Open-Set Classification Approaches to Automatic Bird Song Identification: Towards Non-Invasive Wildlife Monitoring in Brazilian Fauna

Tiago Fernandes Tavares

***Abstract*—Bird song identification has mainly been approached as a closed-set classification problem; that is, all samples are known to be from one of the classes known by the classifier. However, wildlife monitoring using bird songs is closer to an open-set classification setting, as the classifier is required to predict if a sample comes from an unknown origin, like an environmental sound or an unrelated animal. Furthermore, current approaches to bird song classification assume that the model can access the whole dataset and build optimal projections. This is not a realistic scenario in Brazil as the country has thousands of species, and it is unfeasible to build a dataset containing a representative diversity of samples of all of them. This work analyzes algorithms that can be used for the open-set classification of bird songs. The analyzed algorithms can fit models using data from one or from only a few species. The investigation revealed many current technical difficulties and highlighted several opportunities for future work in this field.**

***Index Terms*—Wildlife monitoring, open-set classification, bird song identification, Brazilian fauna.**

I. Introduction

Parallel to the unquestionable increase in important development aspects such as life expectancy [1] and literacy rates [2] all over the world in the last few centuries, the Earth environment has also suffered from diverse consequences of unplanned human action [3]. The need to maintain the human development achieved so far and simultaneously mitigate the unwanted environmental consequences calls for planned, sustainable development endeavors in the future [4] [5].

Sustainable development requires both anticipating and measuring the diverse forms of human impact [4], in special while growing agricultural areas [6]. This can be performed using weather stations or sensor systems, which are able to detect changes in the physical or chemical properties of a particular environment. However, these sensors usually have technical problems, like being slow (as weather, for example, takes years to change in response to human action) [7] or unable to reach a great area (such as electronic detectors, which can only reach their immediate surroundings) [8].

A possible solution to avoid these problems is to use animals as bio-sensors to detect pollution in greater areas [9], which is particularly interesting in countries with large biodiversity like Brazil. This proposal's underlying idea is that animals tend to disappear from particular regions if they sense threats, which might be linked to invasive human presence [10]. Hence, the presence or absence of particular animal species within an area can be indicative of that area's level of pollution [9].

Monitoring wildlife, however, can be difficult because animals can commonly camouflage well among the vegetation [11], thus creating a need to either identify typical animal paths or to perform expeditions to manually count them [12]. However, animals can be identified by automated devices using computer vision on a camera feed or machine listening on a microphone feed [11]. Between these two solutions, computer vision tends to fail more often because it requires direct animal sight, which is often unfeasible, whereas machine listening can deal with omnidirectional acoustic data, which can potentially reach larger distances [13].

There are many animals that use acoustic communication. In the terrestrial fauna, typical examples are insects (like cicadas) [14], amphibians (like frogs and toads) [15], and birds. The latter category has received a great deal of attention from the scientific community in the last few years, and there are sources that make bird sounds available for research.

Such sources have been used in many works that aim to automatically identify bird species solely from their sounds [13], [16]–[18]. They have used a diversity of datasets, and the results are promising. One characteristic of all current methods is that they assume that bird species identification is a closed-set classification problem.

In a closed-set classification problem, all elements yielded to the classifier in the evaluation stage are known to have been presented during the fitting stage. This is the case in human-interactive apps or in software aimed at scientists because in these applications, a user selects what sounds should be classified, and is supposed to know, *a priori*, that a particular sound was emitted by a bird. However, this is not the case in environmental monitoring, as in this application, the classifier is continuously exposed to a diversity of unknown sounds.

This scenario, shown in Fig. 1, is known as an *open-set classification* problem [19]. Open-set classification problems have shown to be much harder than their closed-set counterparts but are closer to the reality of many applications. In open-set problems, the classifier tries to associate each input to a known class and, if necessary, it can label the input as "unknown", indicating that the input probably does not belong to any of the classes learned during fitting.

The difficulty in open-set classification lies in the fact that unknown samples can be different from the known ones in many different ways. Also, unknown classes are not presented

T. F. Tavares is with the Institute for Teaching and Research – INSPER – Rua Quatá, 300, Vila Olímpia, São Paulo-SP, CEP 04546-042 e-mail:tiagoft1@insper.edu.br

