Evaluation of Deep Learning-Based Monitoring of Frog Reproductive Phenology

Kimura, Kaede; Sota, Teiji

Evaluation of Deep Learning-Based Monitoring of Frog Reproductive Phenology

Kaede Kimura[1] and Teiji Sota[1]

[1] Department of Zoology, Graduate School of Science, Kyoto University, Sakyo, Kyoto 606-8502, Japan.
(KK) Email: kaede.kmr@gmail.com. Send reprint requests to this address.
(TS) Email: sota.teiji.88u@st.kyoto-u.ac.jp

To evaluate the utility of a deep-learning approach for monitoring amphibian reproduction, we examined the classification accuracy of a trained model and tested correlations between calling intensity and frog abundance. Field recording and count surveys were conducted at two sites in Kyoto City, Japan. A convolutional neural network (CNN) model was trained to classify the calls of five anuran species. The model achieved 91–100% precision and 75–98% recall per species, with relatively lower performance on less abundant species. Computational experiments investigating the effects of the number and seasonality of the training samples showed that models trained on larger datasets from broader recording seasons performed better. Calling activity was high when males were abundant (Pearson's $r = 0.45–0.66$), although correlations between the calling activity and the number of pairs in amplexus were generally weaker. Our results suggest that deep learning is an effective tool for reconstructing the reproductive phenology of male anurans from field recordings. However, caution is required when applying to rare species and when inferring female reproductive activity.

Monitoring reproductive activity is vital for studying amphibian populations. Passive Acoustic Monitoring (PAM) is a widely used approach to study anuran reproduction since breeding activity is more easily detectable by hearing than by seeing (Dorcas et al., 2009). Through PAM methods, researchers can investigate reproductive phenology (Liu et al., 2022), occurrence (Rowley et al., 2019), and population trends (Weir et al., 2009) of multiple anuran species simultaneously. Recent advancements in machine learning algorithms and low-power devices that can be deployed for extended periods have sparked renewed interest in acoustic monitoring (Gibb et al., 2019). Conventional survey methods relying on direct observation are labor-intensive, limiting both the extent and resolution of studies. Integrating autonomous sensing technologies (such as cameras, recorders, and DNA sequencers) with machine learning to accelerate data acquisition and species identification steps can enable researchers to conduct high-throughput field surveys (Keitt and Abelson, 2021; Besson et al., 2022).

Deep learning, a subset of machine learning algorithms, is particularly promising for automated monitoring and rapidly gaining popularity among ecologists (Borowiec et al., 2022). Among many model architectures utilized in deep learning, Convolutional Neural Networks (CNN) are most widely used for animal classification (Stowell, 2022; Borowiec et al., 2022). CNN models have been successful in identifying species or higher taxa from images (Norouzzadeh et al., 2018; Schneider et al., 2022) or sounds (LeBien et al., 2020; Kahl et al., 2021). Using deep learning in species identification has several advantages over previous machine learning methods. First, deep learning has demonstrated superior classification accuracy compared to other methods (Knight et al., 2017; Mac Aodha et al., 2018). Second, the same procedure can be easily applied to various taxa. The previous approaches usually required study-specific feature engineering, a process in which discriminating features (e.g., call duration, minimum/maximum frequency) are manually designed to detect focal species from other sound sources (Gibb et al., 2019). Instead, the same general-purpose procedure can be applied to different taxa in deep learning because a model automatically learns which features to extract (LeCun et al., 2015). For example, LeBien et al. (2020) presented a model classifying eleven frog and thirteen bird species. User-friendly packages such as fastai (Howard and Gugger, 2020) and Keras (https://keras.io) make this new technique increasingly accessible to biologists.

Despite these advantages, applications of deep learning in bioacoustics are in their early stages and taxonomically biased toward birds and marine mammals (Stowell, 2022). A few studies have demonstrated the ability of CNN models to identify anurans even in situations where multiple species can call simultaneously (Xie et al., 2017; LeBien et al., 2020). However, the utility of this approach in studying reproductive phenology remains unexplored. Much uncertainty still exists about the relationship between indices of calling activity and relative abundance of breeding individuals. For example, indices of calling activity may saturate when males are abundant. Furthermore, if the seasonal activity patterns of males and females are not fully synchronized, call intensity may not reflect the amount of egg deposition, which is often a more relevant aspect of reproduction for population management. It is therefore important to examine how much information about male/female abundance can be extracted from audio recordings.

Here we assessed the effectiveness of a deep-learning approach for monitoring anuran reproductive phenology. Specifically, our objectives were: (1) to evaluate the classification performance of the trained model, (2) to examine how sample size and seasonality in the training datasets affect model performances, and (3) to test relationships between calling activity and the number of individuals observed throughout the breeding seasons. To address these questions, we trained a CNN model to discriminate five anuran species in Japan and compared the estimated calling intensity and the number of observed frogs. The results of our study provided useful information about the effective application and interpretation of deep learning-based phenological monitoring from anuran calls.

