked meta-learning (logistic regression on OOF probabilities). In majority voting, all base learners contribute equally and the final prediction is simply the class receiving the highest number of votes; the procedure does not weight or prioritise machine-learning (MFCC-based) or deep-learning (raw waveform) models. All results are computed using the same train/validation protocol as the main experiments, and reported as mean ± standard deviation across folds (or repeated runs), together with the primary metric used in the paper (see Table A.29).

Table A.28
Ablation of ensemble combination strategies using the same base learners and evaluation protocol (adult vs juvenile classification).

Combination rule	Accuracy (%)	F1 (macro)	AUROC (macro)
Probability averaging	98.02	0.9774	0.9982
Majority vote	97.81	0.9750	0.9982
Stacking (LR, L_2, C via CV)	97.95	0.9766	0.9983

Table A.29
Ablation of ensemble combination strategies using the same base learners and evaluation protocol (female vs male classification).

Combination rule	Accuracy (%)	F1 (macro)	AUROC (macro)
Probability averaging	95.62	0.9543	0.9889
Majority vote	94.94	0.9474	0.9889
Stacking (LR, L_2, C via CV)	95.55	0.9535	0.9902

Table B.30
Specifications of the Raspberry Pi 3 Model B.

Category	Specification
CPU	
SoC	Broadcom BCM2837
Architecture	64-bit ARM Cortex-A53
Core count	Quad-core (4 cores)
Clock speed	1.2 GHz
RAM	
Size	1 GB LPDDR2 SDRAM
Ports	
Video output	Full-size HDMI (up to 1080p60), Composite (3.5mm TRRS)
Audio output	Analog (3.5mm TRRS), Digital (HDMI)
USB ports	4 x USB 2.0
Ethernet	10/100 BaseT RJ45
Wireless	802.11 b/g/n Wi-Fi (2.4 GHz)
Bluetooth	Bluetooth 4.1 Classic, Bluetooth Low Energy (BLE)
GPIO	40-pin header (26 GPIO, Power, Ground)
Camera interface (CSI)	15-pin MIPI CSI
Display interface (DSI)	15-pin MIPI DSI
Storage	MicroSD Card Slot
Power Input	Micro USB (5 V/2.5A recommended)
Software (Commonly Equipped With)	
Primary OS	Raspberry Pi OS (Debian-based Linux)
Desktop environment	LXDE
Included software	Chromium Browser, LibreOffice (optional), Python, Scratch, etc.
Other compatible OS	Ubuntu MATE/Server, Windows 10 IoT Core, LibreELEC, RetroPie, RISC OS, etc.

Appendix B. Specifications of the Raspberry Pi 3 Model B

Table B.30 provides a comprehensive technical overview of the Raspberry Pi 3 Model B. It details various hardware and software aspects, categorised into sections like CPU, RAM, Ports (covering video, audio, USB, Ethernet, wireless, Bluetooth, GPIO, camera, display, storage, and power), and Software (listing the primary OS, desktop environment, included software, and other compatible operating systems). Essentially, it serves as a detailed spec sheet for the Raspberry Pi 3 Model B.

Appendix C. Hardware and software resources

Table C.31 provides a concise overview of the key resources (both hardware and software) utilised in a project involving dataset collection and processing. It lists each resource, such as Raspberry Pi 3, Hydrophone, Python 3.12.3, PyAudio, Laptop (PC), ThinkPad, Keras, Scikit-learn, and TensorFlow, alongside a brief description of its role or function in the project. The table essentially acts as a quick reference guide to the technological stack employed. The details of the computer used is Table C.32.

Table C.31
Hardware and Software resources for dataset collection and processing.

Resources	Description
Raspberry Pi 3	Hardware and software equipped with CPU, Ports, RAM
Hydrophone	Hardware, underwater audio sensor
Python 3.12.3	Software, development environment
PyAudio	Software, library resource for recording sound.
Laptop (PC), ThinkPad	Housing Python, SSH, networking capabilities.
Keras	High-level neural networks API written in Python
Scikit-learn	Python library for ML, built on top of NumPy, SciPy, and Matplotlib, offer some basic tools for building and training deep neural networks
TensorFlow	An open-source ML framework developed by Google, designed to build and train various types of neural networks

Table C.32
Technical specifications of the Lenovo ThinkPad X1 Carbon Gen 13 Aura Edition (PF-2D95WH).