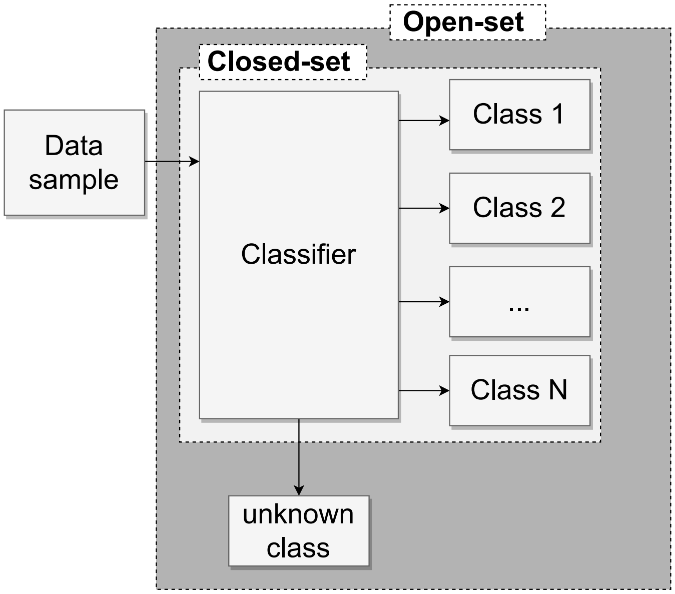


Fig. 1. Closed-set and open-set classification problems. Open-set classifiers are able to identify if an input sample does not belong to any of the known classes.

to the classifier during fitting. Consequently, it is hard to find adequate decision boundaries for the unknown classes.

Open-set classification for bird sounds is especially relevant in Brazil because it is unfeasible to gather a uniform, balanced dataset representative of each of the more than 1800 described species that live within the country. Online databases contain recordings of only a fraction of these species, recorded with different equipment, in different specific environmental conditions. This can lead classifiers to mistake, for example, the sound of a particular river or the distortion of a particular audio recorder with a specific bird's song.

This work investigates the performance of different classification algorithms for open-set bird song classification and compares them with their closed-set counterparts. It also discusses their possible usages in model-sharing ecosystems. Lastly, this work brings forward how open-set classification allows for a sustainable growth model towards a sustainability tool.

II. Related Work

Automatic bird song identification has been researched at least since the 1990s [20]. Until the last five years, most approaches for this task consisted of extracting features from labeled audio files and then using classification algorithms in standard machine learning pipelines [21]. Research work in this field has used a myriad of different datasets and techniques to build data-driven classifiers, and results were compared using standard metrics such as average class accuracy, recall, precision and f1-score.

More recently, deep-learning techniques have been obtaining state-of-the-art results in many classification problems for many years now [22]. However, commercial software for bird song classification still mainly relies on standard techniques such as Hidden Markov Models (HMMs) and handcrafted digital signal processing techniques for noise removal [23]. Deep-learning has been avoided in commercial software because its performance quickly decreases when the number of species in the dataset grows [21], [24], and, furthermore, it acts as a black box, making it impossible to interpret the reasons underlying its predictions, hence harming its use in real environments [24].

This non-interpretability characteristic of more complicated models make its comparison restricted to standard metrics. This means that, although some systems can reach outstanding metrics, it is not possible to anticipate situations and use cases in which the model is likely to fail. Consequently, further development in the field can generate little insight other than improving metrics in a development set.

Most work on automatic bird song classification assumes that any input sample belongs to a class present in the training set, which can lead to very good classification results [25]. Some works have used either an energy threshold or a special class to detect noise [21]. However, both of these scenarios can lead to systems biased towards noise present in the development set.

This work analyses open-set classification systems that can be entirely explained by humans. For such, it combines using simple classification algorithms and feature projections, and assesses the effects of increasing the number of classes in the dataset. It finds that an underlying problem in bird song classification is finding a suitable representation, that is, future efforts should focus more on feature extraction steps than on more complicated classification procedures.

III. Methods

The classification system follows a typical processing pipeline for audio information retrieval [26]–[28], as shown in Fig. 2. First, each audio file is converted to a feature vector that aims to represent its acoustic content. The feature vectors are split into non-overlapping train and test sets, and the train set is used to estimate the classifier parameters. Then, the feature vectors in the test set are yielded the classifier, which predicts their class labels. In the open-set classification setting, one of the possible class labels is "unknown", meaning that the corresponding input sample of the test set does not correspond to any of the classes present in the train set.

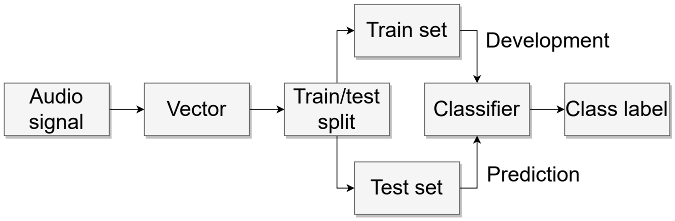


Fig. 2. System overview. Audio signals are converted to vectors which are yielded to classification models. The models generate class labels for each of the inputs.

The dataset that drives this process is described in Section A. The process that converts audio files to their vector representations is discussed in Section B, and the classification algorithms are described in Section D.

A. Dataset

The dataset used in this work consists of audio recordings of birds from Brazil. The dataset was downloaded from the Xenocanto [29] website using webscrapping, comprising all recordings whose location was labeled as "Brazil". The recorded bird's common name was used as a ground truth species label.

