A Successful Crowdsourcing Approach for Bird Sound Classification

PETTERI LEHIKOINEN [iD]

MEERI RANNISTO [iD]

ULISSES CAMARGO [iD]

AKI AINTILA

PATRIK LAUHA [iD]

ESKO PIIRAINEN

PANU SOMERVUO [iD]

OTSO OVASKAINEN [iD]

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

Automated recorders are increasingly used in remote sensing of wildlife, yet automated methods of processing the audio remains challenging. Identifying animal sounds with machine learning provides a solution, but optimizing the models requires annotated training data. Producing such data can require much manual effort, which could be alleviated by engaging masses to contribute to research and share the workload. Birdwatchers are experts on identifying bird vocalizations and form an ideal focal audience for a citizen science project aiming for the required multitudes of annotated avian audio data. For this purpose, we launched a web portal that was targeted and advertised to Finnish birdwatchers. The users were asked to complete two kinds of tasks: 1) classify if a given bird sound belonged to the focal species and 2) classify all the bird species vocalizing in 10-second audio clips. In less than a year, the portal achieved annotations for 244,300 bird sounds and 5,358 clips, and attracted, on average, 70 visitors on daily basis. More than 200 birdwatchers took part in the classification tasks, of which 17 and 4 most dedicated users produced over half of the sound and clip classifications, respectively. As expected of birder experts, the classifications among users were highly consistent (mean agreement scores between 0.85–0.95, depending on the audio type) and resulted in high-quality training data for parameterizing machine learning models. Feedback about the web portal suggested that additional functionality such as increased freedom of choice would increase user motivation and dedication.

INTRODUCTION

Recent increased availability of affordable automated wildlife recorders has extensively enabled access to remotely sense soundscapes for ecological research (Farina, Pieretti, and Piccioli 2011; Hill et al. 2018; Krause and Farina 2016). Although obtaining massive amounts of soundscape audio has become easy, processing the large amount of data to extract the classifications of vocalizing wildlife is challenging, and a prerequisite for ecological inference (Gibb et al. 2019, and the reference therein, Loiselle et al. 2003). Automated classification through machine learning has proven to be time-efficient for this task (LeBien et al. 2020; Ruff et al. 2020; Stowell et al. 2019). Although there are several applications readily available for automated classification of birds (Kahl et al. 2021; Warblr Ltd 2022), their geographical range remains limited to certain countries or continents. Also, variations in recording conditions present challenges to the current methods (Gibb et al. 2019). Some classification methods may work well with high-quality data obtained from directional microphones in a quiet environment but fail when recording quality decreases. Typically, the soundscape data are collected without human intervention using omni-directional microphones that gather data from the environment, including both wanted and unwanted sound sources. Signal-to-noise ratios in such recordings can vary from nearly perfect to the extreme case in which the noise completely masks the sound of interest. To train a well-performing machine learning model, the training data should be representative and as rich as possible (Gibb et al. 2019).

Classification through machine learning requires annotated training data (Camargo, Roslin, and Ovaskainen 2019; Kahl et al. 2021). Such data are readily available in online libraries (Cornell Lab of Ornithology 2022; Xeno-canto Foundation 2022), but the use of these data for training machine learning models is compromised by two features. First, annotations in these libraries are made at the recording-level accuracy, meaning that while it is known that a given species is vocalizing in the recordings, its exact locations and all the other vocal species in the recording remain commonly unspecified. Second, the geographic coverage of the recordings does not necessarily overlap with the region from where the data to be classified has been collected, thus differences in bird dialect can lead to difficulties in classification. While the data available in online libraries can be used to train deep neural networks for species classification (Kahl et al. 2021; Salamon and Bello 2017), having access to fully annotated data that would be representative of the region for which the classifications are to be made would likely increase the classification

ability of the resulting models (Lauha et al. 2022). However, full annotation of recordings for a complete set of species is a laborious task (Gibb et al. 2019; Papadopoulos, Roberts, and Willis 2015).

Recently, citizen science has become more popular among scientists since it has potential to promote education and share knowledge on research processes and results (Franzoni and Sauermann 2014), to increase public engagement and outreach, and to increase productivity as a result of a potentially enormous group of people taking part in the research (Bonney et al. 2014; Van Brussel and Huyse 2019). Citizen science, therefore, offers a framework that has a huge potential for completing large-scale and sometimes onerous tasks by sharing the workload broadly (Franzoni, Poetz, and Sauermann 2021). Bird watching is a common leisure pursuit in developed countries, and the numerous ornithological hobbyists have extensive experience in identifying avian vocalizations. Thus, they are an ideal focal audience for crowdsourcing fully annotated training material used in development of automated bird sound classification methods (Snyder et al. 2022). Furthermore, the expertise of birdwatchers alleviates the concern of data quality, which is often used—albeit seemingly inappropriately—to criticize citizen science projects (Stevenson et al. 2021). These concerns are further reduced when the data consistency among users is studied, controlled, and compared with professionals (Lintott et al. 2008, 2011; Walmsley et al. 2020). For success, citizen science projects should also understand what motivates participants to engage, and should provide further benefits for them (Land-Zandstra, Agnello, and Gültekin 2021). The audiovisual inspection and recognition of bird vocalizations provide a fruitful educational aspect to citizen science audio projects related to birds but have thus far been underutilized by practitioners (Constantine and The Sound Approach 2006).

This project was motivated by the need to develop improved machine learning models for classifying Finnish bird species from soundscape recordings. To support this development task, we launched a web portal for a citizen science project targeted to bird watchers. The aim of the web portal was to produce testing and training data for machine learning purposes from autonomous field recordings—a similar approach to that of Snyder et al (2022). In addition to gaining these data, the web portal was planned to aid users to test their sound identification skills and give feedback of them, and to provide a platform where users could familiarize themselves with audiovisual presentation for the vocalization of the bird species in southern Finland. Here we present 1) the web portal, 2) the engagement process of birdwatchers and its effectiveness, 3) the data acquired and its consistency.

Given that the limited availability of comprehensive expert-validated species acoustics data has been raised as the most fundamental knowledge gap for future development of passive acoustic monitoring (Gibb et al. 2019), we have also published the sound data and metadata on an open source platform (https://zenodo.org/record/7030863#. Y6GgtIRBwuU; Lauha et al. 2022).

MATERIAL AND METHODS

SOUND DATA

Sound data were gathered at ten sites in southern Finland with four recorders at each site. Sites were selected to be in near proximity to standardized line transects of the national bird monitoring scheme (25 km apart) to enable comparisons between traditional monitoring methods and autonomous recordings. Traditional monitoring methods have produced a list of 96 species at these sites (Supplemental File 1: Supplemental Table 1). At sites, the recorders were placed along the transect not closer than 250 m from each other so that they covered all the habitats available along the transect. Recorders were in field from 8 May 2018 until 11 July 2018, a time period that covers approximately the breeding season in southern Finland.