Category	Specification
Performance	Intel Core Ultra (Series 2) processor on Intel vPro, Evo Edition platform; integrated AI engine with up to 13 TOPS AI performance; optimised performance per watt for extended battery life.
AI functions	Local AI features including enhanced video conferencing, document automation, email management, scheduling, improved responsiveness, and enhanced data privacy.
Display	14-inch OLED, 2.8K (2880 × 1800) resolution, 120 Hz refresh rate, 400–500 nits brightness, 100% DCI-P3 colour gamut, DisplayHDR True Black 500 certified.
Memory and storage	Up to 64 GB soldered LPDDR5x (8400 MT/s) RAM; up to 2 TB PCIe Gen5 SSD.
Design and construction	Lightweight design <1 kg; chassis built from recycled carbon fibre, magnesium, and plastic; MIL-STD-810H certified durability.
Security	ThinkShield suite: discrete TPM (dTPM), biometric authentication, AI-enhanced threat detection.
Usability	Haptic TouchPad; TrackPoint with double-tap menu for audio/video settings; Dictation Toolbar with speech-to-text support.
Connectivity	USB-C Thunderbolt 4, USB-A (full-size), and other modern I/O options.
Power	Rapid Charge technology; compact GaN charger for fast charging.

Appendix D. Hardware: Hydrophone features

Table D.33 provides a detailed breakdown of the key characteristics and capabilities of a specific hydrophone device.

Table D.33
Hydrophone specification — Aquarian Audio Model H2d (3.5 mm cable termination).

Specification	Model H2d (Aquarian Audio)
Manufacturer	Aquarian Audio
Model	H2d
Termination	3.5 mm cable termination (note: termination only — electrical/acoustic specifications identical to other termination options)
Sensitivity	−172 dB re: 1 V/μPa (±4 dB, 20 Hz–4 kHz). Approx. sensitivity at 100 kHz: -210 dB re: 1 V/μPa.
Useful range	<10 Hz to >100 kHz (not measured above 100 kHz).
Polar response	Omnidirectional (horizontal).
Operating depth	<80 m.
Output impedance	1.1 kΩ (typical).
Power (current)	0.7 mA (typical).
Minimum PIP requirement	2 V, 0.7 mA.
Dimensions	25 mm × 46 mm.
Mass	105 g.
Specific gravity	5.3.

Note: The "3.5 mm cable termination" refers only to the connector termination. All acoustic and electrical specifications in the table apply to the H2d hydrophone irrespective of connector type unless otherwise specified. Replace any placeholder numbers if you have manufacturer updates.

Data availability

In accordance with FAIR principles, the lobster sound dataset is available on Mendeley Data (DOI: https://doi.org/10.17632/WSM4VFXVSF; landing page: https://data.mendeley.com/datasets/wsm4vfxvsf) Domingos (2025a). The main preprocessing/training code is available at https://github.com/N0736086/lobster_soundscodes Domingos (2025b), and the stacking ensemble script used in this study is available at https://github.com/N0736086/stacking-ensemble-pipeline/blob/main/stack_ensemble_pipelie.pyDomingos (2025).

References

Araújo, M.B., Guilhaumon, F., 2018. Climate change impacts on the distribution of coastal lobsters. Mar. Biol. 165 (12), 1–14. http://dx.doi.org/10.1007/s00227-018-3441-9, URL: https://link.springer.com/article/10.1007/s00227-018-3441-9.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.H., Frommolt, K.H., 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognit. Lett. 31 (12), 1524–1534. http://dx.doi.org/10.1016/j.patrec.2009.09.014, URL: https://www.sciencedirect.com/science/article/pii/S0167865509002487. Pattern Recognition of Non-Speech Audio.

Barroso, V., Xavier, F., Ferreira, C., 2023. Applications of machine learning to identify and characterize the sounds produced by fish. ICES J. Mar. Sci. 80 (7), 1854–1867. http://dx.doi.org/10.1093/icesjms/fsad126.

Bharath, S., et al., 2025. Underwater sound classification using learning-based methods: A review. Expert Syst. Appl. 240, 122236. http://dx.doi.org/10.1016/j.eswa.2024.122236, URL: https://www.sciencedirect.com/science/article/pii/S0957417424013654.

