# Optimizing Machine Learning for the Detection of Splashing Sounds for Automatic Monitoring of Fish Spawning

Line Weiss, Lisa Dörner, Vincent S. Kather, Max de Koning, Kees te Velde, Ine Pauwels, Christian Tudorache, and Hans Slabbekoorn

## Contents



**OPEN ACCESS** with major support from JASCO APPLIED SCIENCES

L. Weiss
Leiden University, Leiden, The Netherlands
e-mail: l.weiss@umail.leidenuniv.nl

L. Dörner (✉) · C. Tudorache · H. Slabbekoorn
Leiden University, Institute of Biology, Leiden, The Netherlands
e-mail: l.doerner@biology.leidenuniv.nl; c.tudorache@biology.leidenuniv.nl; h.w.slabbekoorn@biology.leideuniv.nl

V. S. Kather
Naturalis Biodiversity Center, Understanding Evolution, Leiden, The Netherlands
e-mail: vincent.kather@naturalis.nl

M. de Koning
Wageningen University & Research, Wageningen, The Netherlands
e-mail: max.dekoning@outlook.com

K. te Velde
Naturalis Biodiversity Center, Leiden, The Netherlands
e-mail: kees.tevelde@naturalis.nl

I. Pauwels
Research Institute for Nature and Forest, Brussels, Belgium
e-mail: ine.pauwels@inbo.be



A. N. Popper et al. (eds.), *The Effects of Noise on Aquatic Life IV*,
https://doi.org/10.1007/978-3-031-94229-7_177-1

**Abstract**

Passive acoustic recordings can play an important role in biodiversity monitoring and management of human activities, with growing potential due to artificial intelligence. The twaite shad *Alosa fallax* is a clupeiformes fish found from the northwestern Atlantic to the Mediterranean. It spawns in freshwater and migrates upstream from the sea to rivers in spring. Twaite shad are classed as a threatened species. As part of their spawning behavior, males chase females in circles close to the water surface, producing a distinctive splashing sound that can be used for passive acoustic monitoring. In reintroduction programs, this can be used to measure spawning activity and locate spawning sites. As manual identification of shad splashing events in large audio datasets is time-consuming, an automatic shad splashing detection system is required. Here, embeddings from seven pretrained AI models from different source-domains were used to train a classifier for shad splashing detection. The model BirdNET provided the most informative embeddings. A class imbalance in training data composition was beneficial for classifier performance, and filtering for high-quality positive examples was detrimental. This study shows the value of retraining publicly available, pretrained machine learning algorithms to detect novel challenging sound events against a noisy background.

**Keywords**

Artificial intelligence · Automated acoustic detection · Biodiversity monitoring · Migratory fish · Noise pollution · Sound event detection · Twaite shad · Supervised learning

## Introduction

Ecoacoustics is the study of environmental sounds, collectively referred to as the *soundscape*, and their connection to ecological processes (Alcocer et al. 2022). Many aquatic species perceive and rely on the soundscape for purposes, such as orientation, foraging, defense, communication, mate attraction, and detection of prey and predators (te Velde et al. 2024). Consequently, these species can also be impacted by anthropogenic noise through deterrence, disturbance, distraction, and masking of important sounds.

Underwater noise levels in marine ecosystems have received increased attention in recent decades. Marine mammal strandings in association with sonar exercises triggered the study of noise pollution in aquatic habitats and generated interest from

regulators and the public. In contrast to the marine environment, freshwater systems remain understudied in this regard. Since 2008, European Union (EU) law has banned the addition of energy (including sound) into marine waters if it harms good environmental status. However, no equivalent legal provisions exist for freshwater systems.

Freshwater anthropogenic noise sources vary, including shipping traffic, recreational water activities, urban noise, and land-based traffic from bridges and nearby roads (te Velde et al. 2024). With one-third of global freshwater fish species facing extinction, freshwater ecosystems belong to the most vulnerable ecosystems in the world, highlighting the need to investigate the impact of anthropogenic noise on their resident species.

