forum acusticum 2026
8-12 September · Graz, Austria

DEEP-LEARNING-BASED DETECTORS FOR THE CALLS OF THE SHALLOW WATER COMMON SPADEFOOT TOAD (*PELOBATES FUSCUS*)

Siri Hosøy[1,3], Anders Raagaard[2], Østen Holkestad[1,4], Guillaume Dutilleux[1]

[1] *Acoustics Group, Department of Electronic Systems, NTNU - Norwegian University of Science and Technology, 7491 Trondheim, Norway*
[2] *Department of Biology, Section for Freshwater Biology, University of Copenhagen, Copenhagen, Denmark*
[3] *Now with Framo Flatøy AS, Bergen, Norway*
[4] *Now with Nordic Semiconductor, Trondheim, Norway*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

In the context of the decline of the European common spadefoot toad (*Pelobates fuscus*), long term bioacoustic monitoring of this secretive species that vocalizes underwater is highly relevant. In this study, we present a software detector that features superior performance compared to literature [1]. For this study, hydrophone data was collected in Denmark and Poland with a different recorder than in [1]. Two candidate detectors were designed: a spectrogram-based EfficientNet-B0 classifier and a waveform-based CNN-BiGRU one. Both were trained on existing French hydrophone data and the new Danish Hydromoth data with call-aware losses and post-processing. The spectrogram-based model reached TPR 84.5% at FPR 0.11% with 94% accuracy. The raw-waveform model reached TPR 75% at FPR 0.02% with 93 % accuracy. The new detectors were used to turn long recordings into phenologies.

Keywords: *amphibian monitoring,* Pelobates Fuscus*, detection, deep learning, EfficientNet*

1. Introduction

The European Common spadefoot toad (*Pelobates fuscus*), is one of the most striking illustrations of amphibian decline in Western Europe. The species used to be common from Western Europe to Western Siberia [2]. It is now listed in Annex IV of the European Habitats Directive [3] and in Appendix II of the Bern Convention [4]. This status has prompted conservation actions, including breeding and reintroduction programmes in Italy [5], Germany [6] and Belgium, as well as restoration plans in France where the species is classified as Endangered in the national Red List (IUCN, 2015).

Several aspects of the biology of *P. fuscus* complicate field documentation, beyond the small size of many remaining populations. The species is fossorial and mainly nocturnal, spending much of its life underground and emerging only to breed [7]. Breeding phenology is difficult to predict, and activity in shallow ponds is often secretive [8], with adults remaining invisible in deeper water. Passive acoustic monitoring with hydrophones is thus a natural choice for presence and activity assessment [9].

The vocal repertoire of the species has been described as comprising four to six call types [10], [11], [12]. Among these vocalizations, the species-specific advertisement call is by far the most frequently produced. It is emitted by both males and females during the breeding season [7], [13] and has a relatively simple acoustic structure, making it a suitable target for bioacoustic monitoring [13].

A dedicated feature-based detector for *P. fuscus* advertisement calls is available [1]. Its design prioritised a low false-positive rate (FPR) and achieved an acceptable true-positive rate (TPR, 53-73%) for presence/absence screening, with an FPR near 1.5%. While the achieved FPR was low, the application of this detector to long term datasets led to a non-negligible number of false positives, mostly caused by rain but alsp sympatric species.

In the present study, two deep-learning-based detectors for *P. fuscus* are introduced to improve the precision of the detection of the species. They are trained on an existing dataset collected in France and on new data recordings made in Denmark. The remainder of this paper is organized as follows. Chapter 2 describes the target calls for the two species of interest,