Bouwma, P.E., Herrnkind, W.F., 2009. Sound production in caribbean spiny lobster panulirus argus and its role in escape during predatory attack by Octopus briareus. N. Z. J. Mar. Freshw. Res. 43 (1), 3–13. http://dx.doi.org/10.1080/00288330909509977.

Breiman, L., 1996a. Bagging predictors. Mach. Learn. 24 (2), 123–140. http://dx.doi.org/10.1023/A:1018054314350.

Breiman, L., 1996b. Stacked regressions. Mach. Learn. 24 (1), 49–64. http://dx.doi.org/10.1007/BF00117832.

Breiman, L., 2001a. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breiman, L., 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statist. Sci. 16 (3), 199–231. http://dx.doi.org/10.1214/ss/1009213726.

Briones-Fourzán, P., Domínguez-Gallegos, R., Lozano-Álvarez, E., 2014. Aggressive behaviour of spotted spiny lobsters (Panulirus guttatus) in different social contexts: The influence of sex, size, and missing limbs. ICES J. Mar. Sci. 72 (suppl_1), i155–i163. http://dx.doi.org/10.1093/icesjms/fsu219, arXiv:https://academic.oup.com/icesjms/article-pdf/72/suppl_1/i155/31227438/fsu219.pdf.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Council of the European Union, 2025. The EU's common Fisheries policy and its goals. https://www.consilium.europa.eu/en/policies/common-fisheries-policy-and-its-goals/. (Accessed 19 September 2025).

Davis, S., Mermelstein, P., 1980a. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366. http://dx.doi.org/10.1109/TASSP.1980.1163420.

Davis, S., Mermelstein, P., 1980b. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366. http://dx.doi.org/10.1109/TASSP.1980.1163396.

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Multiple Classifier Systems. In: Lecture Notes in Computer Science, vol. 1857, Springer, pp. 1–15.

Domingos, F.P.F., 2025. stack_ensemble_pipeline – `stack_ensemble_pipelie.py`. URL: https://github.com/N0736086/stacking-ensemble-pipeline/blob/main/stack_ensemble_pipelie.py. GitHub repository file. (Accessed 18 September 2025).

Domingos, F., 2025a. Lobster sound dataset. http://dx.doi.org/10.17632/WSM4VFXVSF, URL: https://data.mendeley.com/datasets/wsm4vfxvsf.

Domingos, F., 2025b. lobster_soundscodes. URL: https://github.com/N0736086/lobster_soundscodes.

Domingos, F.P.F., Lotfi, A., Ihianle, I.K., Kaiwartya, O., Machado, P., 2024. Underwater communication systems and their impact on aquatic life—A survey. Electronics 14 (1), 7. http://dx.doi.org/10.3390/electronics14010007.

Eklund, A., 2023. The Clash of Fisheries: How can the European Union Improve Conservation and Management of Fish Stocks in Union Waters to Ensure Biological Diversity of its Seas?. (Master's thesis). University of Helsinki, URL: https://urn.fi/URN:NBN:fi-fe2023051043337.

European Commission, Directorate-General for Maritime Affairs and Fisheries, 2025. Common Fisheries policy (CFP): Overview. https://oceans-and-fisheries.ec.europa.eu/policy/common-fisheries-policy-cfp_en. (Accessed 19 September 2025).

Gong, Y., Chung, Y.A., Glass, J., 2021. AST: Audio spectrogram transformer. Interspeech 2021 Proceedings / arXiv preprint. URL: https://arxiv.org/abs/2104.01778.

Günen, M.A., 2022. Performance comparison of deep learning and machine learning methods in determining wetland water areas using EuroSAT dataset. Environ. Sci. Pollut. Res. 29 (14), 21092–21106. http://dx.doi.org/10.1007/s11356-021-17177-z.

Han, S., Mao, H., Dally, W.J., 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv Preprint abs/1510.00149. URL: https://arxiv.org/abs/1510.00149.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, http://dx.doi.org/10.1007/978-0-387-21606-5.