Underwater sound conditions and natural acoustic events could provide environmental cues to migratory fish. Fishes may use local soundscapes or distant sound events when making spatial decisions about whether to stay or leave and to select a direction. Rivers often serve as migratory routes for fishes, yet they can also be busy shipping lanes and often flow through noisy urban areas (te Velde et al. 2024). These artificially noisy conditions can mask natural cues and can disturb, deter, and distract fishes on migration through riverine systems. One such migratory species is the anadromous twaite shad *Alosa fallax* (Lacepède), which ranges from the northwestern Atlantic to the Mediterranean. Mature twaite shad start their spawning migrations between February and May (Aprahamian et al. 2003).

As part of their courting behavior, twaite shad males chase females in circles close to the water surface and slap their tails, resulting in loud, arrhythmic splashing sounds lasting between 3 and 10 s. These specific sounds can be used for acoustic monitoring (Aprahamian et al. 2003). Passive acoustic monitoring (PAM) offers a powerful, continuous, and noninvasive means to study this behavior. In-air recordings have already been effectively used to record twaite shad spawning activity (de Koning et al., unpublished). However, manual sound-event detection is a time-consuming process that requires prior knowledge and familiarity with the sound of interest. Additionally, since twaite shad splashing events are difficult to distinguish from other surface water disturbances, it is difficult to recognize them under certain environmental conditions. For brevity's sake, the sound event will be referred to as shad splashing henceforth.

Machine learning models are increasingly used to automate sound-event identification in long recordings, where manual identification of sound-events is too time-consuming (Bianco et al. 2019). Machine learning uses statistical techniques to automatically identify patterns in data and either predict future data or make decisions based on existing uncertain data (Bianco et al. 2019). Supervised machine learning models are trained with pairs of data and corresponding labels to learn prediction mapping. These models usually perform the task they were trained on with high accuracy but generalize poorly to new settings (Bianco et al. 2019). Convolutional neural networks (CNNs) are a common architecture for deep learning models, which are commonly trained in a supervised learning setup. They have been successfully applied for the identification of birds, frogs, and bats based on their vocalizations, but they require large training datasets (Ruff et al. 2021).

Recently, researchers in the field of computational bioacoustics have been evaluating the use of embeddings of existing models (Ghani et al. 2023) for a variety of downstream tasks that the model was not originally given. Embeddings are vector representations that are generated by deep learning models and subsequently passed to the classifiers to predict which classes are present in a given datum. In bioacoustics, embeddings represent the features of audio data in a mathematical format, which can also be used as input to train new classifiers to predict sound events other than the classes the model was originally trained on. For example, a model originally trained on bird song can be trained to predict fish sounds. This process is also referred to as linear probing. This approach benefits from the availability of pretrained models, which have already learned to detect specific sounds against background sounds and therefore vastly reduce the amount of data required to train a classifier for a novel sound (Ghani et al. 2023).

Developing a classifier to monitor shad splashing presents several challenges. Shad splashing is a rare sound event, leading to limited data availability. This often results in class-imbalanced datasets, with more examples for the noise class than the target sound event class, which can bias the classifier toward the majority class (Ghani et al. 2023). When working with such imbalanced datasets, model performance can be assessed with the F1 score, which is the harmonic means of precision and accuracy. When both precision and accuracy have a perfect score of 1.0, so does the harmonic mean. Shad splashing sounds also often resemble other environmental water sounds, further complicating the classification of target sound event versus noise.

Despite these challenges, progress has been made toward developing a shad splashing detector. Diep et al. (2016) developed a Gaussian mixture model (GMM) for shad detection, which achieved moderate performance with an F1 score of 0.69. Sota et al. (2023) later trained a recurrent convolutional neural network (CRNN) to detect shad splashing using a large but imbalanced dataset, achieving an F1 score of 0.52. Despite the large training dataset and more complex architecture, this model also achieved moderate results. Most recently, a convolutional neural network (CNN) developed by McLain and White (unpublished) with a smaller but balanced dataset reportedly achieved high performance metrics, with an F1 score of 0.97. However, testing on a balanced dataset can lead to optimistic performance estimates, which may not reflect the model's effectiveness in realistic, imbalanced scenarios where the target event is rare.

This chapter assesses whether a shad splashing detector trained on embeddings from a pretrained model achieves comparable or improved performance compared to models trained from scratch with limited labelled data. Additionally, suitable training conditions for automatic detection of shad splashing were investigated, comparing existing embedding models to determine which one most effectively extracts relevant acoustic features of shad splashing. Finally, classifier performance was compared across differently structured datasets to assess how training data composition affects classifier performance. The aim was to answer the following questions: Which publicly available pretrained AI model provides the most effective embeddings for training a shad splash detector? In addition, what training dataset composition leads to the best classifier performance?