MATERIALS AND METHODS

Study sites and data collection methods

Study sites.—Audio recording and frog count survey were conducted at two sites in Kyoto City, Japan. The first site was Mt. Uryu (35.04ºN, 135.80º E, alt. 283 m), where *Bufo japonicus japonicus* and *Zhangixalus schlegelii* occurred at a small pond (7 × 3 m) in a mixed evergreen and deciduous broad-leaved forest. The second site was the Kyoto farmstead of Experimental Farm of Graduate School of Agriculture, Kyoto University (35.03ºN, 135.78ºE, alt. 60 m). This experimental farm had a section of rice fields, and three species of anurans, *Dryophytes japonicus*, *Glandirana rugosa*, and *Pelophylax nigromaculatus* bred there.

Field recording and sight count.—We deployed audio recorders from March to April, 2022, at Mt. Uryu and from May to September, 2021 and 2022, at the experimental farm. Environmental sounds were recorded for 50 seconds every hour at a sampling rate of 32 kHz or 48 kHz, using either AudioMoth (Open Acoustic Devices) (Hill et al., 2018), Qriom YVR-R600 (Yamazen), or Song Meter Micro (Wildlife Acoustics). The recorder was enclosed in a sealed bag or a designated waterproof case and placed near the breeding water body.

In 2022, the number of individuals at breeding sites was counted to test the relationships between frog abundance and the intensity of calling activity. Field surveys were conducted almost every day during the breeding season of the target species except for *Z. schlegelii*, which spawns underground (Fukuyama, 1991) and was difficult to visually observe. At Mt. Uryu, an observer (K. K.) searched the breeding pond for *B. japonicus* with a dip net for 10 minutes during the day. It was possible to search almost the whole pond in 10 minutes because of its small size. At the experimental farm, the observer walked slowly along a designated route (215 m long) at night and counted the frogs. Sex was determined by external morphology such as vocal sac, nuptial pads, and coloration.

Classification with Convolutional Neural Network (CNN)

Data processing.—The recorded audio files were first downsampled to 22.05 kHz and cut into 5-second audio segments. These audio segments were converted to mel-scaled spectrograms with the librosa package (v. 0.9.1: McFee et al., 2022) in Python. We cut off the lowest 15 frequency bands (corresponding to 392 Hz) to reduce background noise. The fast Fourier transform window was the Hann window, and the window length was set to 512 with 75% overlap. The spectrograms consisted of grayscale images in portable network graphics (PNG) format with a resolution of 861 × 112 pixels.

Model training.—One of the authors (K. K.) manually annotated a subset of the spectrograms to label all species presented in each 5-second audio segment. Spectrograms lacking our target species were labeled as background. The labeled samples were split into two datasets: one for training a model (2565 samples) and the other for validating the model performance (190 samples; Table 1).

We used the ResNet18 model (He et al., 2015: arXiv 1512.03385) with transfer learning to classify the calls of the five anuran species. Different CNN models (ResNet34 and VGG16) were also tested but not used here because they required longer training time due to the larger number of parameters and nevertheless showed no better performance. In the transfer learning framework, models pre-trained on a large dataset (ImegeNet in our case) are re-trained to adapt to the specific work at hand. This procedure takes advantage of the pre-trained model's ability to extract fundamental features and allows us to efficiently train the model with a relatively small dataset (Christin et al., 2019). We employed the mixup data augmentation method (Zhang et al., 2018: arXiv 1710.09412) to increase the robustness to overlapping calls (Kahl et al., 2021). The head layers of the model were initially trained for ten epochs, and then the entire model was unfrozen and trained for another 150 epochs with the default parameters of the fastai (v. 2.7.10) package (Adam optimizer with a base learning rate of 0.002 and discriminative learning rate; https://www.fast.ai).

Model testing.—The validation samples were selected from the annotated spectrograms to include 40 samples per target class (Table 1). Classification performance was evaluated based on the following metrics.

$$\text{Precision} = \frac{TP}{TP + FP},$$

and,

$$\text{Recall} = \frac{TP}{TP+FN}$$

where *TP*, *FP*, and *FN* indicate true positives, true negatives, false positives, and false negatives, respectively. Precision measures how accurate the model predictions as positive are, and recall measures the ability of the model to find all the positive samples. We removed the background class when calculating these metrics focusing on model performance on the target species.