Hinchcliffe, J., Agnalt, A.L., Daniels, C.L., Drengstig, A., Lund, I., McMinn, J., Powell, A., 2022. European lobster homarus gammarus aquaculture: Technical developments, opportunities and requirements. Rev. Aquac. 14 (2), 919–937. http://dx.doi.org/10.1111/raq.12634, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/raq.12634. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12634.

Huang, S., Tang, J., Dai, J., Wang, Y., 2019. Signal status recognition based on 1DCNN and its feature extraction mechanism analysis. Sensors 19 (9), 2018. http://dx.doi.org/10.3390/s19092018.

Hwang, J., et al., 2019. Unveiling microplastic contamination in seafood: Source, fate and biological effects. Marine Poll. Bull. 146, 110–118. http://dx.doi.org/10.1016/j.marpolbul.2019.05.017, URL: https://www.sciencedirect.com/science/article/pii/S0025326X25008896.

Islam, Y., Das, S.C., Chowdhury, M.J.U., 2025. The role, trends, and applications of machine learning in undersea communication: A Bangladesh perspective. http://dx.doi.org/10.48550/arXiv.2503.00669, arXiv preprint arXiv:2503.00669.

Jézéquel, Y., Bonnel, J., Coston-Guarini, J., Guarini, J.M., Chauvaud, L., 2018. Sound characterization of the European lobster homarus gammarus in tanks. Aquatic Biol. 27, 13–23. http://dx.doi.org/10.3354/ab00692.

Jézéquel, Y., Chauvaud, L., Bonnel, J., 2020a. Spiny lobster sounds can be detectable over kilometres underwater. Sci. Rep. 10 (1), 7943. http://dx.doi.org/10.1038/s41598-020-64830-7.

Jézéquel, Y., Coston-Guarini, J., Chauvaud, L., Bonnel, J., 2020b. Acoustic behaviour of male European lobsters (Homarus gammarus) during agonistic encounters. J. Exp. Biol. 223 (4), jeb211276. http://dx.doi.org/10.1242/jeb.211276.

Jézéquel, Y., Jones, I.T., Bonnel, J., Chauvaud, L., Atema, J., Mooney, T.A., 2021. Sound detection by the American lobster (Homarus Americanus). J. Exp. Biol. 224 (6), jeb240747. http://dx.doi.org/10.1242/jeb.240747.

Kerkhofs, R., 2024. To hum or not to hum: Analyzing and provoking sound production in the American lobster (Homarus Americanus). Bowdoin Library.

Kikuchi, M., Akamatsu, T., Takase, T., 2015. Passive acoustic monitoring of Japanese spiny lobster stridulating sounds. Fish. Sci. 81, 229–234. http://dx.doi.org/10.1007/s12562-014-0835-6.

Looby, A., Vela, S., Cox, K., Riera, A., Bravo, S., Davies, H.L., Rountree, R., Reynolds, L.K., Martin, C.W., Matwin, S., et al., 2023. FishSounds version 1.0: A website for the compilation of fish sound production information and recordings. Ecol. Inform. 74, 101953. http://dx.doi.org/10.1016/j.ecoinf.2022.101953.

Malfante, M., Mars, J.I., Dalla Mura, M., Gervaise, C., 2018. Automatic fish sounds classification. J. Acoust. Soc. Am. 143 (5), 2834–2846. http://dx.doi.org/10.1121/1.5036628.

McGarvey, R., Punt, A.E., Matthews, J.M., Feenstra, J.E., Gardner, C., Burch, P., Hartmann, K., Linnane, A., 2015. Comparing size-limit and quota policies to increase economic yield in a lobster fishery. Can. J. Fish. Aquat. Sci. 72 (9), 1292–1305. http://dx.doi.org/10.1139/cjfas-2014-0405.

Melchiorre, T., 2024. The European union and the arctic: The case of Fishery. In: The Yearbook of Polar Law XV. Brill Nijhoff, pp. 231–257.

Mente, E. (Ed.), 2008. Reproductive Biology of Crustaceans: Case Studies of Decapod Crustaceans, first ed. CRC Press, http://dx.doi.org/10.1201/9781439843345.

Niu, H., Li, X., Zhang, Y., Xu, J., 2023. Advances and applications of machine learning in underwater acoustics. Intell. Mar. Technol. Syst. 1 (8), 1–28. http://dx.doi.org/10.1007/s44295-023-00005-0, URL: https://link.springer.com/article/10.1007/s44295-023-00005-0.