# Methods

## Recording Origins and Annotation

Recordings used in this study were collected along the River Severn, its tributary Teme (UK, 2022), and the Scheldt (BE, 2023, 2024), all of which are known shad spawning rivers. Recordings were made with AudioMoth (model 1.2.0, Open Acoustic Devices) and Song Meters (Micro and SM1, Wildlife Acoustics) (Table 1) at a sampling rate of 48 kHz. The gain of the AudioMoth devices was set to 30.6–31.6 dB and for the Song Meter models to 18 dB. All recording devices were mounted on structures at the shore, directed toward the water. Recordings were manually annotated through auditory and visual inspection using Audacity (version 3.7.4.0, Muse Group), applying a Hann window with a window size of 2048 samples.

A shad splashing event was defined as having a minimum duration of 1.5 s to ensure the audible presence of arrhythmic tail slaps. Shad splashing events with central interruptions longer than 3 s were annotated as two separate events. In ambiguous cases, annotations were reviewed by an experienced second annotator. The remaining parts of the recordings were labelled as noise, which included a diversity of nontarget sounds, such as weirs, water flow, wind, rain, bird chorus, road traffic, and boats. The annotated dataset spanned 28 hours, 40 min, and 38 s of audio material containing 351 annotated shad splashing events (Table 1).

## Evaluation of Embedding Models

The performance of linear classifiers trained with embeddings from seven models was evaluated to determine which model provided the most useful features for shad

**Table 1** Contribution of each sampling location in terms of recording duration and number of positive sound events to the annotated dataset. Abbreviations: shad splashing event (SSE), Upper Lode weir (ULW), AudioMoth (AM), Song Meter Micro (SMM), Song Meter SM1 (SM1)

| Year | River | Location | Recorder | Recording duration | % Total duration | # SSE | % Total SSE count |
|---|---|---|---|---|---|---|---|
| 2022 | Severn | ULW (100 m) | AM | 06:00:00 | 20.8 | 39 | 11.1 |
| 2022 | Severn | ULW (1 km) | SMM | 02:59:48 | 10.4 | 52 | 14.8 |
| 2022 | Teme | Powick | SMM | 03:59:44 | 13.87 | 42 | 12 |
| 2023 | Schelde | Branst (Jetty) | SM1 | 03:59:57 | 13.9 | 163 | 46.4 |
| 2023 | Schelde | Branst (Buoy) | SM1 | 02:51:14 | 9.9 | 17 | 4.8 |
| 2025 | Schelde | Branst (Jetty) | AM | 04:59:35 | 17.3 | 31 | 8.8 |
| 2025 | Schelde | Mariekerke | AM | 03:59:40 | 13.6 | 7 | 2 |

splashing detection. Models were selected based on their expected ability to handle broadband signals (NonBioAVES_ESpecies [Hagiwara 2023], Insect66NET [Faiß and Stowell 2023]), their prior application to aquatic contexts (Google_Whale [Allen et al. 2024], SurfPerch [Williams et al. 2025]), or their embedding capabilities and performance in broader bioacoustics applications (BirdNET [Kahl et al. 2021], Perch_Bird [Ghani et al. 2023], BioLingual [Robinson et al. 2024]). For each model, embeddings were generated and used to train a linear classifier. Embedding extraction, classifier training, and evaluation were performed using the BacPipe pipeline (Python version 3.10.16) available on GitHub (https://github.com/bioacoustic-ai/bacpipe; Kather et al. 2025). This pipeline streamlines the generation of embeddings from full-length recordings and provides evaluation tools when corresponding annotation data are available. All embedding models were evaluated on the entire annotated dataset, which features an approximate noise-to-shad splashing ratio of 100:1.

## Training Conditions Assessment

Seven classifiers (labelled A–G) were trained on training sets, which varied in terms of noise-to-shad splashing ratio, noise subsample, and shad splashing subsample (Table 2) using the BirdNET Analyzer GUI. Noise clips were generated by first creating a label file that segmented noise portions of the recordings into 3-s segments using a Python script (version 3.13.5). These segments were then extracted using Audacity, resulting in 34,300 noise clips. For the training dataset, 281 shad splashing clips (80%) were randomly selected. The same number of noise clips was randomly sampled for the training dataset of classifier A.