Effects of size and seasonal extent of the dataset.—Preparing a training dataset is a time-consuming process and how to optimize it is a practical concern. The number of training samples would affect the

Table 1. Numbers of training and validation samples. Plus signs indicate calls from more than one species are present in a sample. Validation dataset was selected to contain 40 samples for each species. Bjap = *Bufo japonicus*, Zsch = *Zhangixalus schlegelii*, Djap = *Dryophytes japonicus*, Grug = *Glandirana rugosa*, Pnig = *Pelophylax nigromaculatus*.

Label	Training	Validation
Bjap	93	35
Zsch	104	35
Bjap + Zsch	2	5
Djap	143	10
Grug	117	15
Pnig	50	15
Djap + Grug	148	10
Djap + Pnig	99	10
Grug + Pnig	1	5
Djap + Grug + Pnig	25	10
background	1783	40
SUM	2565	190

overall model performance. In addition, recordings from the entire breeding periods include quite heterogeneous soundscapes due to seasonal changes in species composition. Models trained on datasets from narrow seasonal window may poorly perform when applied to long-term recordings. We therefore tested how different training datasets affect model performance by changing dataset size and sampling period. The training dataset was classified into three categories of sampling period: "early one-third" (March 1 – 18 for Mt. Uryu, and May 8 – June 25 for the experimental farm), "early two-thirds" (March 1 – April 3 for Mt. Uryu, and May 8 – August 13 for the experimental farm), and "all" the recording periods. We selected this time periods because vocalization of most species was concentrated on early half of the recording periods in our study sites and the last third of recording period did not contain every species. The number of samples used for training varied between 100 and 1000 for the early one-third category, 100 and 1914 (all training samples within this period) for the early two-thirds category, and 100 to 2000 for the all category. The training samples were randomly selected from the specified period, with which the ResNet18 model was trained. The model performance was evaluated using the macro average precision and recall for the validation dataset (Table 1), and the process was repeated five times for each category and sample size.

Statistical analysis: correlation between calling activity and the number of observed individuals

We performed regression analyses to test the correlation between calling activity and the number of individuals observed during the frog count survey. We conducted regression analysis for both the number of males and amplexing pairs to assess whether acoustics can be used to monitor male vocalization as well as actual breeding events. To quantify the calling activity, the trained ResNet18 model was used to count the number of audio segments containing calls of the target species (the number of call segments; Matsushima et al., 2022). The model was applied to all the spectrograms generated from the field recordings, and then the presence or absence of the calls of each target species was recorded. Since 10 spectrograms were generated from each 50-second recording, the resulting number of call segments per recording and species ranged from 0 to 10. Although the anuran calling index is more popular (Dorcas et al., 2009), its calculation requires subjective judgments of the crowdedness of vocalization. Thus, counting the number of call segments was a more feasible method in our machine learning approach.

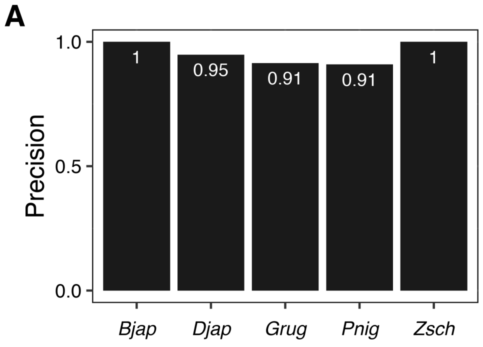
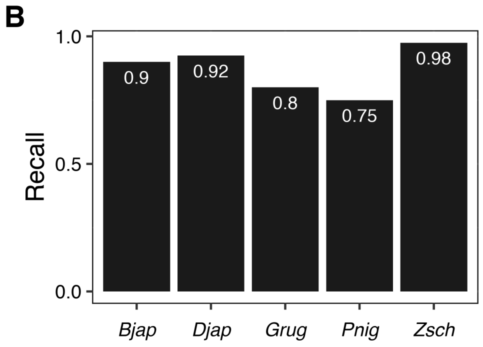

Fig. 1. Performance of the trained ResNet18 model. The values of (A) precision and (B) recall are shown for all species.

The daily average of the number of call segments was assigned as the response variable and the number of observed males or the number of amplexing pairs as the explanatory variable. We used the daily average as an index of call intensity because, in general, it was more strongly correlated with the number of observed males than the number of call segments at a single time point close to the survey time (Supplemental material Fig. S1). To account for the temporal autocorrelation of residuals, we employed generalized least square (GLS) method (Zuur et al., 2009), in which we specified the residual correlation structure using an ARMA(p, q) model, where p and q = 0, 1, or 2. Statistical model fitting was performed using the nlme package (Pinheiro et al., 2022) in the R environment (v. 4.2.2: R Core Team, 2022), and the best model was selected based on the Akaike Information Criterion (AIC).