SELECTION OF ANNOTATION UNITS AND THEIR TASK ALLOCATION FOR USERS

From the total of 1,810,194 minutes of sound data, we selected two audio types that were presented to the users of the web portal: 1) templates and their candidates and 2) 10-second-long clips. For the templates, we selected and manually refined the vocalizations that contained characteristic variation for each species. Thus, a template was a representative sound of the species, and could be a short call or a long phrase of a song. A single species could have many templates to capture its variation in vocalization. The average length (± SD) of a template was 0.80 s ± 0.55 s, and on average, a species had 4.7 ± 4.4 templates.

Based on the templates, the user was asked to score 100 audio candidates (hereafter candidates) of the same duration as the template as either including or not including the focal species of the template. To select these 100 candidates, we scanned the audio data to find the best match between each segment and each template (Ovaskainen, Camargo, and Somervuo 2018). We first asked the user to annotate ten candidates with a varying resemblance to the template. We then chose adaptively the remaining 90 candidates close to the cut-off resemblance. The cut-off resemblance was defined as the cross-correlation value of the candidate above which the users identify the species as present and below which the users identify the species as being absent. We also presented the same candidates for different users in order to get ca. 25% of the candidates to be annotated by all users. Templates were chosen based on the number of user classifications so that the template that had the least classifications was presented to the next user who was classifying templates.

The 10-second clips were centered around a candidate vocalization where a target template species was identified to be present by the users. This was done in order for the clips to include at least one vocalizing bird species. To generate a sufficient level of overlap among clips annotated by different users, 20% of the clips were randomly chosen among those already annotated by another user.

For a detailed description of the selection of annotation units and their task allocation for users, see Supplemental File 2.

ATTRACTING USERS

The focal audience was Finnish birdwatchers who had the necessary experience in identifying bird vocalizations from the geographical region of interest. This recruiting was focused on the forums used by Finnish ornithological experts and hobbyists. Advertising was done to attract visitors, and thus potential users, to the web portal. In the advertisements, we shared news about the use of the web portal, how the classification numbers were developing, and how the data to be generated would benefit the development of improved machine learning algorithms for automated bird classification. Advertising was done in Finnish on Twitter and Facebook, for which we made own profiles for the web portal (Kerttu—kerro tunnistuksesi and @KerroKerttu, respectively). News in Facebook was further shared to birdwatching groups that had both national and local scope. We used email lists and the press media of national and local ornithological societies to disseminate information, and we gave oral presentations in the monthly meetings of these societies. In addition, we actively encouraged users to send in their feedback on the web portal, and we also interviewed some of the users so that we could improve the user experience and reward of using the portal.

The web portal was opened to the public on 19 January 2021 and was first advertised only on an email list of limited people actively participating in the national bird monitoring scheme. This was done to invite a small number of expert test users who were asked to give comments on any apparent flaws in the portal, or on improvements, which could then be revised before advertising the web portal more extensively. We started wider advertising of the web portal on 27 January 2021, but did not share the news in

all the forums at once to avoid any problems arising from excessive traffic at the website. Due to summer holidays, advertising was stopped at the end of June 2021. We used Google Analytics to keep track of the daily page views as a proxy of users visiting the web portal.

WEB PORTAL

The web portal could be accessed through the Finnish Biodiversity Information Facility (FinBIF) webpages. To use the service, login was required as the classifications were stored on a per-user basis. The portal had five sections: Instructions, Declare your level of expertise, Identify letters, Identify recordings, and Results.

Before making any classifications, the user was asked to state their expertise in the *Declare your level of expertise* section. After that, they could move on to produce template classifications in the *Identify letters* section. The section *Identify recordings* was unlocked after the user had made 20 sets of template classifications. This restriction was placed to ensure that the users generated enough template classifications, which required more work than the clip classifications. After the user had unlocked that section, they could freely choose whether to classify templates or clips.

The web service consisted of a back end (https:// bitbucket.org/luomus/kerttu-backend) and a front end (https://github.com/luomus/laji) that were both open source. The back end was developed with the Python Flask framework and the front end with the TypeScript Angular framework. The front end relied on the Web Audio API, which allowed manipulating audio and generating spectrograms on the client side.

Declare your level of expertise

In this section, the user was asked to declare their level of expertise. First, they were asked to evaluate their general level of expertise by answering to two multiple choice questions: "How easily are you able to identify the vocalizations of bird species occurring in Finland?" and "How actively do you engage in birdwatching?" (Figure 1). Both had four options from which to choose, and the options roughly followed cases in which 1) the user had very little experience on bird vocalization and birdwatching, 2) had some experience, 3) had lots of experience and 4) was a professional in these fields.

In addition to the general questions, the user was asked to select all species whose vocalizations they could at least partly identify. They chose the species from a list that contained the most common bird species in Finland, i.e., all species that had 50 or more observations recorded in the FinBIF, which totaled 349 species (Figure 1). The selected species determined which species were shown to the user in the *Identify letters* section.

Identify letters

In this section, the user was asked to classify candidates for the templates, which were called letters in the web portal. The system chose a template for the user that belonged to a species that they could classify according to the *Declare your level of expertise* section and presented a candidate for this template. The user classified the candidate by answering "Yes," "No," or "I don't know" to the question "Is the species vocalizing in the candidate the same as the species in the letter," where the letter meant template of the species that was given on top of the page (Figure 2). After the user answered, the system chose the next candidate. The user had to make 100 classifications (yes or no answers) for a template before moving on to the next. If the user was for some reason not able to make the classifications for a certain template, they could skip it.

The template and candidate were shown in an audio viewer. The audio viewer showed a spectrogram of the audio and allowed user to listen to it. The exact location of both candidate and template were shown with a white box in the spectrogram. The audio viewers had two settings, "Time Buffer" and "Focus Frequency." Time buffer was the time shown before and after the candidate and template. This setting was useful when classifying vocalizations that were hard to classify without a wider context. A broad time buffer allowed the user to see and listen to other vocalizations of the same individual given before and after the candidate. The focus frequency setting allowed selecting a range of frequencies around the candidate and affected both the spectrogram image and the sound. It thus zoomed in to a given frequency range of the spectrogram, and when playing the recording, it attenuated the frequencies outside the candidate with a bandpass filter. This setting helped classification by allowing, for example, for the removal of background noise that did not overlap with the focal frequencies.

Identify recordings

After the user had completed classification of 20 sets of candidates for templates in the *Identify letters* section, they could move on to classify clips. In this section, the user was asked to classify all bird species in a 10-second clip (Figure 3). If the user was unsure about the classification, they could check "occurs possibly" button, otherwise "occurs certainly" was checked by default. The user was also asked to provide additional information about the recording, such as if the recording was of poor quality and whether it contained human speech or other human activity. Users were also asked if the recording contained bird vocalizations they did not recognize.

The clip was shown in an audio viewer. It showed the 10-second clip and 1 second before and after it with a

Figure 1 A view of the web portal section "Declare your level of expertise."

darkened background for a wider context. The audio viewer had a zoom functionality that allowed the user to zoom in freely to the recording. When playing the zoomed-in recording, the frequencies outside the zoomed area were attenuated.

When saving the classifications of a user to the database, the system took into account also those species that the user had listed in the *Declare your level of expertise* section as species that they would be able to identify, but that they had not marked to be present in the focal clip. In this way, the system collected data not only on species presences, but also on species absences.