***Corresponding author**: Guillaume Dutilleux.*

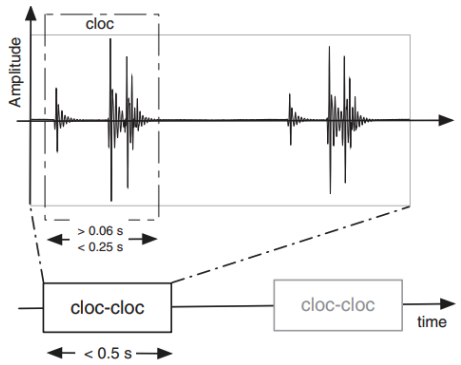


Figure 1: Typical advertisement call of *P. fuscus* and relevant time parameters for the automated detector. Adapted from [1].

the production of a dataset and the architecture and training of the machine learning models. Chapter 3 presents the performance of the detectors and their application to the phenology of the species. These results are discussed in chapter 4. Chapter 5 concludes this paper.

2. Materials and methods

2.1. Advertisement call characteristics and propagation

The target signal was the advertisement call of *P. fuscus*, which consists of a stereotyped sequence of evenly spaced short bursts (Fig. 1). Each burst typically contains two or three similar "cloc" notes. A single note begins with a low-amplitude transient followed by two higher-intensity transients, forming a distinctive temporal pattern. Most acoustic energy produced by male *P. fuscus* is concentrated between 700 and 1200 Hz, with some energy extending up to 2.5 kHz [12], [13]. The calls are emitted in shallow water, which acts as a natural high-pass filter. At a depth of approximately 50 cm, the cut-off frequency is close to 1 kHz, indicating that the received spectrum is strongly influenced by the relative position between the caller and the hydrophone as well as the local water depth.

2.2. Data sources

2.2.1. Existing data

Part of the dataset originally described in [1] was reused. The material consists of long-term hydrophone recordings collected at two breeding sites in north-eastern France, Mothern and Sauer, during the spring seasons of 2015 and 2016. The recording equipment used was SM2 SongMeter programmable recorders (Wildlife Acoustics, Maynard, USA) coupled to HTI-96 hydrophones (High Tech Inc., Long Beach, USA). Audio was sampled at 16 kHz with 16-bit resolution and stored as uncompressed PCM.

2.2.2. Data collected during this study

A new data set was collected during the breeding season in Denmark in May 2024. During this campaign, the recordings were made with Hydromoths (Open Acoustic Devices, Oxford, UK). The transducer was a microphone enclosed in the device's casing. Audio was sampled at 32 kHz with 16-bit resolution and stored as uncompressed PCM.

2.3. Detector design

Two detectors were designed. One operated on spectrogram representations and the other on raw audio waveforms.

2.3.1. Spectrogram-based model (EfficientNet-B0)

The first detector operated on spectrograms. Each waveform was band-limited to the frequency range of interest and transformed into a log-scaled spectrogram, providing a compact representation of both temporal and spectral call characteristics. The detector is based on an EfficientNet-B0 backbone pretrained on ImageNet [14]. This architecture was selected because it is a good trade-off between computational efficiency and representational capacity, making it suitable for large-scale inference on long-term recordings. Spectrograms were resized to 224×224 pixels and replicated across three channels to match the input format expected by the model while preserving the original time–frequency structure [15].

A lightweight feed-forward head was appended to adapt the feature extractor to call detection. The head consisted of two fully connected layers with 512 and 128 units respectively, both using ReLU activation, followed by dropout rates of 0.15 and 0.30 to reduce overfitting [16]. A single sigmoid output neuron produced the probability of an advertisement call being present in each one-second segment.

2.3.2. Waveform-based model

The second detector operated directly on the raw audio waveform without any intermediate time–frequency transformation. This approach seemed quite relevant considering the transient nature of the signal to be detected whose localization along the frequency axis is not very accurate. A one-dimensional convolutional front end learned feature representations directly from the waveform, enabling the model to capture subtle temporal variations in amplitude and phase that may be lost in spectrogram-based approaches. Residual and dilated convolutional blocks were used to extract local temporal patterns at multiple receptive field scales, allowing the network to detect both short transients and slower oscillatory structures within each one-second segment. The convolutional stack was followed by bidirectional gated recurrent unit (BiGRU) layers that aggregated temporal dependencies across the window in both forward and backward directions [17]. This architecture emphasized temporal continuity and microstructure while avoiding the constraints of a fixed spectrogram transform, offering a complementary representation to the spectrogram-based detector [16].