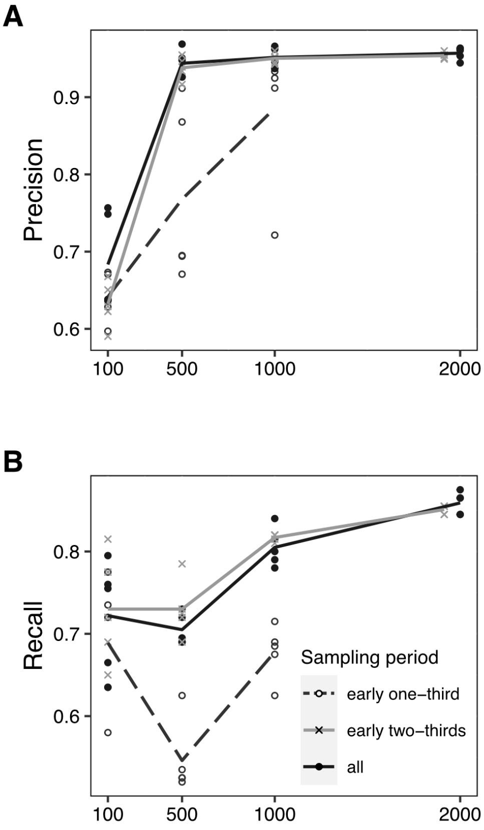

Fig. 2. Effects of dataset size and seasonal extent on model (A) precision and (B) recall. Training samples are selected from either early one-third, early two-thirds, or all of the recording period. Lines connect average values.

To quantify the strength of the relationships, we calculated the correlation coefficients (Pearson's r) between the daily average number of call segments and the number of males or the number of amplexing pairs.

RESULTS

Data collection.—Environmental sounds were successfully recorded for most of the study period, but half of the recordings from May 9–10 and May 15–17 were missing due to device problems. We conducted a frog sight count 21 times at Mt. Uryu and 110 times at

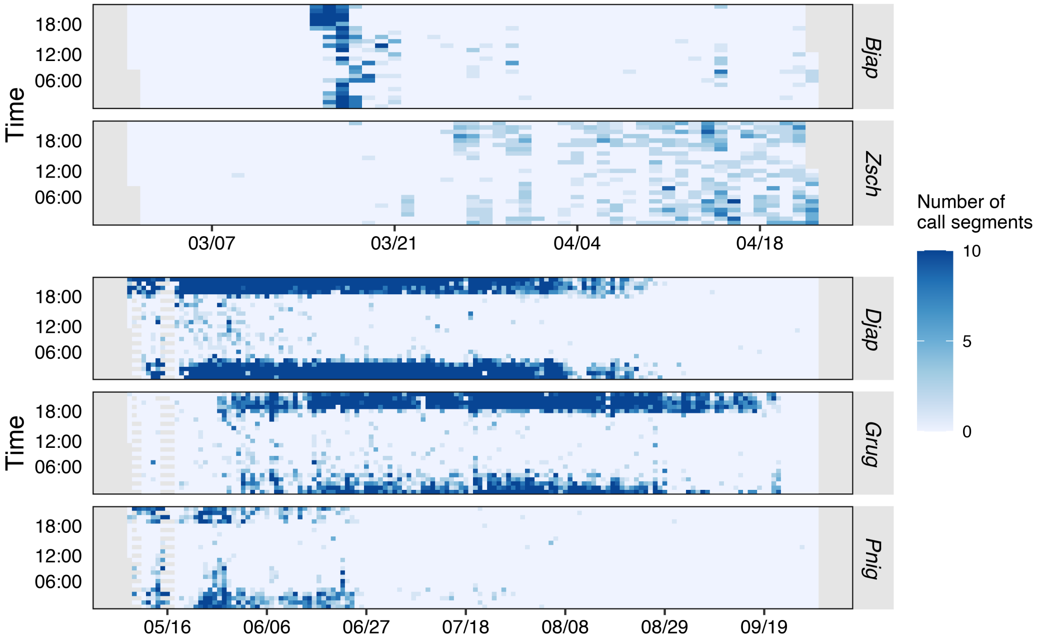

Fig. 3. Calling activity of the five anuran species in Kyoto City, Japan, inferred by the trained model. Darker blue represents a greater number of audio segments in which calls are detected, and gray color represents missing recordings.

the experimental farm of Kyoto University. A small number of *B. japonicus* were found at Mt. Uryu (3 males and 2 pairs at most), while *Z. schlegelii* was not observed although their calls were heard. At the experimental farm, *D. japonicus* was the most abundant species (on average 21.3 males per day), followed by *G. rugosa* (9.5 males) and *P. nigromaculatus* (1.2 males).