Web portal results

The *Results* section showed how many classifications had been made in total, how many of them had been done by the signed-in user, and how many by all users who had participated (as a list) (Figure 4). The user could
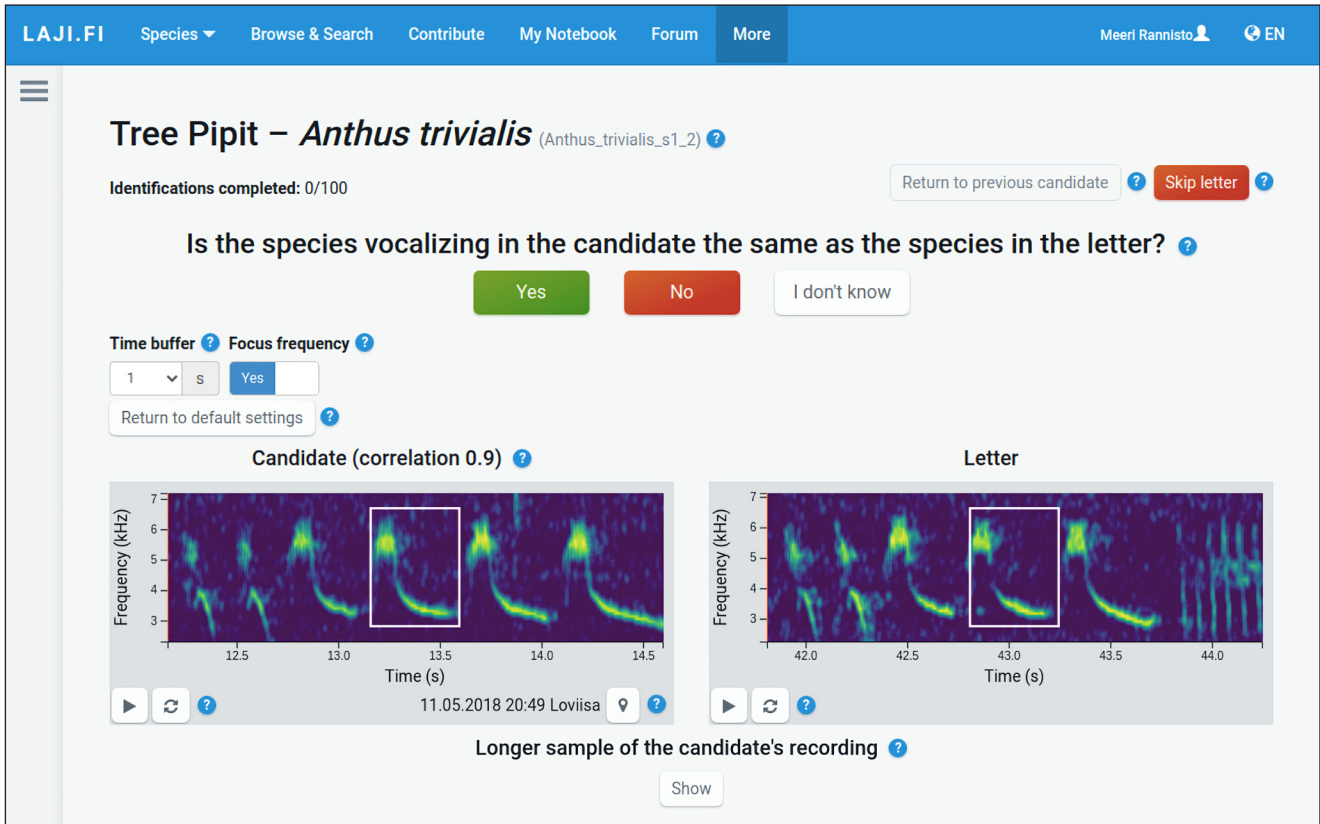
Figure 2 A view of the web portal section "Identify letters" (note that templates were called letters in the web portal).
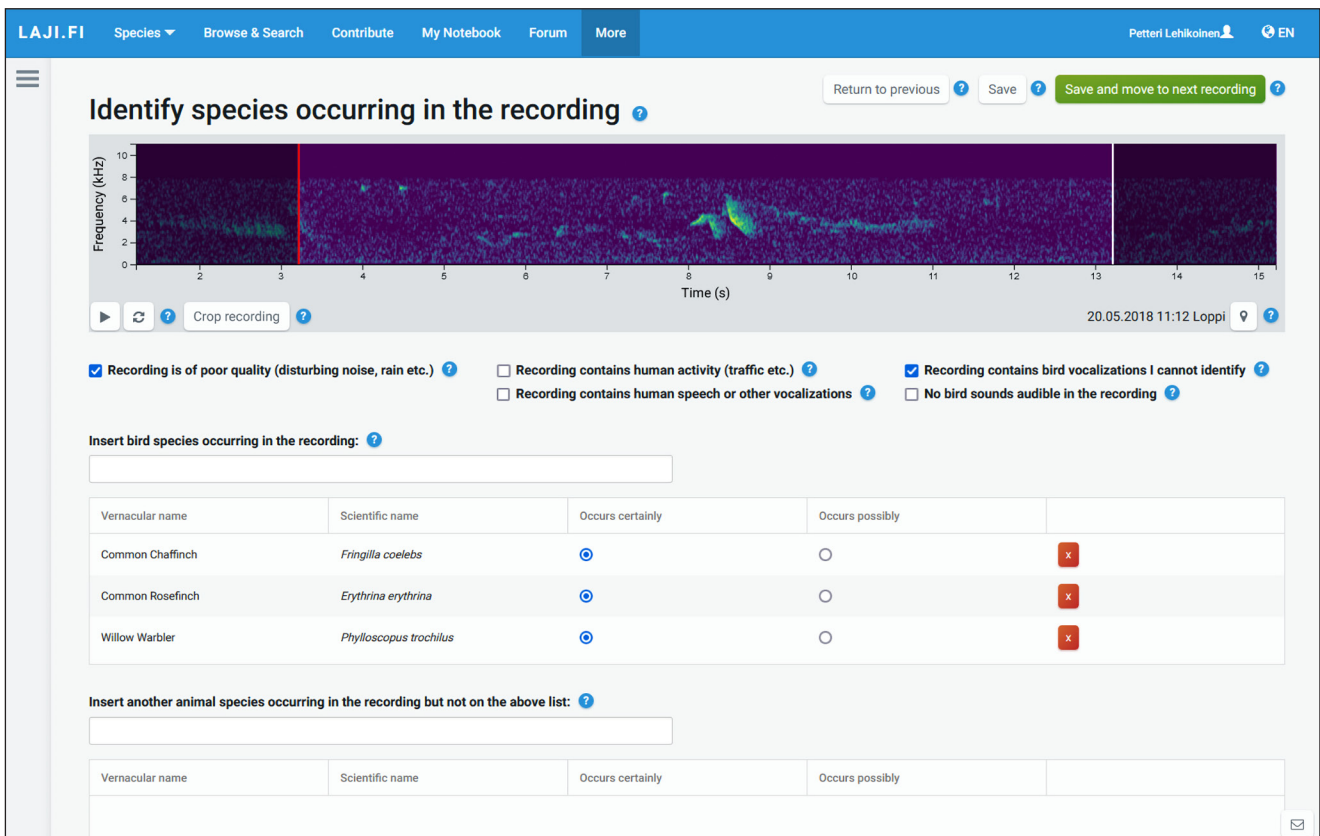