In total, the dataset contains data from 1539 species, but, as shown in Fig. 3, most of them have only a few recordings. Using too few recordings for a species can harm the fitting and evaluation process because environmental noises or recording conditions can easily become confounding factors and generate misleading results. To mitigate this situation, only species with more than 50 recordings were used, resulting in a total of 93 unique labels.

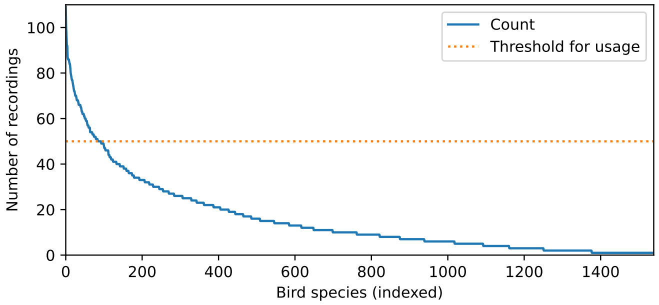


Fig. 3. Number of recordings for each bird species (indexed to improve visualization clarity) in the dataset. Although the dataset is wide regarding the number of species, most of them have only a few recordings. Only species with 50 or more recordings were used in the experiments.

B. Vector Representation for Audio

Currently, there are many techniques that allow mapping audio files to unique vectors, which are applicable to a myriad of situations. More recently, deep learning techniques have been used to derive representations directly from data. In this work, the model-sharing idea imposes important restrictions on the vectorization.

In the model-sharing paradigm, the dataset is, a priori, unknown, except for a few points or classes. Consequently, it makes little sense to use auto-encoders, as they are evidently dependent on data related to all classes of the dataset. Also, this restricts the possibility of feature-wise standardization, as the mean and standard deviation of each feature can greatly change among classes.

Consequently, this scenario calls for a fixed, well-defined feature calculation process. Such process is accomplished by first re-sampling recordings to a sampling frequency of 22050 Hz and then normalizing audio samples to zero mean and unit variance (thus mitigating the effects of recording birds at different distances) and then calculating frame-wise, predefined features in sliding windows of known length (in this work, the window is 2048 samples long and the step between two subsequent windows in 512 samples) and then calculating statistics that measure the feature behaviors through time. After that, feature vectors are independently normalized to unit norm, which brings the absolute value of all elements in the vector to the range $[-1, 1]$.

It is hard to devise feature sets that are suitable discriminators for both known and unknown bird species, but it is reasonable to assume that different bird songs provide different stimuli to the human ear. Because of that, the framewise features calculated for this work are the Mel-Frequency Cepstral Coefficients (MFCCs) from 2 to 12 (that is, 11 coefficients). The first MFCC is excluded because it represents the signal energy, which is obviously non-representative of the signal content as recording conditions can vary.

After calculating the framewise MFCCs, their first and second order differentials are calculated, resulting in 33 framewise features. The first and second order differentials are important because they model sequential dependecies between frames. The features are summarized into a single vector containing the mean and standard deviation of the framewise MFCCs and the standard deviation of the differentials, resuting in a 44-dimension summary vector for each audio file. The mean of the differentials is not used because it accounts for drifts and non-oscillatory behavior in the audio files, which is not expected to occur in samples longer than a few seconds. Their standard deviation, however, is used because it accounts for the variation in each MFCC, which is related to types of variations in each critical band.

Importantly, the feature set cannot be arbitrarily large. This is because there is a very limited amount of recordings available for each species in the dataset, thus using feature vectors with higher dimension can increase the probability that elements unrelated to birds, like environment sounds or equipment filters, become confounding factors in the classification process. For this reason, this work refrains from using much of the existing features in the literature, even if they could potentially lead to greater benchmark results.

C. Experimental Setup

This works compares the results in two different experimental setups: the closed-set experiment, which is similar to most of the literature, and the open-set experiment, which is more rarely explored. These setups are described next.

1) Closed-set Experiment: The most commonly studied scenario in bird species identification is the closed-set classification. In this case, all samples in the test set are known to be drawn from one of the classes that exist in the train set. Although this is not a realistic use case for environmental monitoring, it provides an upper bound to the expected performance in the open-set scenario, and can be used to draw important insight from data.

The closed-set experiments discussed in this section comprise randomly selecting N species from the whole dataset, and then dividing the samples from the selected species into train and test sets. Then, each classification algorithm is trained using data from the train set, and evaluated using normalized accuracy in the test set.

The experiments were performed for increasing values of N, highlighting the effects of using more classes, thus progressively increasing the task's difficulty.

2) Open-set experiment: In the open-set classification experiment, classifiers required to either label an input as belonging to one of the known classes or to label that input as belonging to a "unknown" class. This is clearly harder than the closed-set experiment.