Model performance.—The ResNet18 model trained on the full set of training data (*n* = 2565; Table 1) classified five anuran species with 91–100% precision and 75–98% recall rates for the 190 validation samples (Fig. 1). Although the precision was sufficiently high for *B. japonicus*, *Z. schlegelii*, and *D. japonicus* (> 95%), it was relatively low for *G. rugosa* and *P. nigromaculatus* (91%). Similarly, the recall rates were lower for the latter two species (75–80%) than the former three (>90%) (Fig. 1).

Effects of size and seasonal extent of the dataset.—When the ResNet18 model was trained using various subsets of training data (*n* = 100–2000), model performance generally increased as the number of training samples increased (Fig. 2). However, models trained on the early one-third category samples exhibited lower performance than those trained on samples from a broader recording period (i.e. early two-thirds or all). Even with 1000 early one-third samples, the models did not, on average, perform better than the models trained on 500 early two-thirds or all samples. When trained on sub-samples from early two-thirds or all the recording period, precision quickly approached the asymptotic value, although larger sample sizes (> 1000) were required to achieve high recall values (Fig. 2B).

Correlation between calling activity and the number of observed individuals.—In total, 46870

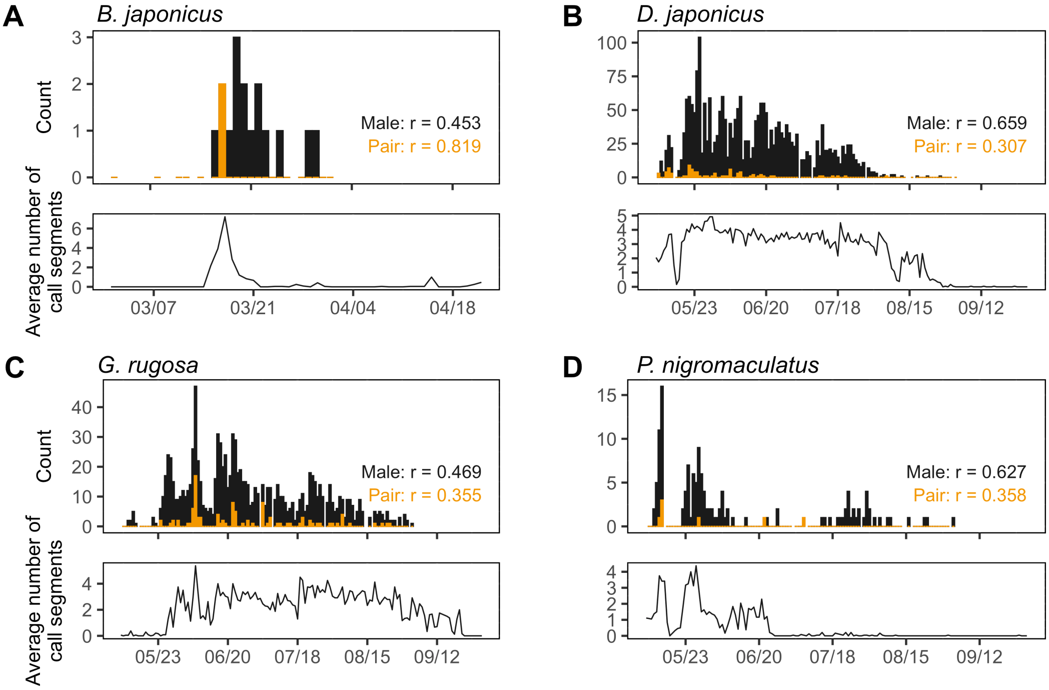

Fig. 4. Relationships between the number of males or pairs and calling activity, quantified by the daily average number of call segments. Black bars represent the numbers of males, and orange represent the numbers of pairs in amplexus. Pearson's correlation coefficients (*r*) between abundance and calling activity are shown.

spectrograms were analyzed with the trained ResNet18 model to reveal the calling activity of the target species (Fig. 3). The pattern of calling activity varied among species. *Bufo japonicus* had a short breeding season, while other species, especially *D. japonicus* and *G. rugosa*, showed prolonged breeding activity. The spring breeders (*B. japonicus* and *Z. schlegelii*) called both day and night, but the summer breeders (*D. japonicus, G. rugosa,* and *P. nigromaculatus*) called mostly at night (Fig. 3).

Figure 4 shows the relationships between calling activity and the number of males observed; these variables were positively correlated with each other (*r* = 0.453–0.659). For all species, the GLS regression analysis showed that there was a significant correlation between the number of males and the number of call segments (Table 2).