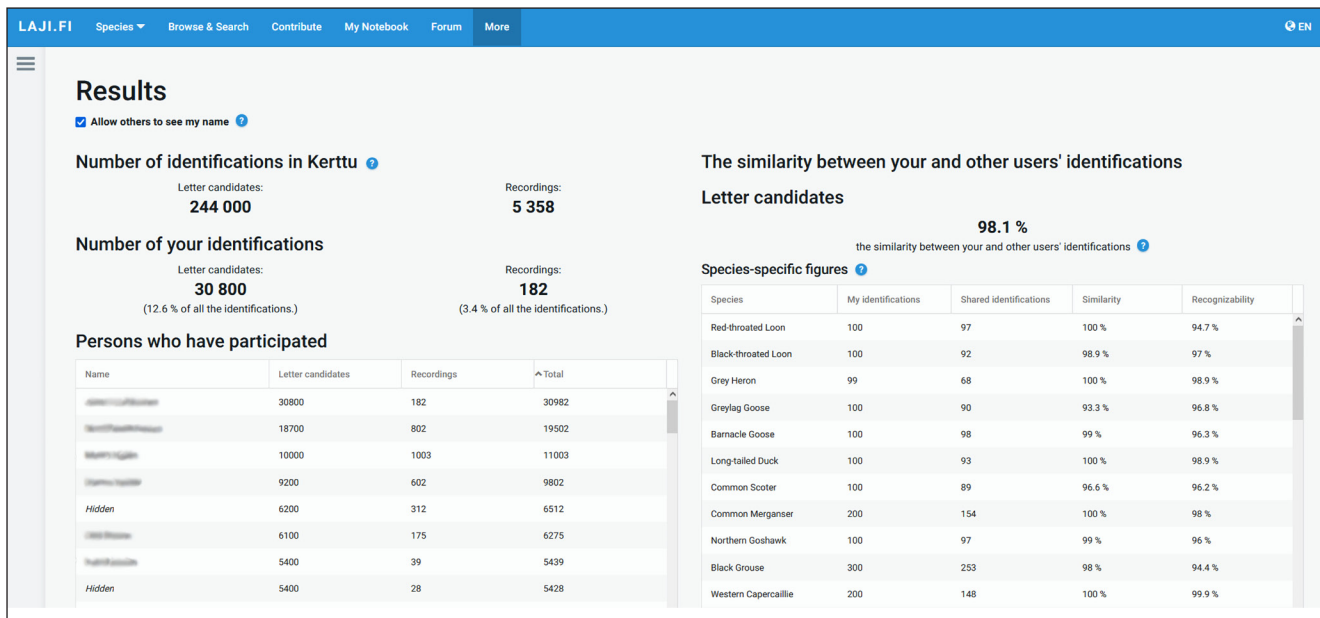**Figure 3** A view of the web portal section "Identify recordings."

Figure 4 A view of the web portal section "Results," which presents the number of all classifications made in the portal as well as those made by the user themselves and other users who have approved public visibility of their names. The results also show how well the users' classifications (identifications in the portal) correspond to the classifications of other users of the same recordings given at species level, and, they show a species-specific recognizability, which is the average proportion of unanimous classifications by candidates among all users.

choose whether their name was shown or if they remained anonymous.

The *Results* page also aimed to provide additional value to the user by showing how similar their template classifications were compared with others. It showed two values to the user for each species: similarity and recognizability (Figure 4). Similarity was the percent of user classifications that matched the majority vote (only candidates that had classifications from at least three different users were taken into account). Recognizability represented how well the species was recognized in general by showing the average proportion of unanimous classifications by candidates among all users. These values gave the user information on which species their classifications coincided or did not coincide with classifications made by other users. The similarity and recognizability were introduced to the portal only three months after it was made public (on 21st April) to ensure that adequate data had been received for the calculation of the similarity and recognizability values and to add a new feature to the portal, which could give additional value for the existing users to continue producing new classifications.

DATA QUALITY EVALUATION

The consistency of the classifications made by different users was evaluated separately for candidates and clips as an agreement score. For each candidate, a majority vote of the classifications was calculated by discarding uncertain votes and calculating the share of positive votes among all positive (1) and negative (0) votes. The agreement score was evaluated for all candidates with at least three positive or negative classifications by calculating the proportion of agreeing votes, that is $\max(m, 1\text{-}m)$, where m denotes the majority vote of the candidate. The candidates included in the majority vote calculation were classified by an average of 9.1 ± 10.3 (SD) users, and the majority vote could be calculated to 18,870 (24.3%) out of all 77,743 classified candidates.

For clips, the users had listed all species that they were able to recognize in the recording and noted whether there were any other vocalisations that they were not able to recognize. The agreement score was evaluated for all clips that at least three people had labelled. As described above in the section "Identify recordings," we generated data on species absences on the basis of the species the users had declared that they could identify. We discarded these data on absences if a user had marked that the clip contained vocalisations that they were unable to identify. Based on the data, species-specific majority votes were calculated for each clip. The agreement score was calculated as explained above for all species that at least one person had classified for the clip.

We also calculated user-specific classification accuracy indices for candidates and clips. For candidates, the accuracy score was calculated as the proportion of candidates, for which the user's classification matched the rounded majority vote of the letter. The same similarity value was

shared personally with the signed-in user in the *Results*. For clips, we calculated precision score, which described what proportion of user's all-positive votes were correct according to the majority vote, and recall score, which described the proportion of the positive classifications given by the user among all species that (according to the majority vote) were present in the recordings the user had labelled. The user-specific clip accuracy score was calculated by taking the mean of the precision and the recall score.

RESULTS

AMOUNT OF PAGE VIEWS, USERS, AND IDENTIFICATIONS

Between the 19th of January and end of June 2021, altogether 11,429 daily page views were obtained (mean 70.1 ± SD 113.9). The number of page views increased often directly after the advertising events, in particular after advertisements in mailing lists and in social media (Figure 5). In general, the number of daily visitors decreased with days from the release of the web portal, and accordingly the effects of new advertising events became weaker (Figure 5).

On average, a user declared to identify at least partly the vocalization of 130.6 ± 86.5 (SD) bird species of the 349 listed in the *Declare your level of expertise* section. Given the long list of species, we checked if the users got tired of listing their expertise before they got to the end. We found

no evidence for this since the number of users who could identify the species did not decrease towards the end of the species list. Until the end of 2021, 203 distinct users had participated in the web portal, classifying a total of 244,300 candidates (sets of 100 candidates for 2,443 templates). Out of these 203 users, 43 had also classified bird species vocalizing in 5,358 clips (Figure 6). Thus, an average user had classified 12.0 ± 28.4 (SD) sets of candidates and 124.6 ± 211.6 (SD) clips. For candidates, 17 users (8.4% of all users who classified candidates) were responsible for more than 50% of the classifications, whereas four persons (9.3% of all users who classified clips) did classifications for more than 50% of all clip classifications (Figure 6).