The open-set experiments discussed in this section comprise randomly selecting N species from the whole dataset to form the "known" set, and then dividing samples from these species it into train and test sets. The train set is used to estimate both the classifier parameters and the rejection threshold. In prediction, the algorithms can output "unknown" as a prediction. Each classification algorithm is evaluated using normalized accuracy in the test set.

Also, the classification algorithms are yielded samples from classes that are not part of the "known" set. In this second task, the classifiers are expected to always predict "unknown". Each classification algorithm is evaluated using the normalized accuracy of these predictions. Similarly to the closed-set experiments, the open-set ones were performed using increasing values of N.

D. Classification Algorithms

The classification algorithms used in this work (K-Nearest Neighbors, Naive Bayes, and Support Vector Classifier) are well-known and broadly used in many fields. They were chosen over more complicated models because they can be fully explained using simple principles of linear algebra and statistics. However, their usage in the open-set experiments require some adaptations, as discussed next.

1) K-Nearest Neighbors: The K-Nearest Neighbors classifier (KNN) works by storing all feature vectors v in the training set together with their corresponding class. Then, for prediction, it calculates the Euclidean distance between the input vector o and each of the stored vectors, selects the K ones that are closer to the input and returns the class that appears most frequently among these K nearest neighbors. This work used $K = 5$ based on experiences with other audio datasets.

This simple, cost-effective solution can achieve competitive results. Due to its non-parametric nature, a larger KNN model can be built by simply using the stored vectors of smaller models. Henceforth, a large model for bird detection can be built by simply sharing feature vectors, that is, sharing models instead of sharing raw data.

In a closed-set experiment, the classifier is forced to associate all input samples with a known class. However, the open-set setting requires the model to reject samples, that is, associate samples with the "unknown" class. Because the model relies on Euclidean distances, it is possible to find a rejection threshold l such that, if

$$\min_i ||v_i - o||^2 > l, \tag{1}$$

then the input can be considered too distant from the known classes, hence it can be labeled as belonging to the "unknown" class. The rejection threshold was estimated by first calculating $\min_i ||v_i - o||^2$ for all samples in a validation set, and then using l as the 90-percentile of these values.

2) Naive Bayes: The underlying idea of a Naive Bayes classifier is to use training data do estimate a probability distribution for the observed feature vectors o given their known class C_i, that is, $P(o|C_i)$, and then, in the prediction stage, use Bayes' Theorem to estimate the probability of a sample belonging to a class given its feature vector, that is, the posterior probability $P(C_i|o)$. In a closed-set experiment, the classifier simply predicts the class with higher $P(C_i|o)$, but in an open-set configuration it is necessary to estimate when a sample does not belong to any of the known classes. Assuming that samples from unknown classes present a lower posterior probability in relation to the known classes, sample rejection can be performed using a threshold l such that, if

$$\max_i P(C_i|o) < l, \tag{2}$$

then the sample is assumed to be too different from any known classes, thus its class is unknown.

The threshold l was estimated by calculating the minimum posterior probability for each sample in the training set, then selecting the 10-percentile of that series. This means that, if the train and test sets have the same sample distribution in the feature space, we can expect to wrongly associate samples with unknown classes around 10% of the time.

Because all distributions $P(o|C_i)$ are estimated independently, a different Naive Bayes classifier model can be trained for each class, and then all models can be combined in a multi-class classifier. Similarly to the KNN classifier, this allows for distributed model training, which favors building communities based on sharing models instead of sharing data.

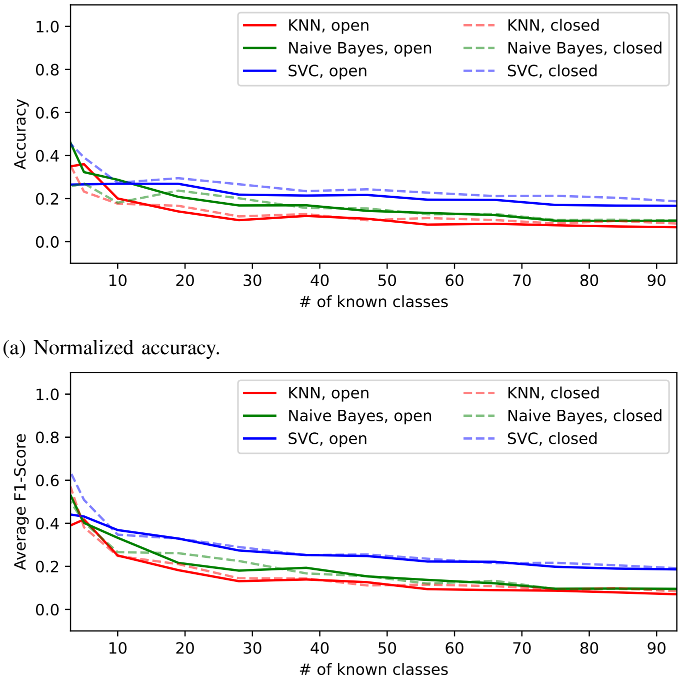
3) Support Vector Classifier: The Support Vector Classifier (SVC) [30] works by combining risk minimization and regularization to find an optimal hyper-plane that separates samples in a high-dimensional projection. The SVC uses less computational resources that modern neural networks, but frequently achieves competitive results. The SVC allows to estimate posterior probabilities relating classes to observed vectors ($P(C_i|o)$) based on normalizing the distance between the observed vector and the hyper-planes. This work used the hyperparameters $\mu = 10^{-7}$ and $C = 100$ based on experiments with other audio datasets.

