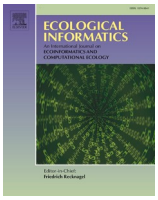
Ecological Informatics 83 (2024) 102823

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Acoustic fingerprints in nature: A self-supervised learning approach for ecosystem activity monitoring

Dario Dematties [a,b,*], Samir Rajani [a], Rajesh Sankaran [a,b], Sean Shahkarami [a,b], Bhupendra Raut [a,c], Scott Collis [a,c], Pete Beckman [a,b], Nicola Ferrier [a,b]

[a] Northwestern Argonne Institute of Science and Engineering, Northwestern University, Hogan Biological Sciences Bldg., 2205 Tech Drive, First Floor, Suite 1160, Evanston 60208, IL, USA
[b] Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Lemont 60439, IL, USA
[c] Environmental Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Lemont 60439, IL, USA

ARTICLE INFO

2000 MSC:
0000
1111

PACS:
0000
1111

Keywords:
Self-supervised learning
Biodiversity
Edge computing
Avian diversity monitoring
Deep learning
Joint embedding

ABSTRACT

According to the World Health Organization, *healthy communities rely on well-functioning ecosystems*. Clean air, fresh water, and nutritious food are inextricably linked to ecosystem health. Changes in biological activity convey important information about ecosystem dynamics, and understanding such changes is crucial for the survival of our species. Scientific edge cyberinfrastructures collect distributed data and process it in situ, often using machine learning algorithms. Most current machine learning algorithms deployed on edge cyberinfrastructures, however, are trained on data that does not accurately represent the real stream of data collected at the edge. In this work we explore the applicability of two new self-supervised learning algorithms for characterizing an insufficiently curated, imbalanced, and unlabeled dataset collected by using a set of nine microphones at different locations at the Morton Arboretum, an internationally recognized tree-focused botanical garden and research center in Lisle, IL. Our implementations showed completely autonomous characterization capabilities, such as the separation of spectrograms by recording location, month, week, and hour of the day. The models also showed the ability to discriminate spectrograms by biological and atmospheric activity, including rain, insects, and bird activity, in a completely unsupervised fashion. We validated our findings using a supervised deep learning approach and with a dataset labeled by experts, confirming competitive performance in several features. Toward explainability of our self-supervised learning approach, we used acoustic indices and false color spectrograms, showing that the topology and orientation of the clouds of points in the output space over a 24-h period are strongly linked to the unfolding of biological activity. Our findings show that self-supervised learning has the potential to learn from and process data collected at the edge, characterizing it with minimal human intervention. We believe that further research is crucial to extending this approach for complete autonomous characterization of raw data collected on distributed sensors at the edge.

1. Introduction

Understanding ecosystems is critical for human survival, as they provide vital resources and regulate Earth's processes (Cianfagna et al., 2021; Marselle et al., 2021). Ecosystem degradation can lead to reduced agricultural productivity and increased risks of natural disasters (AbdelRahman, 2023; Gomiero, 2016; Kato and Huang, 2021; Paz et al., 2020; Walz et al., 2021; Wickramasinghe, 2021). Biodiversity loss is an indicator of declining ecosystem health (Biodiversity and Ecosystem Stability | Learn Science at Scitable, n.d.; Ashford et al., 2021).

Recent advancements in environmental monitoring (Stephenson, 2020) have led to the development of Sage, a software-defined sensor network for artificial intelligence (AI)-enabled edge computation (Beckman et al., 2019; Catlett et al., 2019; Catlett et al., 2022). This platform enables continuous ecosystem monitoring and biodiversity characterization by processing full-resolution sensor data in situ using efficient, low-power computation at the edge.

AI, particularly deep learning methods, allows for the automated

* Corresponding author at: Northwestern Argonne Institute of Science and Engineering, Northwestern University, Hogan Biological Sciences Bldg., 2205 Tech Drive, First Floor, Suite 1160, Evanston 60208, IL, USA.
E-mail address: dario.dematties@northwestern.edu (D. Dematties).

https://doi.org/10.1016/j.ecoinf.2024.102823
Received 25 July 2023; Received in revised form 9 September 2024; Accepted 9 September 2024
Available online 20 September 2024

detection and classification of environmental data (Abeßer, 2020; Kahl et al., 2021). However, AI models deployed at the edge often use curated datasets for training, which may not reflect real-world conditions (Al-Atat et al., 2023; Fan et al., 2023; Hao et al., 2023; Lin et al., 2022; Murshed et al., 2021; Wu et al., 2019). Additionally, the statistical distribution of live data streams can change over time (Ackerman et al., 2022), further complicating training and inference.