To increase the noise-to-shad splashing ratio to 2:1 for classifier B and 5:1 for classifier C, additional randomly selected noise clips were introduced. For classifiers D, E, and F, noise clips were randomly resampled, which resulted in a

**Table 2** Overview of training dataset variants. Abbreviations: shad splashing event (SSE)

| Model ID | Training set Noise:SSE ratio | Training set # clips | Noise subsample | SSE subsample | Test set # clips |
|---|---|---|---|---|---|
| A | 1:1 | 281:281 | Fraction of B's sample | Base | 700:70 |
| B | 2:1 | 562:281 | Fraction of C's sample | Base | 700:70 |
| C | 5:1 | 1405:281 | Base | Base | 700:70 |
| D | 2:1 | 281:281 | Different sample 1 (overlap with C's sample) | Base | 700:70 |
| E | 2:1 | 281:281 | Different sample 2 | Base | 700:70 |
| F | 2:1 | 281:281 | Different sample 3 | Base | 700:70 |
| G | 2:1 | 316:158 | Fraction of B's sample | Without low-quality SSEs | 700:70 |

partial overlap of six noise clips between the training datasets of classifiers C and D and no overlap for classifiers E and F. For classifier G, low-quality shad splashing events, barely visible in the spectrogram with low signal-to-noise ratios, were removed from the sound subset. The noise class for training classifier G was based on classifier B's noise samples, maintaining a 2:1 noise-to-shad splashing ratio. All classifiers were evaluated on the same test dataset, comprising 3-s clips and a rare-occurrence noise-to-shad splashing ratio of 10:1 between the clips.

Manual performance evaluation was conducted using the confidence values output by the GUI's batch analysis tool for the test dataset (Table 2). The minimum confidence threshold was set to 0.05. A confidence table was constructed by assigning a confidence value of 0 to clips with predicted confidence below this threshold. Receiver operating characteristic (ROC) and precision–recall (PR) curves were computed using the R packages *pROC* and *PRROC* (Fig. 4) to illustrate the trade-off between the model's ability to avoid false positives and identify true positives. The optimal classification threshold was defined as the value that maximized the F1 score, which is the harmonic mean of precision and recall and balances false positives and false negatives. Mel spectrograms were generated for all true positive, false positive, and false negative predictions using the librosa library in Python (version 3.13.5) with a fast Fourier transform (FFT), with a window size of 2048 and a hop length of 128. These parameters resulted in a frequency bin width of ~23.44 Hz and a time step of 2.67 ms at a sampling rate of 48 kHz. A preliminary qualitative performance evaluation was conducted by inspecting predictions on 5 hours of continuous, unedited field recordings, generated with a sliding window step size of 1.5 s.

## Results

### Evaluation of Embedding Models for Acoustic Feature Extraction

BirdNET outperformed all other embedding models with a macro F1 score above 0.8 and a macro accuracy above 0.7. The micro metrics and AUC values near 1 for all classifiers indicate strong overall performance (Fig. 1). However, while all classifiers achieved high per-class accuracy for noise, their accuracies for shad splashing events were substantially lower. BirdNET performed best, correctly identifying over 50% of the positive sound events (Fig. 2).

### Classifier Performance Across Training Data Variants Using BirdNET-Derived Features

Classifiers D and G achieved the highest true positive count (D (tp) = 60, G (tp) = 64) but also produced higher false positive counts compared to the other classifiers (D (fp) = 79, G (fp) = 164), except for classifier E (Fig. 3). Classifier E yielded the highest false positive count without improving on true positives and false