The number of amplexing pairs was significantly related to the number of call segments for most species, and for the pairs of *D. japonicus*, it was marginally significant (*p* = 0.07) (Table 2). The correlations between calling activity and the number of pairs were lower than those for males (*r* = 0.307–0.358), except for *B. japonicus* (*r* for males = 0.453 vs. *r* for pairs = 0.819) (Fig. 4). The peak of calling activity of *B. japonicus* clearly coincided with the sole day when amplexing pairs were observed (Fig. 4A). Although models with similar AIC values were present, the estimates of the top models were essentially identical and did not influence our conclusion (Supplemental material Tables S1, S2).

DISCUSSION

We evaluated the utility of deep learning in automatically monitoring reproductive phenology of anurans and found that the CNN model trained on a

Table 2. Generalized least squares (GLS) regression analyses testing the relationships between the number of call segments and either the number of males or the number of amplexing pairs. Bold font indicates statistical significance.

Species	Response variable	Correlation structure	Estimate	Std. Error	*t*	*df*
Bufo japonicus	male	MA(1)	0.905	0.283	3.197	19
	pair	ARMA(1,1)	1.567	0.136	11.493	19
Dryophytes japonicus	male	ARMA(2,1)	0.011	0.004	2.624	108
	pair	ARMA(2,1)	0.077	0.043	1.786	108
Glandirana rugosa	male	ARMA(1,1)	0.071	0.011	6.397	108
	pair	ARMA(1,1)	0.180	0.031	5.840	108
Pelophylax nigromaculatus	male	AR(2)	0.130	0.025	5.133	108
	pair	AR(2)	0.457	0.141	3.244	108

few thousand spectrogram images was capable of identifying five anuran species in field recordings. This automatic identification greatly reduced the number of manually analyzed audio segments from 46870 to 2755, a 94% reduction. In the field, our acoustic monitoring required only a few minutes to replace an SD card and battery once every three months. This was in contrast to the direct observation (visual encounter survey and dip netting), which took 10 to 60 minutes every day and 67 hours in total. Test performance based on our validation dataset showed that the trained model identified the target species with high precision (91–100%) and moderate to high recall (75–98%) values (Fig. 1). Relatively lower model performance on *P. nigromaculatus* and *G. rugosa* could be the results of low signal-to-noise ratios; calls from these species may have been masked by the calls from the dominant and intensely vocalizing species, *D. japonicus*. Although an acceptable level of identification errors may differ depending on the purpose, it would be safe to say that species with both precision and recall values higher than 90% can be analyzed practically. Even for species with moderate recall values, automatic identification models would be highly informative in reconstructing reproductive phenology if the detection rate does not vary substantially among seasons. Indeed, the estimated patterns of calling activity shown in Figure 3 were generally consistent with previously reported reproductive phenology of these species (Okuno, 1985; Fukuyama, 1991; Shimoyama, 1993; Chang, 1994; Matsui and Maeda, 2018). In addition, the model detected an unexpected calling behavior of *Bufo japonicus* on April 15, almost one month after the peak breeding period (Fig. 3). Identity of the calls was later confirmed by the authors. Such sporadic calling activity may not be detected without a comprehensive analysis of the long-term recordings. These results suggest that the deep learning approach is useful in detecting anuran calls from field recordings and revealing details of their breeding activities.

The number of call segments was high when males were abundant at the breeding sites (Table 2). The moderate correlations between these two variables (Fig. 4) suggest that field recordings can provide not only binary information of presence or absence, but also some quantitative estimates of relative male abundance. Interestingly, some peaks in the calling activity coincided with the increases in male abundance on those days (Fig. 4). This result was consistent with a previous study reporting that the number of individuals estimated from calls was significantly correlated with the number of males captured at the same night (Shirose et al., 1997). However, there are some caveats in our analysis. First, it was difficult to distinguish non-breeding males with secondary sexual characteristics from actively breeding males. For example, most male *Pelophylax nigromaculatus* observed after late June might not be reproductively active because few calls were detected during that period (Fig. 4D). Our sight-count data included both breeding and possibly non-breeding males, which can affect the correlations between the number of males and the detected calling activity. Second, the studied experimental farm was composed of sections of rice fields with different irrigation schedules. Timing of irrigation can affect reproductive phenology of anurans (e.g. Shimada et al., 2013). In our study, anuran breeding may have occurred asynchronously within and outside the study route, and this may have obscured the correlations between sight and call counts. Third, the observed male abundance would be subject to stochastic errors arising from incomplete detection. While we have no information on the detection rates for our target species, detection would likely be high for *B. japonicus* due to the small size of the breeding pond, but relatively low for the three species inhabiting the experimental farm where we conducted visual encounter surveys. These points can be the source of uncertainty in the relationship between male abundance and the calling activity. We also find that, although the number of amplexing pairs was high on the nights of intense male vocalization (Table 2), the correlations were not strong for most species (Fig. 4). In *B. japonicus*, however, the correlation of call and amplexing pair counts was higher than that of call and male counts (Fig. 4A); this was attributed to the fact that non-breeding males remained at the pond even after they stopped calling.