These probabilities can be used similarly to those discussed in the Naive Bayes classifier (Section D2). Likewise, a threshold can be found using the 10-percentile of the posteriors of the training set, and samples that satisfy Expression 2 are considered to belong to an unknown class.

Differently from the Naive Bayes classifier, the SVC is capable of using the multiple classes in the training set to generate separation boundaries. This commonly leads to better classification results, but makes it necessary to re-train the model's parameters if a new class is added to the model.

IV. Results

Fig. 4 shows the normalized accuracy (that is, the mean of the class-wise accuracy) and the unweighted class-wise average f1-score for the closed-set and the open-set classifiers considering only the known classes. Clearly, the closed-set classifiers have a greater performance, as they do not need to classify samples as "unknown". With the increase of the number of known species, the performances of the closed-set classifiers decrease and become closer to those of their open-set counterparts.



(a) Normalized accuracy.

(b) F1-Score.

Fig. 4. Unweighted class-wise average results for closed- and open-set problems for all tested algorithms.

In general, KNN performed worse, and SVC performs best among these classifiers. This has been observed in other audio-related classification problems. However, the performance difference decreases as the number of classes increase in the experiment.

Fig. 5 shows the accuracy for the samples of unknown classes for each classifier. In this test, only samples from unknown classes were presented to the classifier, hence this rate can be interpreted as a one-class prediction accuracy. Interestingly, this recall rate does not vary significantly when the number of classes increase.

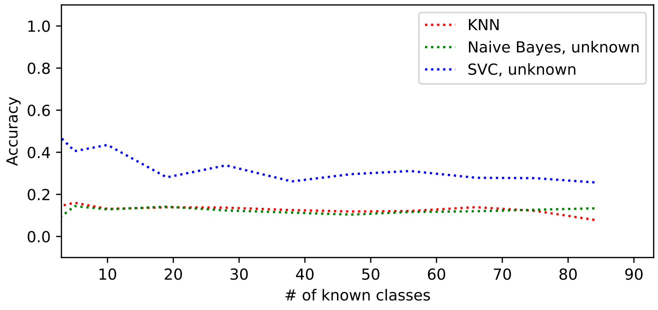


Fig. 5. Classification accuracy for samples of unknown classes.

It is usual to use confusion matrices to detect biases in classifiers, but our dataset has 96 classes, which would make the confusion matrix too large to display and discuss. Because of that, we use the histogram of the class-wise F1-Score, as shown in Fig. 6, as basis for our discussion. As it can be seen, KNN and Naive Bayes tend to present poor performance in a large number of classes, while this tendency decreases for SVC.

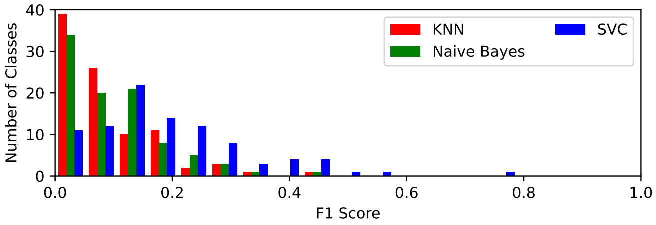


Fig. 6. Histogram of class-wise F1-Scores in the closed-set experiment.

Although classification accuracy is an important benchmark measure, the sample organization in the feature space can give important insight towards future work. Fig. 7 shows a 2D Principal Component Analysis (PCA) projection of the feature space using $N = 5$ classes. It is possible to see that classes do not form specific clusters, which can explain the relatively low classification results.