DATA CONSISTENCY

The data set of candidates classified by at least three users contained 18,870 candidate classifications belonging to 230 templates of 73 bird species. The mean of the agreement score of candidates was 0.958 and the median was 1.00. When aggregating over the templates by calculating the average agreement score of all candidates corresponding to one template, the mean of template-specific agreement scores was 0.95. The agreement scores were not affected by the number of users who had classified the candidate. The number of clips labelled by at least three users was 303, and on average, these clips were classified by 3.7 ± 1.6 (SD) users. According to the majority vote, these clips contained, on average, 2.6 bird species. In these data, the
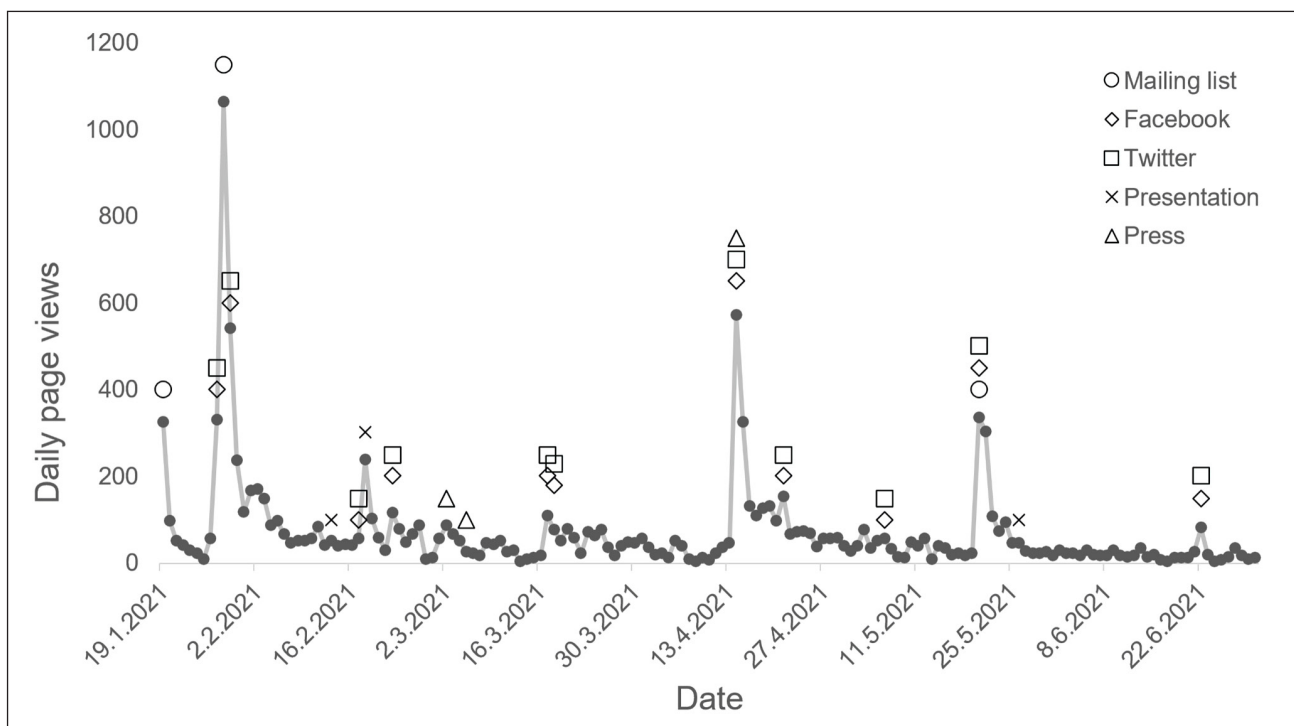

Figure 5 The number of daily page views on the web portal over time. Dates and types of advertisements are expressed as point shapes on top of the daily page views.

mean agreement score among all species that at least one person had classified in a clip was 0.850.

For candidates, the mean of user-specific classification accuracy was 0.9484, and 85% of the users had classification index greater than 0.9 (Figure 7). Users'

classifications were also highly consistent for the clips, even though these contained slightly more variation than that of the candidates. Out of 37 users, who had labelled at least some clips with two or more persons, 73% had both precision and recall higher than 0.8. We also studied
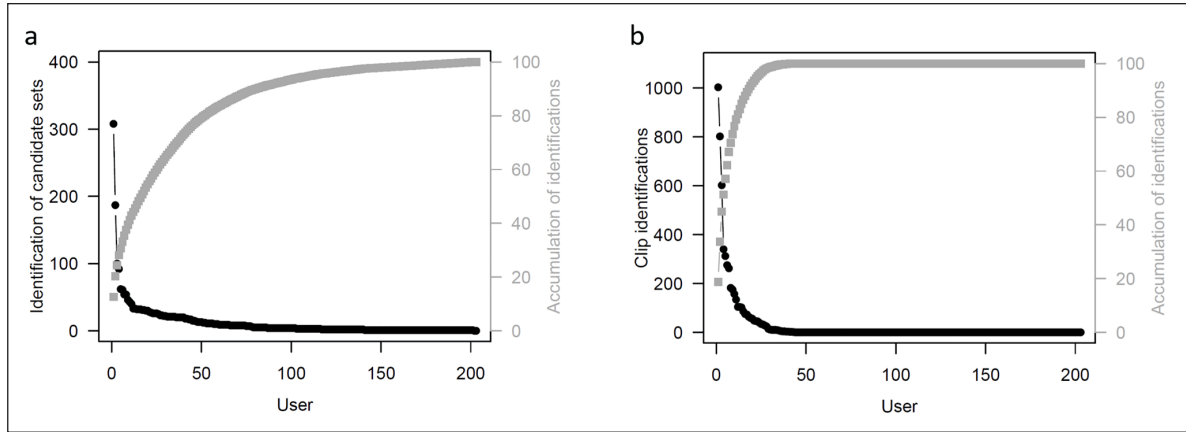