Noda, J.J., Travieso, C.M., Sánchez-Rodríguez, D., 2016. Automatic taxonomic classification of fish based on their acoustic signals. Appl. Sci. (ISSN: 2076-3417) 6 (12), http://dx.doi.org/10.3390/app6120443, URL: https://www.mdpi.com/2076-3417/6/12/443.

Patek, S.N., 2002. Squeaking with a sliding joint: mechanics and motor control of sound production in palinurid lobsters. J. Exp. Biol. 205 (16), 2375–2385. http://dx.doi.org/10.1242/jeb.205.16.2375.

Phillips, B.F., Wahle, R.A., Ward, T.M., et al., 2015. Lobsters: ocean icons in changing times. ICES J. Mar. Sci. 72 (suppl_1), i1–i5. http://dx.doi.org/10.1093/icesjms/fsv002, URL: https://academic.oup.com/icesjms/article/72/suppl_1/i1/622676.

Radhakrishnan, E., et al., 2020. Ecology and global distribution pattern of lobsters. In: Lobsters: Biology, Fisheries and Aquaculture. Springer, pp. 87–112. http://dx.doi.org/10.1007/978-981-32-9094-5_5, URL: https://link.springer.com/chapter/10.1007/978-981-32-9094-5_5.

Rountree, R.A., et al., 2023. Applications of machine learning to identify and characterize the acoustic signatures of fishes: A review. ICES J. Mar. Sci. 80 (7), 1854–1875. http://dx.doi.org/10.1093/icesjms/fsad085, URL: https://academic.oup.com/icesjms/article/80/7/1854/7246570.

Shabani, S., Kamio, M., Derby, C.D., 2009. Spiny lobsters use urine-borne olfactory signaling and physical aggressive behaviors to influence social status of conspecifics. J. Exp. Biol. 212 (15), 2464–2474. http://dx.doi.org/10.1242/jeb.026492.

Sill, J., 2009. Feature-weighted linear stacking. arXiv:0911.0460. arXiv.

Staaterman, Claverie, Patek, 2010. Disentangling defense: The function of spiny lobster sounds. Behaviour 147 (2), 235–258. http://dx.doi.org/10.1163/000579509X12523919243428.

Strubell, E., Ganesh, A., McCallum, A., 2019. Energy and policy considerations for deep learning in NLP. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3645–3650. http://dx.doi.org/10.18653/v1/P19-1355.

Sze, V., Chen, Y.H., Yang, T.J., Emer, J., 2017. Efficient processing of deep neural networks: A tutorial and survey. Proc. IEEE 105 (12), 2295–2329. http://dx.doi.org/10.1109/JPROC.2017.2761740.

Taylor, J., Patek, S., 2010. Crustacean seismic communication: Heard but not present. pp. 9–23, The use of vibrations in communication: properties, mechanisms and function across taxa. Research Signpost, Trivandrum, Kerala.

Verma, P., Berger, J., 2021. Audio transformers: Transformer architectures for large scale audio understanding. Adieu convolutions. arXiv preprint arXiv:2105.00335. URL: https://arxiv.org/abs/2105.00335. (Accessed 17 September 2025).

Weilgart, L., 2018. The impact of ocean noise pollution on fish and invertebrates. Report for OceanCare, Switzerland https://api.semanticscholar.org/CorpusID:49364618.

Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5 (2), 241–259.

Yadav, R., et al., 2024. The unseen threat of the synergistic effects of microplastics and heavy metals in aquatic ecosystems. Curr. Pollut. Rep. 10, 1–20. http://dx.doi.org/10.1007/s40726-024-00298-7, URL: https://link.springer.com/article/10.1007/s40726-024-00298-7.

Yue, H., Zhang, L., Wang, D., Wang, Y., Lu, Z., 2017. The classification of underwater acoustic targets based on deep learning methods. In: 2017 2nd International Conference on Control, Automation and Artificial Intelligence. CAAI 2017, Atlantis Press, pp. 526–529. http://dx.doi.org/10.2991/caai-17.2017.118.

Zhou, Z.H., 2012. Ensemble Methods: Foundations and Algorithms. Chapman & Hall/CRC, Boca Raton, FL.