Our computational experiments on the effects of different training datasets on model performance suggest some strategies for efficient data annotation. Models trained on seasonally biased datasets performed worse than those trained on more diverse data (Fig. 2), implying that the training dataset should include data from broad recording dates. The lower performance of the model can be attributed to the limited number of training samples in the *Z. schlegelii* class, as they rarely called during the first third of our recording period (see Fig. 3). The decrease in call frequency of our target species during the final third of the recording period may have resulted in a small difference between the models trained on two-thirds and all category samples. Regarding the number of training samples, 1000 samples were required to achieve 95% precision and 80% recall. In our experience, this corresponded to about 10 working hours for an experienced person to label data. It should be noted that the recall rate did not saturate with our 2000 samples (Fig. 2C). The model performance would be improved by adding further training samples, especially from species with low recall rates such as *G. rugosa* and *P. nigromaculatus*.

In conclusion, this study demonstrated the ability of a CNN model to accurately identify calls of five anuran species in complex field recordings, and to provide some information on the relative abundance of males at breeding sites. However, caution is needed to link the number of call segments and the number of pairs in amplexus that would reflect female reproductive activity. These results will serve as a base for future studies adopting a deep-learning approach to monitoring anuran reproduction. This method reduces the time required for data analysis and allows researchers to handle long-term recordings from multiple sites in a standardized manner. One of the main challenges would be the limited availability of open acoustic databases, especially for Asian frogs (Womack et al., 2022). Moreover, the existing databases are mostly of focal species recordings, which may perform poorly when applied to omnidirectional soundscape recordings (Kahl et al., 2021). Another possible area of future work would be to improve abundance estimates from acoustic data. Call counts to estimate population abundance have sometimes been used to study marine mammals (Marques et al., 2013), but this may be difficult for some anurans that perform dense and continuous chorusing. Rapid advances in sound source separation of overlapping calls may help to overcome this problem (Bermant, 2021; Denton et al., 2022).

DATA ACCESSIBILITY

Supplemental material is available at https://www.ichthyologyandherpetology.org/h2023018 and code to train the CNN models is available at https://github.com/kaede-kmr/evaluate-DL-monitoring. Unless an alternative copyright or statement noting that a figure is reprinted from a previous source is noted in a figure caption, the published images and illustrations in this article are licensed by the American Society of Ichthyologists and Herpetologists for use if the use includes a citation to the original source (American Society of Ichthyologists and Herpetologists, the DOI of the Ichthyology & Herpetology article, and any individual image credits listed in the figure caption) in accordance with the Creative Commons Attribution CC BY License.

LITERATURE CITED

Bermant, P. C. 2021. BioCPPNet: automatic bioacoustic source separation with deep neural networks. Scientific Reports 11:23502.

Besson, M., J. Alison, K. Bjerge, T. E. Gorochowski, T. T. Høye, T. Jucker, H. M. R. Mann, and C. F. Clements. 2022. Towards the fully automated monitoring of ecological communities. Ecology Letters 25:2753–2775.

Borowiec, M. L., R. B. Dikow, P. B. Frandsen, A. McKeeken, G. Valentini, and A. E. White. 2022. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution 13:1640–1660.

Chang, J. C. W. 1994. Multiple spawning in a female *Rana rugosa*. Japanese Journal of Herpetology 15:112–115.

Christin, S., É. Hervet, and N. Lecomte. 2019. Applications for deep learning in ecology. Methods in Ecology and Evolution 10:1632–1644.

Denton, T., S. Wisdom, and J. R. Hershey. 2022. Improving bird classification with unsupervised sound separation. p. 636–640. *In*: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Dorcas, M. E., S. J. Price, S. C. Walls, and W. J. Barichivich. 2009. Auditory monitoring of anuran populations. p. 281–298. *In*: Amphibian Ecology and Conservation: a Handbook of Techniques. C. K. Dodd (ed.). Oxford University Press, Oxford.

Fukuyama, K. 1991. Spawning behaviour and male mating tactics of a foam-nesting treefrog, *Rhacophorus schlegelii*. Animal Behaviour 42:193–199.

Gibb, R., E. Browning, P. Glover-Kapfer, and K. E. Jones. 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods in Ecology and Evolution 10:169–185.

Hill, A. P., P. Prince, E. Piña Covarrubias, C. P. Doncaster, J. L. Snaddon, and A. Rogers. 2018. AudioMoth: evaluation of a smart open acoustic device for monitoring biodiversity and the environment. Methods in Ecology and Evolution 9:1199–1211.