Figure 6 The numbers of classifications by the participating users arranged in descending order from left to right (*black dots*) and their accumulation curves (*grey squares*) for **(a)** the sets of candidates for templates and (b) the 10-second clips. In total, 203 users contributed to candidate classification (a), whereas 43 users also classified clips **(b)**.
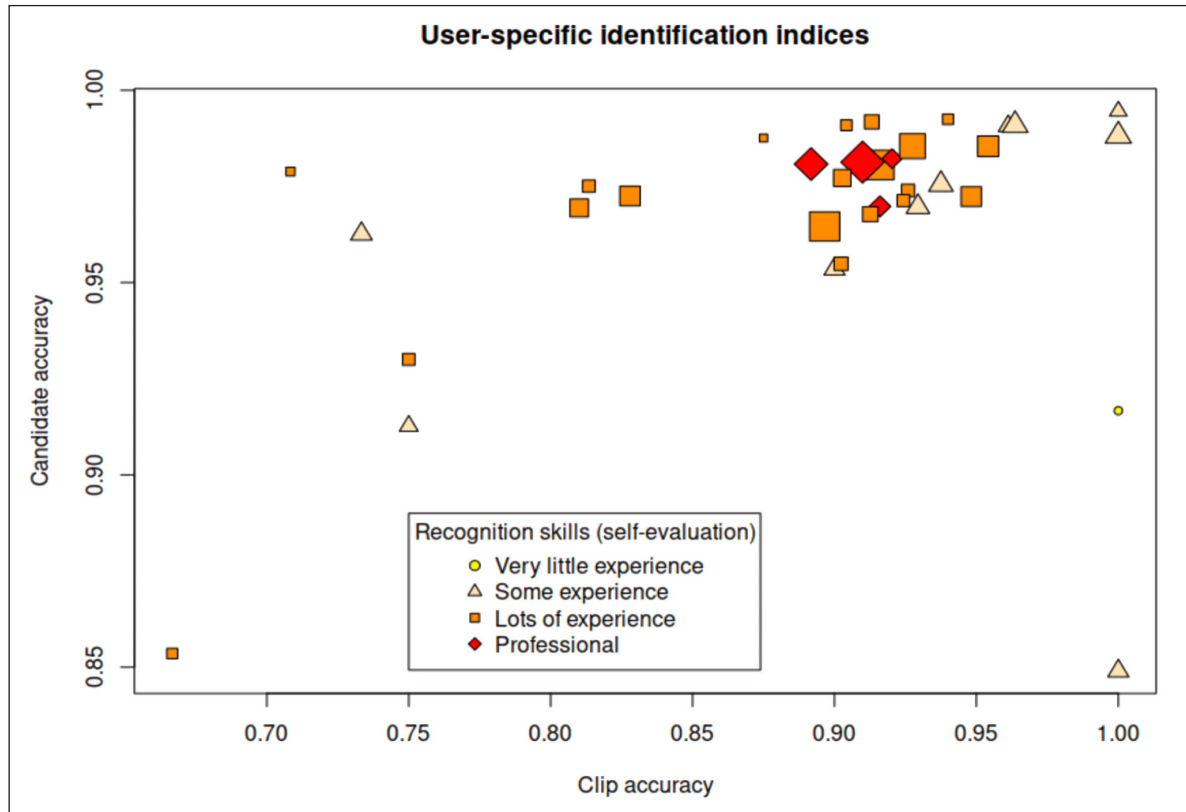

Figure 7 Variation among users in their quality of classifications. The plot shows the relation between users' accuracy in candidates and clips. For candidates, the accuracy score was calculated as the proportion of candidates for which the user's classification matched the rounded majority vote of the letter. For clips, the accuracy score was calculated by taking the mean of precision (proportion of user's all-positive votes that were correct according to the majority vote) and recall score (proportion of user's positive classifications among all species that were present in the recordings). The self-evaluation of the user's expertise is indicated by the symbol type, and the size of the symbol indicates the number of classifications made by the user (varied from 409 to 31,895 including both candidates and clips).

whether user-specific candidate accuracy scores, clip precision, or clip recall correlated with user's evaluation of their own skills or birding activity but found no strong connection (Figure 7). Clip precision had a weak negative (but not statistically significant) correlation with both user's own skill (–0.396) and level of birding activity (–0.232), while candidate accuracy had a weak positive correlation with user's skill (0.269) and level of birding activity (0.287).

DISCUSSION

Because the project generated a good amount of high-quality data, we consider this citizen science project highly successful, although we faced some challenges. Consequently, the data generated by this project were sufficient for parameterizing machine learning models that provided much improved classifications compared with earlier available methods, as described in detail by Lauha et al. (2022).

Concerning the quality of the data, our results showed all the data to be highly consistent among the users, suggesting that essentially all users provided high-quality data, even if there were many more users than just a few dedicated experts. In particular, the high precision and recall of most users suggests that only a minority of the users were either overconservative or overconfident while naming the species for the clips. The high consistency of the data is probably due to the fact that the majority of users, and likely especially the most active users, were experienced birders who are very familiar with the vocalizations of the species occurring in the recordings. Classifications of clips showed a slightly higher variation compared with candidates, which was expected, because finding and recognizing all vocalizing species from a clip is a much harder task than giving a binary label for a specific target species and a short vocalization. Room for discrepancy among the users arises especially when vocalizations of several species overlap, and/or some of them are very faint, and/or masked by different sources of background noise. We found no strong connection with the users' self-evaluation and classification accuracy, precision, or recall.

In our advertising and communication with users, we encouraged and directly asked for feedback, and especially suggestions on how to improve user experience. We aimed to quickly implement all suggestions as long as they were technically feasible, to increase the motivation and reward of use for the participants, which cannot be disregarded in a successful citizen science project (Land-Zandstra, Agnello, and Gültekin 2021). The vast majority of the feedback was positive. In particular, most users commented that the

portal was easy and intuitive to use, some even mentioning that it was addictive. Also, the easy access to audiovisual representation of the sounds collected appraisal, and several users mentioned that it was rewarding to learn how to identify common bird vocalizations solely based on the spectrogram.

LIMITATIONS IN THE DATA, PORTAL, AND ADVERTISING

Negative correlation between users' experience and precision score in clips hints that it is possible that in some cases only the most experienced birders might have found all vocalizing species and labelled the clips correctly, but their classification indices may have gotten penalized from disagreeing with the majority vote. Furthermore, the lack of clear connection between users' self-evaluated expertise and their classification indices could be explained by the fact that the majority of the data consists of common species, which are most likely identified even by less experienced birders. In addition to the high consistency of the data, the manner of data collection improved the data quality. The web portal was planned so that it produced presence-absence and not presence-only data. Presence-only data is common for citizen science projects, but can be a major constraint for further use of the data (Callaghan et al. 2021; Mengersen et al. 2017).

The advertising was targeted quite strictly to birdwatchers, who form a rather small group of people. It is likely that had the portal been advertised to the general public, without birdwatching background, it would have produced much more classification data. It remains unknown if such data would have been of high quality. Unreliable classifications would have made the use of data for automated bird sound identification modelling much more complicated, and it may not have been appropriate for our purpose, despite the increase in data. The results of Snyder et al. (2022) may indicate that using experts is more time efficient. They launched a similar system to that of the section *Identify letters* in our web portal, where expert citizen scientists delineated bounding boxes from spectrograms and assigned them to certain bird species (i.e., corresponding to our templates). These were then used to find, computationally, regions of interest (ROIs). ROIs were provided to users with both minimal or no bird vocalization knowledge and to experts in bird vocalization for classification if the focal species was either present or absent in an ROI (i.e., corresponding to our candidates). Whereas users with little knowledge of bird vocalization provided a lot of information on species absence, the vast majority of species presence classifications were provided by the bird vocalization expert users (Snyder et al. 2022).