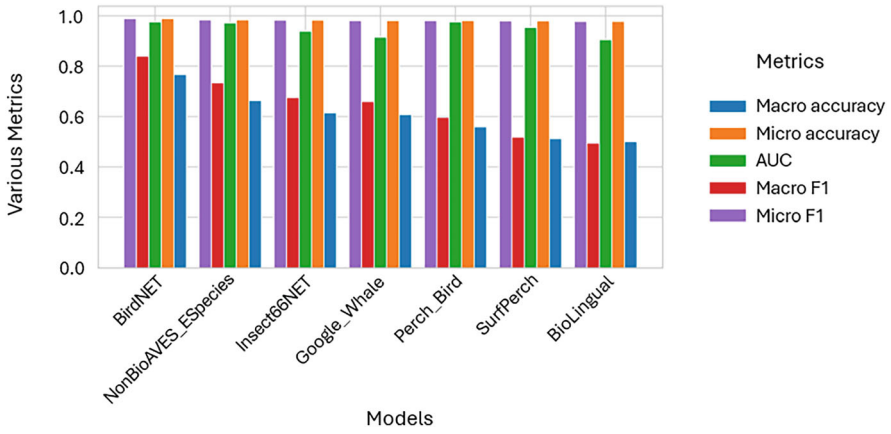


**Fig. 1** Evaluation metrics (macro accuracy, micro accuracy, ROC AUC, macro F1 score, micro F1 score) achieved by classifiers trained with embeddings from seven pretrained models

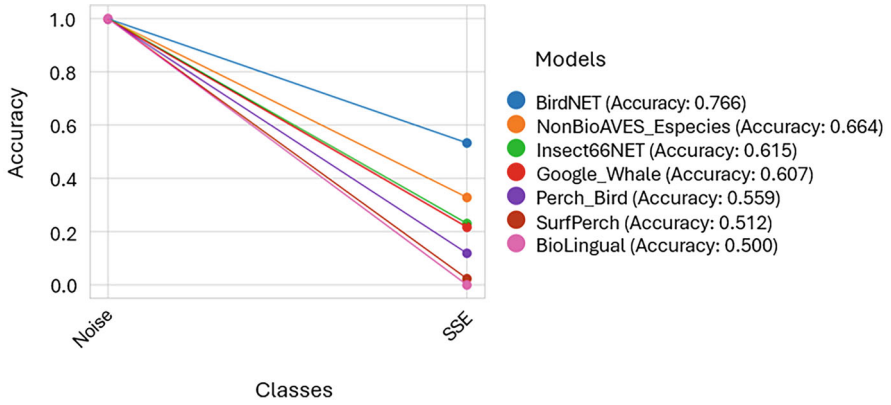


**Fig. 2** Per-class accuracy achieved by classifiers trained with embeddings from seven pretrained models

negatives (E (tp) = 51, (fp) = 272). The lowest false positive counts were achieved by classifiers C and F but were accompanied by higher false negative counts (C (tp) = 41, (fp) = 8; F (tp) = 33, (fp) = 4).

Classifier C achieved the highest area under the precision–recall curve (PR AUC = 0.63) and the highest F1 score (= 0.69) (Fig. 4, Table 3). This indicates that classifier C has a high true-positive detection rate while minimizing false positive detection. These values fall within the performance range of models developed by Diep et al. (2016) and Sota et al. (2023) (Table 4).

A subset of prediction outcomes by classifier C were visually and acoustically inspected. True positive spectrograms had clearly visible shad splashing events (Fig. 5a–c). False positive spectrograms, with a high confidence score of 1 by the