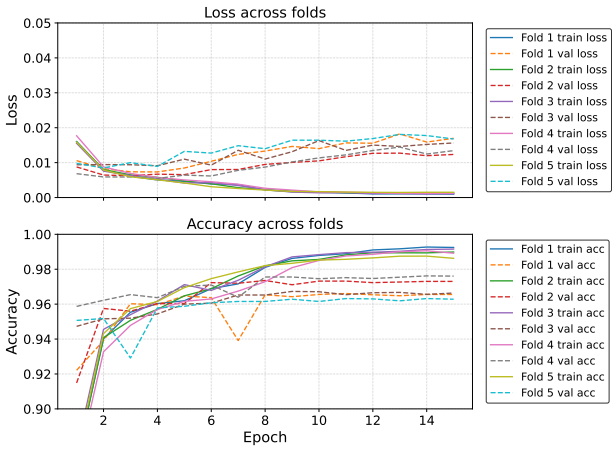


Figure 2: Training and validation curves for the spectrogram-based model. Fold 5 features the best performance.

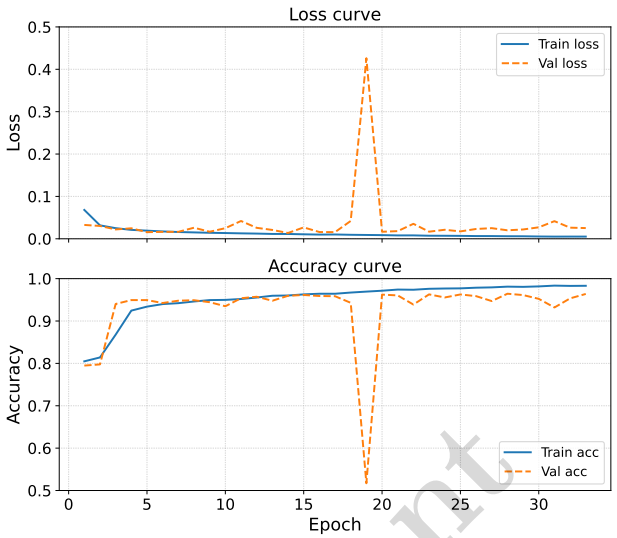


Figure 3: Training and validation curves for the raw audio model.

2.3.3. Training procedure

Both models were trained using manually-annotated French and Danish data and evaluated on independent test data before being applied to the full recordings. The Adam optimizer was used for training. The spectrogram-based detector was tuned through k-fold cross-validation on the combined training and validation data, while the raw-waveform detector used a fixed split and was selected based on validation results. Model evaluation used accuracy, true positive rate (TPR), false positive rate (FPR), precision, and the F_β score with $\beta = 0.5$, which places greater emphasis on precision.

3. Results

3.1. Spectrogram-based model

The spectrogram-based detector was trained using 5-fold cross-validation. Training accuracy approached 99%, with validation accuracy stabilizing between 96–98% across folds. Loss values converged towards zero for training and approximately 0.04% for validation. Fold 5 achieved the best balance between precision and recall while minimizing false positives and was selected as the final detector. Figure 2 shows the corresponding training curves.

Table 1 summarizes detection performance using both a single threshold ($p > 0.95$) and a dual-threshold approach, where weaker predictions ($0.60 < p < 0.95$) were re-evaluated. The dual-threshold strategy significantly improved recall while maintaining a low FPR.

Table 1: Detection performance on the evaluation set using single vs. dual threshold.