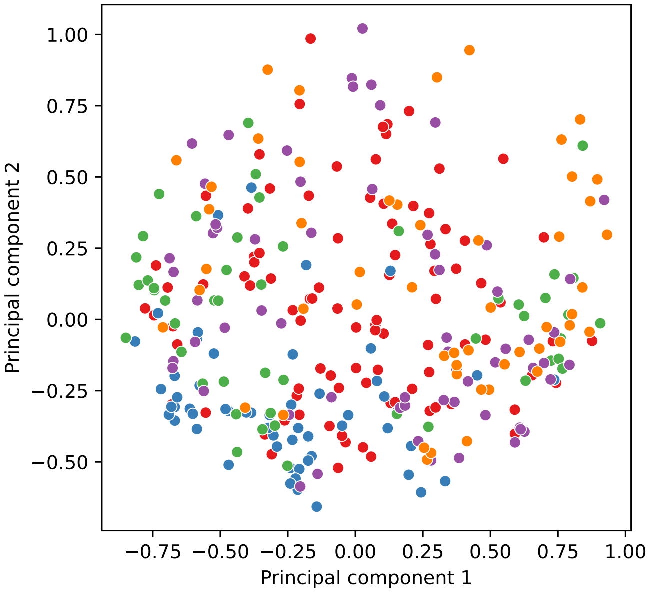


Fig. 7. PCA projection of feature space when using 5 species. The calculated silhouette score was -0.08.

However, the PCA projection can be misleading, as class clusters might be organized in manifolds. This situation can be analyzed using a T-SNE projection [31], as shown in Fig. 8. The T-SNE projection shows that the data seems to form small clusters that are primarily populated with samples from one or two classes.

Both PCA and T-SNE projections show a high overlap between classes, agreeing with their low silhouette scores,

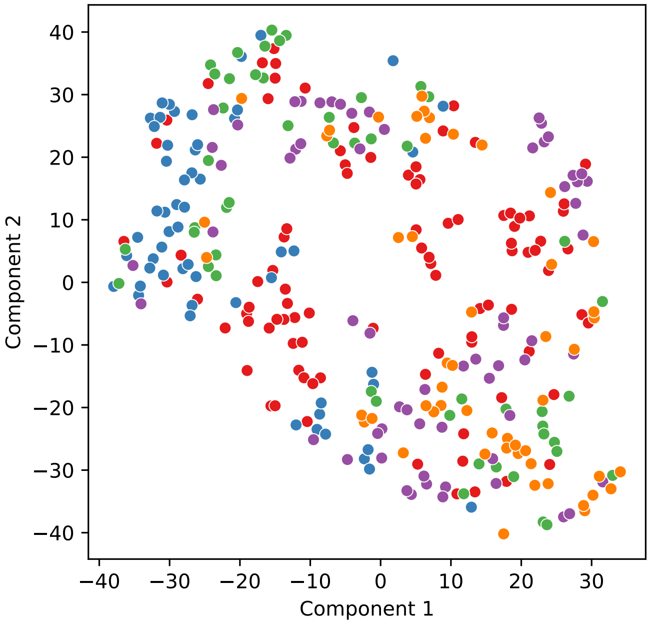


Fig. 8. T-SNE projection of feature space when using 5 species. The calculated silhouette score was -0.07.

which indicates the dataset is hard to classify. However, not all classes overlap with all others. The projections indicate, for example, a stronger separation between blue and orange points than between purple and green points.

This highlights a property that emerges jointly from the audio files in the development set and their representation: some species are more easily identified than others. This corroborates with the results shown in Fig. 6, which clearly display some classes having higher F1-Scores for the SVC classifier. Importantly, this is a characteristic of the features that represent audio files, hence other feature sets could lead to different results, meaning that results are highly dependent on the specific species that are present in the dataset, and consequently, building classification systems requires analyzing each specific bird song dataset.

V. Discussion

Bird species identification from audio is a known problem and has had several proposed solutions in the last years, but it has been consistently approached as a closed-set problem. Although there is a finite (even if slowly changing due to evolution and extinction) number of existing bird species in the world, it is necessary to have at least some tenths of audio samples related to each species from a particular location to build a dataset in order to be able to account for different recording environments and equipments. As shown in Fig. 3, Xenocanto, one of the largest bird sound archives in the world, cannot provide such amount of material for a large country such as Brazil.

Importantly, all data used in this work comprises birds from Brazil. This means that results are only applicable to Brazilian birds. However, the applied techniques do not rely on any assumptions that are exclusive to Brazilian birds, henceforth they can be tested in songs acquired in other locations.

A characteristic of current datasets that harms the development of wildlife monitoring systems is that they focus on labeling samples according to bird species, but have little to no metadata regarding the types of noises also present in the recording. It is likely that some types of noises are more harmful towards identifying particular bird species, that is, some are more harmed by city noises, others are more likely to be confused with frogs, and so on. Such fine-grained evaluation can be important to anticipate system failures if bird identifiers are used in automatic, open-air settings.

Regardless of the data availability, there are several technical challenges that can already be tackled. As shown in Section IV, one of the most important ones is to find feature spaces that lead to a better separation of species without requiring data from all (or several) species. This would allow changing MFCCs, which are inspired in the human hearing system, with other features that could be more closely related to the bird sound production process.