Concerning the amount of data generated, the majority of the users visited the portal before May, after which the visitor numbers decreased. The main reason for this was likely that the high birding season in Finland begins in mid to late April with arriving migrants, and users were more likely to spend their time listening to birds *in vivo*. Furthermore, the summer holiday season starts in June, when people are even less likely to find themselves sitting in front of computer listening to birds *in vitro*. However, partly for the same reasons (i.e., field work and holidays), there was also less advertising done in May and June, which also likely affected the lower numbers of users. Without the few most dedicated experts, the amount of data would have been drastically smaller. Reaching out to a large audience could thus be crucial for such projects to find such dedicated experts. In this project, the target audience was rather small, as there are only 14,000 registered birdwatchers in Finland (BirdLife Finland 2022). However, advertising through national and local birding NGOs made it fairly easy to connect with the target audience.

The majority of the negative feedback and suggestions for improvement concerned the malfunctioning of the portal. The main reason for such malfunctioning was that some users had operational systems in which we had not tested the portal properly before it was launched. The portal was coded on Windows operational system, and despite being in theory compatible also for iOS, some malfunctions occurred occasionally, and especially on iOS. Some of the improvement suggestions would have demanded massive changes to the functional structure of the portal and thus could not be fulfilled. The overarching theme in these was the increase for the users' freedom of choice. Providing variability in tasks in a similar project resulted in participants increasingly engaged with the project (Snyder et al. 2022). This could have been done by allowing the users to choose the species of the candidates they were to classify. This would have also provided information for the users about which species are included in the templates, and thus also the expected extent of the study outcomes. In addition, allowing users to classify clips without first needing to classify candidates for 20 templates would have increased the freedom of choice. This might have increased the number of classified clips, since some users saw the classification of candidates as highly time consuming and not highly rewarding, and consequently did not finish 20 templates. Therefore, these users did not have the opportunity to classify clips, and it remains unclear how many clips would have these users eventually classified and if they would have classified as many templates as they did if they had also had the opportunity to classify clips simultaneously.

Given the small amount of data needed for fine-tuning the classification models to local conditions (Lauha et al. 2022), the sufficient number of candidates for each template could have been smaller, which would have decreased the laboriousness of this task. Replacing the sections of *Identify letters* and *Identify recordings* with a single section in which users could classify clips and use drawing tools to annotate the representative sounds with bounding boxes for each species could have increased the freedom of choice and simultaneously produced the required training and testing material for classification model development.

A clear flaw in our planning of the portal was that it did not give a time stamp for the classification, which would have allowed us to study how the different advertisement methods affected the number of classifications, and how much the number of visitors explained the actual number of classifications. It would have also provided a tool to monitor the laboriousness of classifying templates among all users by studying the variation of time needed for completing a set of 100 candidates. Furthermore, to achieve a better picture of how different advertising methods affected both visitors and classifications, an *a priori* study on advertising could have been set up.

CONCLUSIONS

Crowdsourcing was effective for generating a large amount of high-quality data in this project. The success of the project was largely due to experienced users, some of whom were highly dedicated to the project, providing massive amounts of data. In this project, we generated training data that was useful for parameterizing much improved bird classification models for Southern Finland (Lauha et al. 2022), but equally importantly, we learned lessons for how to better implement similar citizen science projects in the future. In particular, we have recently launched the *Bird Sounds Global* portal (https://bsg.laji.fi/) for generating global training data to be used to parameterize machine learning models for classifying birds from autonomous audio recordings. In *Bird Sounds Global*, we have taken into account the feedback from the Finnish pilot project reported here by combining the two separate steps of template and clip annotation into annotation of clips only. We have furthermore given the user the opportunity to not only list the species, but also draw bounding boxes around the vocalizations, which we hope will make the system even more intuitive and attractive for the users.

DATA ACCESSIBILITY STATEMENT

To promote the open use of the data generated by this citizen science campaign, we published the data in Zenodo;

https://zenodo.org/record/7030863#.Y6GgtIRBwuU, Lauha et al. 2022). There are no universal projects gathering fully annotated data, although such repositories have been identified important for future development of passive acoustic monitoring (Gibb et al. 2019). This includes those one-minute audio segments from which training data were generated either for templates or for clips, and text files that describe the annotation units (as bounding boxes on time-frequency space), as well as the annotations made by the users, together with relevant metadata such as the expertise levels of the users. The user identities were anonymized in the data publication.

SUPPLEMENTARY FILES

The supplementary files for this article can be found as follows:

- **Supplemental File 1.** Table 1. List of observed bird species along the routes. DOI: https://doi.org/10.5334/cstp.556.s1
- **Supplemental File 2.** Detailed description of annotation units and allocation to users. DOI: https://doi.org/10.5334/cstp.556.s2

ETHICS AND CONSENT

The research conducted does not require acceptance of the human ethics committee in Finland. The users were required to accept the privacy policy of The Finnish Biodiversity Information Facility (FinBIF) (https://laji.fi/privacypolicy) when registering and also agree to the use and publication of their anonymized classifications in this study.

ACKNOWLEDGEMENTS

FUNDING INFORMATION

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

UC collected all the audio data. AA and PLe built-up the templates, for which OO coded a platform in MatLab. All participated in the design of the web portal; MR constructed the portal, PLe did the advertising, and PLa analyzed the data consistency. All contributed to the writing of the manuscript.

AUTHOR AFFILIATIONS

Petteri Lehikoinen orcid.org/0000-0002-2272-024X
Finnish Museum of Natural History, University of Helsinki, Finland; Research Programme in Organismal & Evolutionary Biology, Faculty of Biological and Environmental Sciences, University of Helsinki, FI

Meeri Rannisto orcid.org/0000-0001-7026-3680
Finnish Museum of Natural History, University of Helsinki, FI

Ulisses Camargo orcid.org/0000-0002-3718-3344
Finnish Museum of Natural History, University of Helsinki, FI

Aki Aintila
Research Programme in Organismal & Evolutionary Biology, Faculty of Biological and Environmental Sciences, University of Helsinki, FI

Patrik Lauha orcid.org/0000-0001-7204-8255
Research Programme in Organismal & Evolutionary Biology, Faculty of Biological and Environmental Sciences, University of Helsinki, FI

Esko Piirainen
Finnish Museum of Natural History, University of Helsinki, FI

Panu Somervuo orcid.org/0000-0003-3121-4047
Research Programme in Organismal & Evolutionary Biology, Faculty of Biological and Environmental Sciences, University of Helsinki, FI

Otso Ovaskainen orcid.org/0000-0001-9750-4421
Department of Biological and Environmental Science, University of Jyväskylä, Finland; Research Programme in Organismal & Evolutionary Biology, Faculty of Biological and Environmental Sciences, University of Helsinki, Finland; Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim N-7491, Norway

REFERENCES

BirdLife Finland. 2022. Liity jäseneksi. Available at: https://www.birdlife.fi/liitytaitue/liity/. [Last accessed 24 August 2022].