Approach	Acc.	TPR	FPR
Single ($p > 0.95$)	0.8836	0.6950	0.0004
Dual threshold	0.9403	0.8451	0.0011

3.2. Waveform-based model

The waveform-based model was trained for 30 epochs. Training accuracy increased towards 99%, with validation accuracy stabilizing near 96%. A jump in validation loss occurred between epochs 18–20, likely due to a small validation set and lack of augmentation. The final model was used for evaluation and the training curves are shown in Figure 3. The predicted probability distribution follows the same structure as for the spectrogram based model.

The model was evaluated on 2061 positive and 5381 negative windows. Most true positives clustered near $p > 0.8$, while true negatives near zero. Using a dual-threshold configuration with $p > 0.75$ for strong predictions, and $0.4 < p < 0.75$ for weak predictions, the model achieved 93% accuracy, 75% TPR, and a FPR of 0.02%, as shown in Table 2.

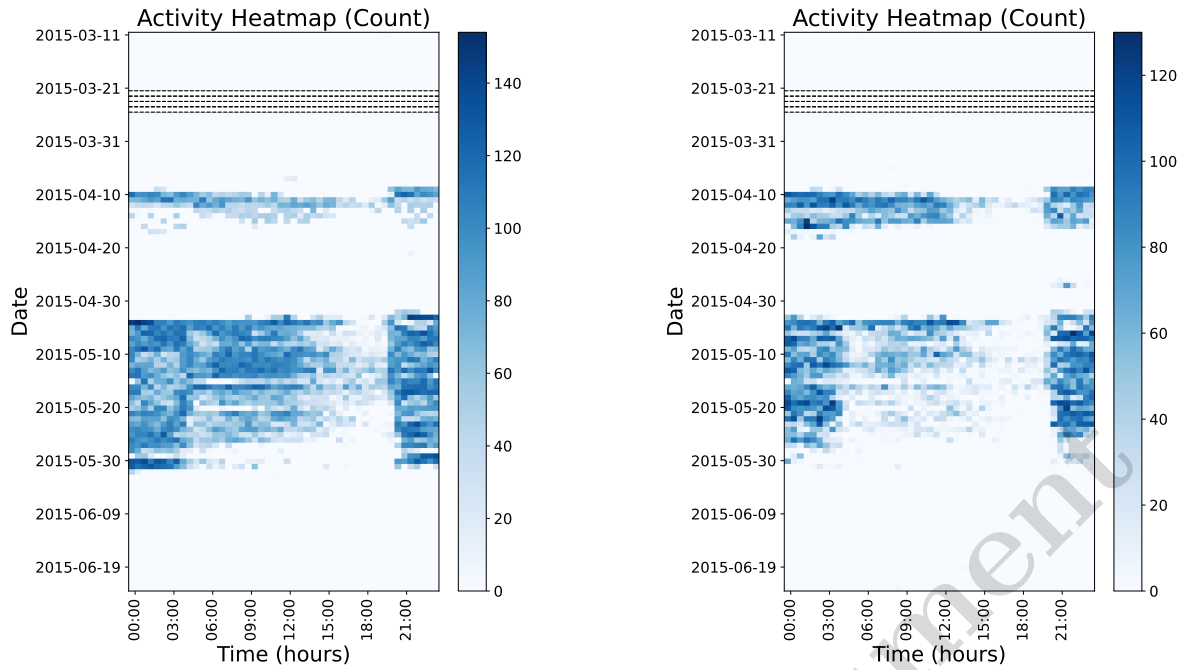
Table 2: Evaluation metrics for the raw audio model

Approach	Acc.	TPR	FPR
Single ($p > 0.75$)	0.9249	0.7306	0.0002
Dual threshold	0.9312	0.7533	0.0002

3.3. Data preparation and annotation

Audio recordings were divided into fixed 1 s analysis windows using a sliding window with a 0.5 s hop, applied consistently across the training, validation, and test partitions. Manual annotations indicated the presence of advertisement calls. A window was labeled as positive if it overlapped an annotated call by at least 50%.