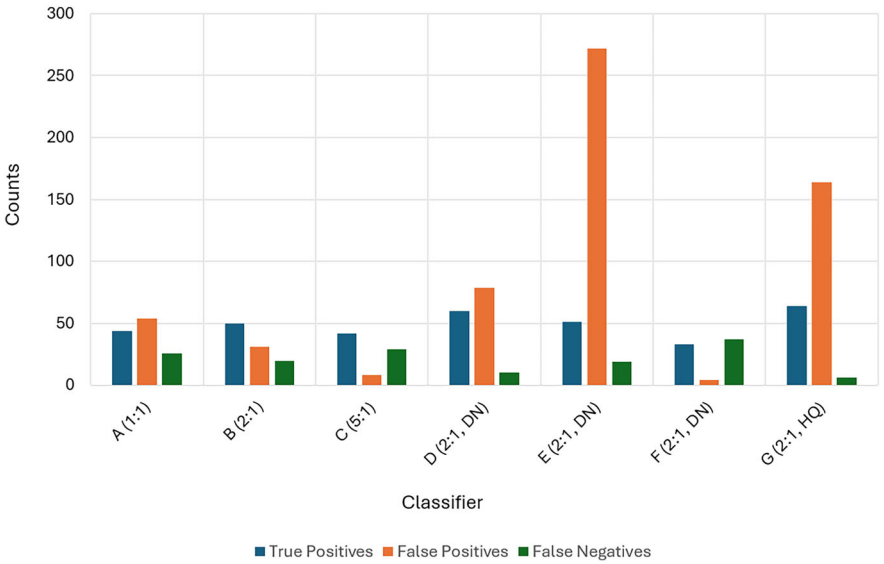


**Fig. 3** Total counts of true positive, false positive, and false negative predictions achieved by seven classifiers (A–G) trained on different training dataset variants and evaluated on the same test dataset. The information within the brackets indicates noise-to-shad splashing ratio, noise subsample, and shad splashing subsample of the training dataset. Abbreviations: different noise subsample (DN), high-quality shad splashing subsample (HQ)

classifier, displayed acoustic patterns of water sounds like gurgling, which resembles shad splashing (Fig. 5d). False positives associated with lower confidence values were visually distinct from shad splashing (Fig. 5e, f). The false negatives were generally noisy and included acoustic activity like bird calls or other ambient sounds (Fig. 5g–i). Shad splashing events were either masked by these sounds or distant.

### Preliminary Performance Assessment on Real-World Data

From an unedited 5-hour field recording, classifiers E and G generated the highest number of predictions, and classifiers C and F generated the fewest (Table 5). The exploratory visual and auditory inspection of selected predictions suggested that the relative performance trends observed on the test dataset were maintained. However, the number of false positive predictions was elevated.

## Discussion

Detection of shad splashing is challenging due to its acoustic similarity to environmental noise and its rarity in recordings. To address the limited availability of annotated data and the need for robust shad detection, this study evaluated whether

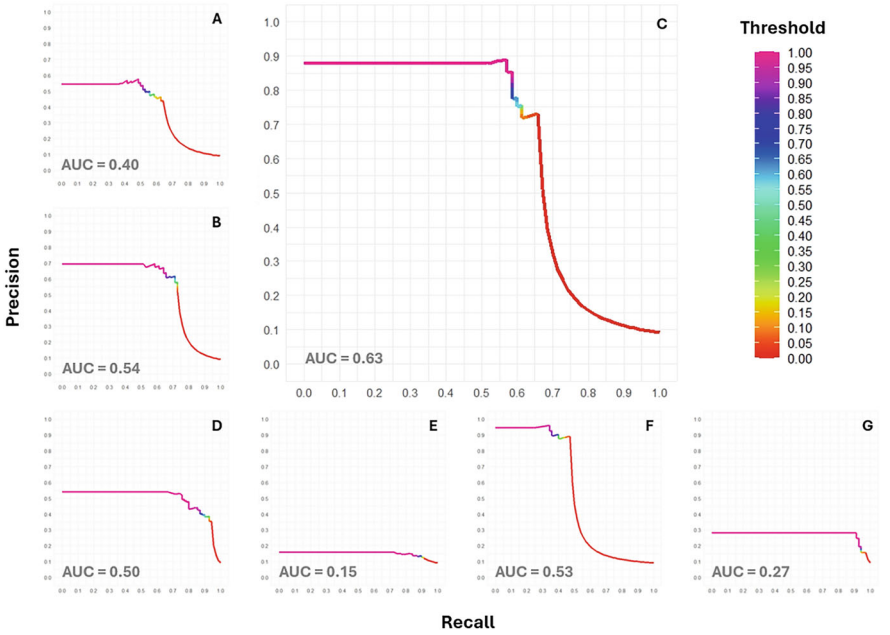


**Fig. 4** Precision–recall curves for seven classifiers (A–G) trained on different dataset variants across individual decision thresholds

**Table 3** Performance metrics for each classifier at individual decision threshold

| Model ID | Threshold | Recall | Precision | F1 score |
|---|---|---|---|---|
| A | 0.1 | 0.63 | 0.45 | 0.52 |
| B | 0.7 | 0.71 | 0.62 | 0.66 |
| C | 0.95 | 0.59 | 0.84 | 0.69 |
| D | 0.95 | 0.86 | 0.43 | 0.57 |
| E | 1 | 0.73 | 0.16 | 0.26 |
| F | 0.05 | 0.47 | 0.89 | 0.62 |
| G | 1 | 0.91 | 0.28 | 0.43 |