Bonney, R, Shirk, JL, Phillips, TB, Wiggins, A, Ballard, HL, Miller-Rushing, AJ and **Parrish, JK.** 2014. Next Steps for Citizen Science. *Science*, 343(6178): 1436–1437. DOI: https://doi.org/10.1126/science.1251554

Callaghan, CT, Poore, AGB, Mesaglio, T, Moles, AT, Nakagawa, S, Roberts, C, Cornwell, WK, et al. 2021. Three Frontiers for the Future of Biodiversity Research Using Citizen Science Data. *BioScience,* 71(1): 55–63. DOI: https://doi.org/10.1093/biosci/biaa131

Camargo, U, Roslin, T and **Ovaskainen, O.** 2019. Spatio-temporal scaling of biodiversity in acoustic tropical bird communities. *Ecography,* 42(11): 1936–1947. DOI: https://doi.org/10.1111/ecog.04544

Constantine, M and **The Sound Approach.** 2006. *The Sound Approach to birding*. Dorset, UK: The Sound Approach.

Cornell Lab of Ornithology. 2022. Macaulay Library. Available at https://www.macaulaylibrary.org/. [Last accessed 24 August 2022].

Farina, A, Pieretti, N and **Piccioli, L.** 2011. The soundscape methodology for long-term bird monitoring: A Mediterranean Europe case-study. *Ecological Informatics,* 6(6): 354–363. DOI: https://doi.org/10.1016/j.ecoinf.2011.07.004

Franzoni, C, Poetz, M and **Sauermann, H.** 2021. Crowds, citizens, and science: a multi-dimensional framework and agenda for future research. *Industry and Innovation*, 1–34. DOI: https://doi.org/10.1080/13662716.2021.1976627

Franzoni, C and **Sauermann, H.** 2014. Crowd science: The organization of scientific research in open collaborative projects. *Research Policy,* 43(1): 1–20. DOI: https://doi.org/10.1016/j.respol.2013.07.005

Gibb, R, Browning, E, Glover-Kapfer, P & **Jones, KE.** 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2): 169–185. DOI: https://doi.org/10.1111/2041-210X.13101

Hill, AP, Prince, P, Piña Covarrubias, E, Doncaster, CP, Snaddon, JL and **Rogers, A.** 2018. AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution,* 9(5): 1199–1211. DOI: https://doi.org/10.1111/2041-210X.12955

Kahl, S, Wood, CM, Eibl, M and **Klinck, H.** 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics,* 61: 101236. DOI: https://doi.org/10.1016/j.ecoinf.2021.101236

Krause, B and **Farina, A.** 2016. Using ecoacoustic methods to survey the impacts of climate change on biodiversity. *Biological Conservation,* 195: 245–254. DOI: https://doi.org/10.1016/j.biocon.2016.01.013

Land-Zandstra, A, Agnello, G and **Gültekin, YS.** 2021. Participants in Citizen Science. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The Science of Citizen Science* (pp. 243–259). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-58278-4_13

Lauha, P, Somervuo, P, Lehikoinen, P, Seibold, S, Geres, L, Richter, T and **Ovaskainen, O.** 2022. Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods in Ecology and Evolution,* 13: 2799–2810. DOI: https://doi.org/10.1111/2041-210X.14003

LeBien, J, Zhong, M, Campos-Cerqueira, M, Velev, JP, Dodhia, R, Ferres, JL and **Aide, TM.** 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, 59: 101113. DOI: https://doi.org/10.1016/j.ecoinf.2020.101113

Lintott, C, Schawinski, K, Bamford, S, Slosar, A, Land, K, Thomas, D, Vandenberg, J, et al. 2011. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies*. *Monthly Notices of the Royal Astronomical Society*, 410(1): 166–178. DOI: https://doi.org/10.1111/j.1365-2966.2010.17432.x

Lintott, CJ, Schawinski, K, Slosar, A, Land, K, Bamford, S, Thomas, D, Vandenberg, J, et al. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey*. *Monthly Notices of the Royal Astronomical Society*, 389(3): 1179–1189. DOI: https://doi.org/10.1111/j.1365-2966.2008.13689.x

Loiselle, BA, Howell, CA, Graham, CH, Goerck, JM, Brooks, T, Smith, KG and **Williams, PH.** 2003. Avoiding Pitfalls of Using Species Distribution Models in Conservation Planning. *Conservation Biology,* 17(6): 1591–1600. DOI: https://doi.org/10.1111/j.1523-1739.2003.00233.x

Mengersen, K, Peterson, EE, Clifford, S, Ye, N, Kim, J, Bednarz, T, Hunter, V, et al. 2017. Modelling imperfect presence data obtained by citizen science. *Environmetrics,* 28(5): e2446. DOI: https://doi.org/10.1002/env.2446

Ovaskainen, O, Moliterno de Camargo, U and **Somervuo, P.** 2018. Animal Sound Identifier (ASI): software for automated identification of vocal animals. *Ecology Letters*, 21(8): 1244–1254. DOI: https://doi.org/10.1111/ele.13092

Papadopoulos, T, Roberts, S and **Willis, K.** 2015. Detecting bird sound in unknown acoustic background using crowdsourced training data. *arXiv preprint arXiv:1505.06443.*

Ruff, ZJ, Lesmeister, DB, Duchac, LS, Padmaraju, BK and **Sullivan, CM.** 2020. Automated identification of avian vocalizations with deep convolutional neural networks. *Remote Sensing in Ecology and Conservation,* 6(1): 79–92. DOI: https://doi.org/10.1002/rse2.125

Salamon, J and **Bello, JP.** 2017. Deep convolutional neural networks and data augmentation for environmental sound

classification. *IEEE Signal processing letters,* 24(3): 279–283. DOI: https://doi.org/10.1109/LSP.2017.2657381

Snyder, R, Clark, M, Salas, L, Schackwitz, W, Leland, D, Stephens, T, Clas, K, et al. 2022. The Soundscapes to Landscapes Project: Development of a Bioacoustics-Based Monitoring Workflow with Multiple Citizen Scientist Contributions. *Citizen Science: Theory and Practice*, 7(1): 24. DOI: https://doi.org/10.5334/cstp.391

Stevenson, RD, Suomela, T, Kim, H and **He, Y.** 2021. Seven Primary Data Types in Citizen Science Determine Data Quality Requirements and Methods. *Frontiers in Climate*, 3. DOI: https://doi.org/10.3389/fclim.2021.645120

Stowell, D, Wood, MD, Pamuła, H, Stylianou, Y and **Glotin, H.** 2019. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods in Ecology and Evolution,* 10(3): 368–380. DOI: https://doi.org/10.1111/2041-210X.13103

Van Brussel, S and **Huyse, H.** 2019. Citizen science on speed? Realising the triple objective of scientific rigour, policy influence and deep citizen engagement in a large-scale citizen science project on ambient air quality in Antwerp. *Journal of Environmental Planning and Management,* 62(3): 534–551. DOI: https://doi.org/10.1080/09640568.2018.1428183

Walmsley, M, Smith, L, Lintott, C, Gal, Y, Bamford, S, Dickinson, H, Wright, D, et al. 2020. Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society*, 491(2): 1554–1574. DOI: https://doi.org/10.1093/mnras/stz2816

Warblr Ltd. 2022. Warblr: Identify UK bird songs. 25.1.2022. Available at https://www.warblr.co.uk/. [Last accessed 24 August 2022].

Xeno-canto Foundation. 2022. Xeno-Canto – sharing bird sounds from around the world. 25.1.2022. Available at https://xeno-canto.org/. [Last accessed 24 August 2022].