The dataset was divided 80/10/10 into training, validation, and test sets at the file level to prevent temporal leakage between partitions. For the spectrogram-based detector, an additional k-fold cross-validation was performed on the combined training and validation data to support model selection before evaluating



(a) Spectrogram-based detector

(b) Raw-audio detector

Figure 4: Daily calling activity heatmaps from Mothern 2015 showing a 30-minute binning and colour scale. The spectrogram-based detector is shown on the left and the raw-audio detector on the right for direct comparison.

performance on the held-out test set. The distribution of positive and negative chunks in each fold is summarized in table 3.

For the raw audio model, the training data contained 8010 positive and 34 698 negative examples, giving a total of 42 708 windows [16]. The validation data contained 2020 positive and 7836 negative examples, giving a total of 9856 windows. These values reflect the natural class imbalance present in the recordings, with non-call segments occurring far more frequently than advertisement calls.

Table 3: Number of positive and negative chunks used for training and validation in each fold.

	Training set		Validation set	
	Pos.	**Neg.**	**Pos.**	**Neg.**
F1	6684	17 859	1638	3818
F2	6670	17 585	1652	4092
F3	6576	18 074	1746	3603
F3	6588	15 294	1634	6383
F5	6670	17 896	1652	3781

3.4. Application to large scale recordings

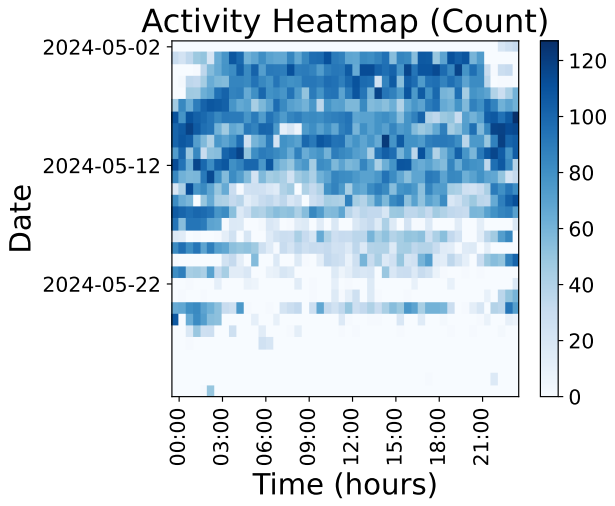
The two detectors proposed deliver time-stamped call detections as label files. This data can be turned into call rates. The call rate is likely correlated with the number of callers within reach of the acoustic sensor. Figure 4 shows detected vocal activity for Mothern during the 2015 breeding season, to be compared with the validated detection pattern presented in [1]. Figure 5 presents vocal activity detected by the two detectors for the recordings made in Denmark in May 2024.

4. Discussion

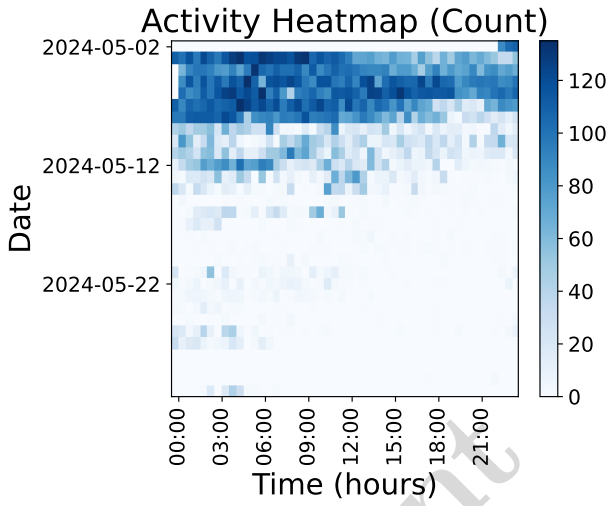
The raw model was less sensitive to faint calls but produced fewer false positives and better-aligned detection intervals. Fig. 6 and 7 illustrate the inherent limitation of detection based on the raw waveform, even in environment where the amount of background noise is fairly low.