It is important to highlight that the output probabilities and distances yielded by the Naive Bayes and the KNN could be calibrated using Pratt scaling. However, because Pratt scaling consists of fitting data to a monotonic curve, the percentiles would remain the same, which means that the results would not change. Nevertheless, Pratt scaling could lead to a more elegant mathematical formulation, and could be used in future implementations.

Last, it could be relevant to further study algorithms for the open-set classification problem. There are some more recent proposals that could inspire interesting solutions for this problem, such as the Siamese networks and their capability to learn manifolds from data, or using transfer learning to devise features from data in other domains. These can be interesting challenges for future work.

VI. Conclusion

This work investigates the open-set classification problem of identifying bird species from audio recordings of their songs, which is related to wildlife monitoring applications. It also investigates the constraint that models for identifying particular species can be created without access to a large amount of data. This technical constraint is necessary for bird identification models that can be used in Brazilian wildlife monitoring in a short term, as building a larger dataset is a task that can take many years.

Results indicate that devising features more related to bird sound production processes could lead to improved accuracy, even if the used classifiers are not changed. Also, the evaluation methodology can be improved by assessing the impacts of using different environmental sounds as inputs to the classifier. Last, future work could comprise using modern classification algorithms.

The open-set classification problem for bird song identification enabling automatic wildlife monitoring is yet far from being completely solved. This work proposes a step towards this solutions, and has highlighted some of the technical challenges that appear with these constrains. They can be tackled in the future, in further steps towards using digital signal processing and machine learning to aid in sustainable development.

References

[1] World Bank, "Life expectancy at birth, total (years) | Data." [Online]. Available: https://data.worldbank.org/indicator/SP.DYN.LE00.IN

[2] ——, "Literacy rate, adult total (% of people ages 15 and above) | Data." [Online]. Available: https://data.worldbank.org/indicator/SE.ADT.LITR.ZS

[3] H. Yang, M. Ma, J. R. Thompson, and R. J. Flower, "Waste management, informal recycling, environmental pollution and public health," *Journal of Epidemiology and Community Health*, vol. 72, no. 3, pp. 237–243, Mar. 2018. [Online]. Available: https://jech.bmj.com/lookup/doi/10.1136/jech-2016-208597

[4] S. Fawzy, A. I. Osman, J. Doran, and D. W. Rooney, "Strategies for mitigation of climate change: a review," *Environmental Chemistry Letters*, vol. 18, no. 6, pp. 2069–2094, Nov. 2020. [Online]. Available: https://link.springer.com/10.1007/s10311-020-01059-w

[5] G. Plumecocq, T. Debril, M. Duru, M.-B. Magrini, J. P. Sarthou, and O. Therond, "The plurality of values in sustainable agriculture models: diverse lock-in and coevolution patterns," *Ecology and Society*, vol. 23, no. 1, 2018. [Online]. Available: https://www.jstor.org/stable/26799066

[6] L. A. Martinelli, R. Naylor, P. M. Vitousek, and P. Moutinho, "Agriculture in brazil: impacts, costs, and opportunities for a sustainable future," *Current Opinion in Environmental Sustainability*, vol. 2, no. 5-6, pp. 431–438, Dec. 2010. [Online]. Available: https://doi.org/10.1016/j.cosust.2010.09.008

[7] C. D. Kolstad and F. C. Moore, "Estimating the Economic Impacts of Climate Change Using Weather Observations," *Review of Environmental Economics and Policy*, vol. 14, no. 1, pp. 1–24, Jan. 2020. [Online]. Available: https://www.journals.uchicago.edu/doi/10.1093/reep/rez024

[8] Y. Zhang, Y. Zhu, Z. Zeng, G. Zeng, R. Xiao, Y. Wang, Y. Hu, L. Tang, and C. Feng, "Sensors for the environmental pollutant detection: Are we already there?" *Coordination Chemistry Reviews*, vol. 431, p. 213681, Mar. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010854520307438

[9] M. A. A. Salahuddin, I. S. Rohayani, and D. A. Candri, vol. 913, no. 1, p. 012058, nov 2021. [Online]. Available: https://doi.org/10.1088/1755-1315/913/1/012058

[10] J. dos Santos Cerqueira, H. N. de Albuquerque, and F. de Assis Salviano de Sousa, "Impact of the functioning of a thermeletry in the bird fauna of the brazilian semiarid," *Revista Ibero-Americana de Ciências Ambientais*, vol. 9, no. 2, pp. 71–83, Sep. 2017. [Online]. Available: https://doi.org/10.6008/cbpc2179-6858.2018.002.0007

[11] R. Gula, J. Theuerkauf, S. Rouys, and A. Legault, "An audio/video surveillance system for wildlife," *European Journal of Wildlife Research*, vol. 56, no. 5, pp. 803–807, Oct. 2010. [Online]. Available: https://doi.org/10.1007/s10344-010-0392-y