Howard, J., and S. Gugger. 2020. Fastai: a layered API for deep learning. Information. An International Interdisciplinary Journal 11:108.

Kahl, S., C. M. Wood, M. Eibl, and H. Klinck. 2021. BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics 61:101236.

Keitt, T. H., and E. S. Abelson. 2021. Ecology in the age of automation. Science 373:858–859.

Knight, E. C., K. C. Hannah, G. J. Foley, C. D. Scott, R. M. Brigham, and E. Bayne. 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12:14.

LeBien, J., M. Zhong, M. Campos-Cerqueira, J. P. Velev, R. Dodhia, J. L. Ferres, and T. M. Aide. 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Ecological Informatics 59:101113.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature 521:436–444.

Liu, G., R. T. Kingsford, C. T. Callaghan, and J. J. L. Rowley. 2022. Anthropogenic habitat modification alters calling phenology of frogs. Global Change Biology 28:6194–6208.

Mac Aodha, O., R. Gibb, K. E. Barlow, E. Browning, M. Firman, R. Freeman, B. Harder, L. Kinsey, G. R. Mead, S. E. Newson, I. Pandourski, S. Parsons, J. Russ, A. Szodoray-Paradi … K. E. Jones. 2018. Bat detective-Deep learning tools for bat acoustic signal detection. PLoS Computational Biology 14:e1005995.

Marques, T. A., L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack. 2013. Estimating animal population density using passive acoustics. Biological Reviews of the Cambridge Philosophical Society 88:287–309.

Matsui, M., and N. Maeda. 2018. Encyclopedia of Japanese frogs. Bun-ichi Sogo Shuppan, Tokyo, Japan.

Matsushima, N., M. Hasegawa, and J. Nishihiro. 2022. Effects of landscape heterogeneity at multiple spatial scales on paddy field-breeding frogs in a large alluvial plain in Japan. Wetlands 42:106.

McFee, B., A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, … Thassilo. 2022. librosa: 0.9.1. https://doi.org/10.5281/zenodo.6097378

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences of the United States of America 115:E5716–E5725.

Okuno, R. 1985. Studies on the natural history of the Japanese Toad, *Bufo japonicus japonicus*. VIII. Climatic factors influencing the breeding activity. Japanese Journal of Ecology 35:527–535.

Pinheiro, J., D. Bates, and R Core Team. 2022. nlme: linear and nonlinear mixed Effects models. R package version 3.1-162. https://cran.r-project.org/web/packages/nlme/index.html

R Core Team. 2022. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rowley, J. J. L., C. T. Callaghan, T. Cutajar, C. Portway, K. Potter, S. Mahony, D. F. Trembath, P. Flemons, and A. Woods. 2019. FrogID: Citizen scientists provide validated biodiversity data on frogs of Australia. Herpetological Conservation and Biology 14:155–170.

Schneider, S., G. W. Taylor, S. C. Kremer, P. Burgess, J. McGroarty, K. Mitsui, A. Zhuang, J. R. deWaard, and J. M. Fryxell. 2022. Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. Methods in Ecology and Evolution 13:346–357.

Shimada, T., A. Imamura, and N. Ohnishi. 2013. A study of larval phenologies of five anuran species in

Japanese paddy fields. Japanese Journal of Herpetology 2013:77–85.

Shimoyama, R. 1993. Female reproductive traits in a population of the pond frog, *Rana nigromaculata*, with prolonged breeding season. Japanese Journal of Herpetology 15:37–41.

Shirose, L. J., C. A. Bishop, D. M. Green, C. J. MacDonald, B. R. J., and N. J. Helferty. 1997. Validation tests of an amphibian call count survey technique in Ontario, Canada. Herpetologica 53:312–320.

Stowell, D. 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10:e13152.

Weir, L., I. J. Fiske, and J. A. Royle. 2009. Trends in anuran occupancy from northeastern states of the North American Amphibian Monitoring Program. Herpetological Conservation and Biology 4:389–402.

Womack, M. C., E. Steigerwald, D. C. Blackburn, D. C. Cannatella, A. Catenazzi, J. Che, M. S. Koo, J. A. McGuire, S. R. Ron, C. L. Spencer, V. T. Vredenburg, and R. D. Tarvin. 2022. State of the amphibia 2020: a review of five years of amphibian research and existing resources. Ichthyology & Herpetology 110:638–661.

Xie, J., R. Zeng, C. L. Xu, J. L. Zhang, and P. Roe. 2017. Multi-label classification of frog species via deep learning. p. 187–193. *In*: 2017 IEEE 13th International Conference on e-Science (e-Science).

Zuur, A. F., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. Mixed effects models and extensions in ecology with R. Springer, New York.