Despite their good overall performance, both detectors exhibited characteristic error patterns. Most false positives occurred during episodes of heavy rain, wind, or mechanical noise, which produced broadband transients resembling the temporal envelope of *P. fuscus* calls. These artefacts often appeared as vertical striations in spectrograms and triggered the spectrogram model more readily than the raw audio model. False negatives were mainly caused by faint calls close to the noise floor or by vocalisations straddling chunk boundaries. Some calls were also missed when multiple individuals called in quick succession, creating overlapping bursts that confused the model's internal timing assumptions.

Beyond model-level factors, environmental conditions impose natural limits on detectability. The underwater propagation of *P. fuscus* vocalisations is strongly influenced by water depth and substrate, leading to attenuation and distortion at greater distances from the hydrophone. Consequently, even an ideal detector cannot guarantee 100% recall in real field conditions. Recognising these constraints is essential for interpret-



(a) Spectrogram-based detector



(b) Waveform-based detector

Figure 5: Daily calling activity heatmaps from Denmark showing 30-minute binning and colour scale. The spectrogram-based detector is shown on the left and the raw-audio detector on the right for direct comparison.

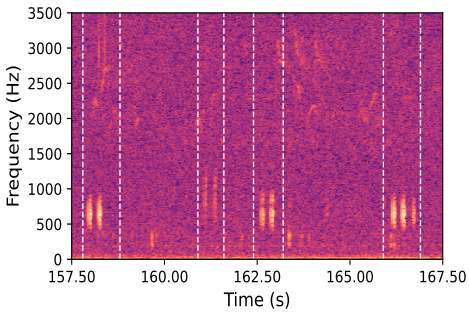


Figure 6: Spectrogram snippet showing four calls.

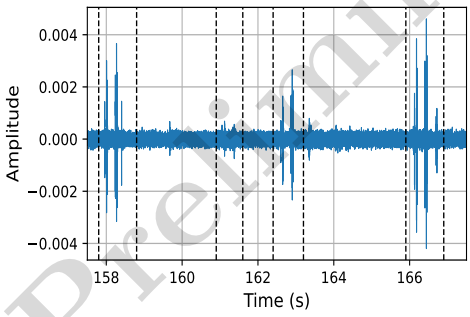


Figure 7: Raw waveform from the same snippet as in Fig. 6, where only three among the four calls are easy to spot.

ing absence data and for designing future monitoring campaigns with optimal hydrophone placement and sampling intervals.

The calling activity patterns shown in Figure 4 are quite similar to the validated detections presented in [1]. Further work should include the direct validation of the detections from the machine-learning-based detector for a more accurate picture. Figure 5 suggests that the breeding season was already close to its peak in early May 2025 on the Danish site. Like for the French recordings, vocal activity varies significantly from one day to the next during the breeding season.

5. Conclusion

In this paper, we presented two deep-learning-based detectors for the rapidly declining *P. fuscus*. The first one builds on EfficientNet-B0 with a custom head. It takes spectrograms as inputs. The second one is based on CNN-BiGRU and processes the audio signal directly. Both outperform the existing feature-based detector both when it comes to true positive rate and false positive rate. The spectrogram-based detector achieves a higher true positive rate likely because the time-frequency representation is better at highlighting calls when the signal-to-noise ratio is low. The raw-waveform model features the lowest false positive rate. It is indeed better at discriminating between the transient-like call of the species and other transients like rain drops and shocks.

Further work should evaluate if machine learning makes it possible to provide information about the abundance of callers in the detection range of the acoustic sensor, instead of merely presence data.

References

[1] G. Dutilleux and C. Curé, "Automated acoustic monitoring of endangered common spadefoot toad populations reveals patterns of vocal activity," *Freshwater Biology*, vol. 65, no. 1, pp. 20–35, Jan. 2020.