[12] D. P. Munari, C. Keller, and E. M. Venticinque, "An evaluation of field techniques for monitoring terrestrial mammal populations in amazonia," *Mammalian Biology*, vol. 76, no. 4, pp. 401–408, Jul. 2011. [Online]. Available: https://doi.org/10.1016/j.mambio.2011.02.007

[13] R. Shrestha, C. Glackin, J. Wall, and N. Cannings, "Bird Audio Diarization with Faster R-CNN," in *Artificial Neural Networks and Machine Learning – ICANN 2021*, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds. Cham: Springer International Publishing, 2021, vol. 12891, pp. 415–426.

[14] J. SUEUR, "Cicada acoustic communication: potential sound partitioning in a multispecies community from Mexico (Hemiptera: Cicadomorpha: Cicadidae)," *Biological Journal of the Linnean Society*, vol. 75, no. 3, pp. 379–394, 10 2008. [Online]. Available: https://doi.org/10.1046/j.1095-8312.2002.00030.x

[15] R. S. Schmidt, "Central Mechanisms of Frog Galling," *American Zoologist*, vol. 13, no. 4, pp. 1169–1177, Nov. 1973. [Online]. Available: https://academic.oup.com/icb/article-lookup/doi/10.1093/icb/13.4.1169

[16] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *The Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 1209–1219, Aug. 1996. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.415968

[17] M. T. Lopes, L. L. Gioppo, T. T. Higushi, C. A. Kaestner, C. N. Silla Jr., and A. L. Koerich, "Automatic bird species identification for large number of species," in *2011 IEEE International Symposium on Multimedia*, 2011, pp. 117–122.

[18] R. H. Zottesso, Y. M. Costa, D. Bertolini, and L. E. Oliveira, "Bird species identification using spectrogram and dissimilarity approach," *Ecological Informatics*, vol. 48, pp. 187–197, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954118300888

[19] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.

[20] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *The Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 1209–1219, Aug. 1996. [Online]. Available: https://doi.org/10.1121/1.415968

[21] N. Priyadarshani, S. Marsland, and I. Castro, "Automated birdsong recognition in complex acoustic environments: a review," *Journal of Avian Biology*, vol. 49, no. 5, pp. jav–01447, May 2018. [Online]. Available: https://doi.org/10.1111/jav.01447

[22] W. Yu, K. Yang, Y. Bai, T. Xiao, H. Yao, and Y. Rui, "Visualizing and comparing alexnet and vgg using deconvolutional layers," in *Proceedings of the 33 rd International Conference on Machine Learning*, 2016.

[23] J. Marchal, F. Fabianek, and Y. Aubry, "Software performance for the automated identification of bird vocalisations: the case of two closely related species," *Bioacoustics*, vol. 31, no. 4, pp. 397–413, Jul. 2021. [Online]. Available: https://doi.org/10.1080/09524622.2021.1945952

[24] X. Dong and J. Jia, "Advances in automatic bird species recognition from environmental audio," *Journal of Physics: Conference Series*, vol. 1544, no. 1, p. 012110, May 2020. [Online]. Available: https://doi.org/10.1088/1742-6596/1544/1/012110

[25] R. A. Bistel, A. Martinez, and G. B. Mindlin, "Neural networks that locate and identify birds through their songs," *The European Physical Journal Special Topics*, vol. 231, no. 3, pp. 185–194, Dec. 2021. [Online]. Available: https://doi.org/10.1140/epjs/s11734-021-00405-5

[26] S. Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 619–625, 2000.

[27] X.-l. Li, Z.-l. Du, and Y.-f. Zhang, "Kernel-based audio classification," in *2006 International Conference on Machine Learning and Cybernetics*, 2006, pp. 3313–3316.

[28] Y. Zhu, Z. Ming, and Q. Huang, "Automatic audio genre classification based on support vector machine," in *Third International Conference on Natural Computation (ICNC 2007)*, vol. 1, 2007, pp. 517–521.

[29] Xenocanto Foundation, "Xenocanto," https://xeno-canto.org/.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: https://doi.org/10.1007/bf00994018

[31] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

Tiago F. Tavares has a Bachelor's degree (2008) in Computer Engineering from the University of Campinas (Unicamp). He has a MsC degree in Electrical Engineering (2010) from Unicamp with a focus on Digital Signal Processing. He has a PhD degree in Electrical Engineering (2013) from Unicamp with a focus on Machine Learning, with a sandwich internship at the University of Victoria (Canada) in 2011-2012. He has worked as a post-doctoral fellow at the Interdisciplinary Nucleus for Sound Studies (NICS-Unicamp) from 2013 to 2015, and from 2015 to 2021 he was assistant professor at Unicamp. He has been with Insper since 2022.