**Table 4** Performance metrics for different shad splash detectors

| Model ID | Threshold | Recall | Precision | F1 score |
|---|---|---|---|---|
| Diep et al. (2016) | 0.23 | 0.69 | 0.68 | 0.69 |
| Sota et al. (2023) | – | 0.64 | 0.44 | 0.52 |
| McLain and White (unpublished) | – | 0.97 | 0.97 | 0.97 |
| Classifier C | 0.95 | 0.59 | 0.84 | 0.69 |

classifiers trained using pretrained embedding models can match the performance of models trained from scratch. A linear classifier trained on BirdNET embeddings from in-air recordings achieved an F1 score of 0.69, performance comparable to

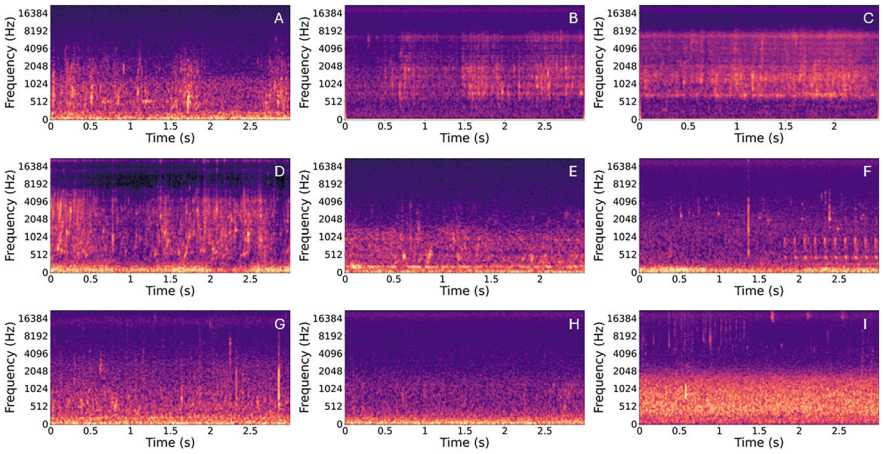


**Fig. 5** Mel spectrograms for different prediction outcomes of classifier C. (**a**–**c**) True positives; (**d**–**f**) false positives; (**g**–**i**) false negatives

**Table 5** Number of predicted shad splashing events in 5 hours of continuous recordings per classifier

| Classifier ID | # Predictions |
|---|---|
| A | 248 |
| B | 403 |
| C | 149 |
| D | 933 |
| E | 5283 |
| F | 53 |
| G | 7210 |

published shad splashing detectors (F1 = 0.69 Diep et al. 2016; F1 = 0.52 Sota et al. 2023), despite using a substantially smaller training dataset than Sota et al. (2023). Furthermore, an imbalanced training dataset enhanced performance. This outcome aligns with Ghani et al. (2023), who found that using embeddings extracted from models trained on bird vocalizations produced higher-quality classification results consistently compared to embeddings from models trained on general audio datasets. This may be due to the inherent diversity and complexity of bird vocalizations, which allows such models to generalize better to other bioacoustic tasks.

Training datasets with an increasing imbalance in the noise-to-shad splashing ratio, from ratios of 1:1 to 1:5, led to improved classifier performance. There was a consistent reduction in false positive predictions as the imbalance in the training data increased. This trend may be explained by (1) a stronger bias toward the majority class ("noise"), (2) a smaller prior shift between training and test datasets, or (3) a larger number of negative examples included during training. Classifiers trained on different noise subsamples exhibited different prediction patterns, ranging from conservative to confident, which may be due to the random undersampling of the majority class and risks discarding valuable examples. This observation suggests

that classifiers might not have been exposed to sufficient negative examples. Filtering the dataset based on shad splashing event quality further reduced the already limited sizes of both the negative and the positive class and removed ambiguous cases, potentially explaining the decline in performance.

The relative performance trends of the different classifiers were largely maintained during preliminarily evaluation on real-world recordings. However, the proportion of false positive predictions appeared to be elevated, possibly due to the relative infrequency of shad splashing events in the real-world data compared to the test dataset. To obtain more reliable performance estimates, nested cross-validation could be employed in future work, which may enable decision threshold tuning and evaluation on multiple folds. Alternatively, the decision threshold could be adjusted by using continuous real-world data and subsequently evaluating classification performance on a separate set of real-world recordings. Furthermore, performance may improve by increasing the size of both the positive class and the negative class. In conclusion, our study demonstrates that shad splashing events can be detected using machine learning through a classifier trained on top of BirdNET, which provides high potential for automatic twaite shad monitoring.