[2] C. Eggert, D. Cogălniceanu, M. Veith, G. Dzukic, and P. Taberlet, "The declining Spadefoot toad, *Pelobates fuscus* (Pelobatidae): Paleo and recent

environmental changes as a major influence on current population structure and status," eng, *Conservation genetics*, vol. 7, no. 2, pp. 185–195, 2006, ISSN: 1566-0621.

[3] EU, "Directive 92/43/eec on the conservation of natural habitats and of wild fauna and flora," European Commission, Brussels, BE, Council Directive 92/43/EEC, May 1992.

[4] Council of Europe, "Convention on the conservation of european wildlife and natural habitats - Appendix II strictly protected fauna species," Council of Europe, Bern, CH, European treaty 104, Sep. 1979.

[5] A. Giovannini, D. Seglie, and C. Giacoma, "Identifying priority areas for conservation of spadefoot toad, *Pelobates fuscus insubricus* using a maximum entropy approach," *Biodiversity and conservation*, vol. 23, no. 6, pp. 1427–1439, 2014, ISSN: 0960-3115.

[6] L. ten Hagen et al., "Vocalizations in juvenile anurans: Common spadefoot toads *(Pelobates fuscus)* regularly emit calls before sexual maturity," *Die Naturwissenschaften*, vol. 103, no. 9-10, pp. 75–8, 2016.

[7] A. Nöllert, *Die Knoblauchkröte*. Neue Brehm-Bücherei, Ziemsen-Verlag, 1990.

[8] C. Eggert and R. Guyétant, "Reproductive behaviour of spadefoot toads (Pelobates fuscus): Daily sex ratios and males' tactics, ages, and physical condition," eng, *Canadian journal of zoology*, vol. 81, no. 1, pp. 46–51, 2003, ISSN: 0008-4301.

[9] B. A. DeGregorio, P. J. Wolff, and A. N. Rice, "Evaluating hydrophones for detecting underwater-calling frogs," *Herpetological Conservation and Biology*, vol. 16, no. 3, pp. 513–524, Dec. 2021.

[10] B. Müller, "Bio-akustische und endokrinologische Untersuchungen and der knoblauchkröte *Pelobates fuscus fuscus*," *Salamandra*, vol. 20, no. 2-3, pp. 121–142, 1984.

[11] F. Andreone and R. Piazza, "A bioacoustic study of *Pelobates fuscus insubricus* (Amphibia, Pelobatidae)," *Bolletino di zoologia*, vol. 57, no. 4, pp. 341–349, 1990.

[12] F. Stanescu, L. R. Forti, D. Cogalniceanu, and R. Marquez, "Relase and distress calls in european spadefoot toads, genus pelobates," *Bioacoustics*, vol. 28, no. 3, pp. 224–238, Feb. 2019. DOI: 10.1080/09524622.2018.1428116.

[13] K. H. Frommolt, M. Kaufmann, S. Mante, and M. Zadow, "Die Lautäußerungen der Knoblauchkröte (*Pelobates fuscus*) und Möglichkeiten einer akustischen Bestandserfassung der Art," *Rana*, no. Sonderheft 5, pp. 101–112, 2008.

[14] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.11946.

[15] Ø. F. Holkestad, "A deep-learning-based approach to detect the common spadefoot toad (Master's thesis)," NTNU, Trondheim, Norway, Master's thesis, Jun. 2021.

[16] S. Hosøy, "Machine learning for acoustic monitoring of Pelobates fuscus (Master's thesis)," NTNU, Trondheim, Norway, Master's thesis, Jul. 2025. [Online]. Available: https://hdl.handle.net/11250/3222423.

[17] A. Falch, "Raw audio end-to-end deep learning architectures for sound event detection (Master's thesis)," UiO, Oslo, Norway, Master's thesis, May 2023.