**Acknowledgments** The authors would like to extend their special thanks to the Agentschap Maritime Dienstverlening en Kust Shipping Assistance Division, and Aertssen for making the fieldwork for this study possible by granting access to vessels and field sites. Further thanks to Ioannis Sidiropoulos and Chloe Kenney for their contributions during the fieldwork. We also thank Tim Smyth and Franz Hölker for the helpful exchanges during this study.

Funded by the European Union under the Horizon Europe Programme, Grant Agreement No. 101135471 (AquaPLAN).

**Competing Interest Declaration** The author(s) has no competing interests to declare that are relevant to the content of this manuscript.

## References

Alcocer I, Lima H, Sugai LSM, Llusia D (2022) Acoustic indices as proxies for biodiversity: a meta-analysis. Biol Rev 97(6):2209–2236. https://doi.org/10.1111/brv.12890

Allen AN, Harvey M, Harrell L, Wood M, Szesciorka AR, McCollough JLK, Oleson EM (2024) Bryde's whales produce biotwang calls, which occur seasonally in long-term acoustic recordings from the central and western North Pacific. Front Mar Sci 11:1394695. https://doi.org/10.3389/fmars.2024.1394695

Aprahamian MW, Baglinière J-L, Sabatié MR, Alexandrino P, Thiel R, Aprahamian CD (2003) Biology, status, and conservation of the anadromous Atlantic Twaite Shad Alosa fallax fallax. Biodivers Status Conserv World's Shads 35:103–124

Bianco MJ, Gerstoft P, Traer J, Ozanich E, Roch MA, Gannot S, Deledalle C-A (2019) Machine learning in acoustics: theory and applications. J Acoust Soc Am 146(5):3590–3628. https://doi.org/10.1121/1.5133944

Diep D, Nonon H, Marc I, Lebel I, Roure F (2016) Automatic acoustic recognition of shad splashing using a smartphone. Aquat Living Resour 29(2):204. https://doi.org/10.1051/alr/2016015

Faiß M, Stowell D (2023) Adaptive representations of sound for automatic insect recognition. PLoS Comput Biol 19(10):e1011541. https://doi.org/10.1371/journal.pcbi.1011541

Ghani B, Denton T, Kahl S, Klinck H (2023) Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci Rep 13(1):22876. https://doi.org/10.1038/s41598-023-49989-z

Hagiwara M (2023) AVES: animal vocalization encoder based on self-supervision. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095642

Kahl S, Wood CM, Eibl M, Klinck H (2021) BirdNET: a deep learning solution for avian diversity monitoring. Eco Inform 61:101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Kather VS, Ghani B, Stowell D (2025) Clustering and novel class recognition: evaluating bioacoustic deep learning feature extractors. In: 11th Convention of the European Acoustics Association, Málaga, Spain, April 9. https://doi.org/10.48550/arXiv.2604.11560

Robinson D, Robinson A, Akrapongpisak L (2024) Transferable models for bioacoustics with human language supervision. In: ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1316–1320. https://doi.org/10.1109/ICASSP48485.2024.10447250

Ruff ZJ, Lesmeister DB, Appel CL, Sullivan CM (2021) Workflow and convolutional neural network for automated identification of animal sounds. Ecol Indic 124:107419. https://doi.org/10.1016/j.ecolind.2021.107419

Sota A, Guyot P, Alix F, Xhika K (2023) CRNN-based splash audio event detection for fish monitoring. In: Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023, pp 5159–5162. https://doi.org/10.61782/fa.2023.1008

te Velde K, Mairo A, Peeters ETHM, Winter HV, Tudorache C, Slabbekoorn H (2024) Natural soundscapes of lowland river habitats and the potential threat of urban noise pollution to migratory fish. Environ Pollut 359:124517. https://doi.org/10.1016/j.envpol.2024.124517

Williams B, van Merriënboer B, Dumoulin V, Hamer J, Fleishman AB, McKown M, Munger J, Rice AN, White C, Hobbs C, Razak T, Curnick D, Jones KE, Denton T (2025) Using tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. Philos Transact Royal Soc B 380(20240280). https://doi.org/10.1098/rstb.2024.0280