While numerous studies have utilized supervised learning for audio classification, self-supervised learning (SSL) approaches have been limited. Recent machine learning strategies have demonstrated the separation of soundscape components using various clustering methods (McGinn et al., 2023; Michaud et al., 2023; Morales et al., 2022; Morita et al., 2022; Rowe et al., 2021; Sun et al., 2022; Thomas et al., 2021). These techniques range from neighborhood-based dimensionality reduction based on spectrograms (Thomas et al., 2021) to combinations of unsupervised and supervised deep learning (Michaud et al., 2023; Morales et al., 2022).

Our approach differs from previous research in several key aspects. We explore learning from datasets that are neither well-curated nor balanced, mimicking real-world edge computing scenarios. We employ two SSL approaches under the joint embedding umbrella (self-DIstillation with NO labels (DINO) (Caron et al., 2021a) and variance-invariance-covariance regularization (VICReg) (Bardes et al., 2022)) to simulate edge computing training without preprocessing. We process soundscapes collected from multiple devices at different locations, addressing the challenge of varying statistical features over time and space.

The experimentation of our approach include the visualization of the output space to extract various aspects of the data, such as recording time, date, and location. Clustering properties are derived from morphological features in spectrograms, including birdsongs, environmental conditions (e.g., rain), insect sounds, and human activity. Clusters are labeled using an alternative supervised-learning architecture (BirdNET (Kahl et al., 2021)). The clustering properties are validated using a human-expert labeled dataset (NIPS4Bplus (Morfi et al., 2018)). Acoustic indices are utilized to add explainability to clustering properties extracted from the output feature space.

Our method demonstrates the ability to separate different spectrogram features, allowing for the identification and analysis of various soundscape components. This capability is crucial for understanding the complex and dynamic nature of ecosystems through acoustic data.

The novel contributions of our approach to the field include:

- A framework for processing and learning from distributed data collected by edge cyberinfrastructure, addressing the challenges of real-world, uncurated datasets.
- Application of SSL to soundscape analysis, enabling the extraction of meaningful features without needing extensive labeled data.
- Integration of multiple data sources and temporal variations, providing a more comprehensive understanding of ecosystem dynamics.
- A method for adding explainability to SSL-derived features through acoustic indices, bridging the gap between machine learning outputs and ecological interpretations.
- Demonstration of the effectiveness of SSL in analyzing poorly curated and unlabeled datasets collected by edge infrastructures like Sage, opening new possibilities for large-scale ecosystem monitoring and biodiversity assessment.

By addressing these challenges and providing these novel capabilities, our approach paves the way for more robust and adaptable methods of ecosystem monitoring using edge computing and AI technologies.

2. Materials and methods

Our methodology, illustrated in Fig. 1, comprises three main stages: training, inference, and post-processing. We convert soundscapes to spectrogram images, then train a neural network to produce vector representations of each image using self-supervised learning techniques. During training, each input image is augmented to create two versions, which are processed through an encoder to produce embedding vectors. A distance function between these vectors is computed as a loss for network training. In the inference stage, all images are processed through the pre-trained network to obtain embedding vectors, forming a multidimensional output feature space. Post-processing involves various techniques to study the feature space, including dimensionality reduction (PCA, t-SNE) for visualization and clustering methods (DBSCAN, k-means, KNN).

2.1. Experimentation roadmap

Fig. 2 presents our experimentation roadmap, organized by datasets (columns) and experimental stages (rows).

We use three datasets: the Morton Arboretum dataset, a high bird density filtered subset, and the NIPS4Bplus dataset (Morfi et al., 2018). The experimental stages include:

Pre-processing: This stage involves converting audio soundwaves into spectrogram images. The procedure is described in detail in Section 2.3. This step transforms the raw audio data into a format suitable for neural network processing.

Training: In this stage, we train the networks using self-supervised learning procedures. The process involves using the spectrogram images without labels to train the network to produce meaningful embeddings. This stage is elaborated in Sections 2.4.

Inference: After training, we use the pre-trained models with frozen parameters to process the complete set of spectrogram images. This stage produces a set of embedding vectors for each image, which are used in subsequent analysis and validation steps. Importantly, even though the model may see the same data as during training, it hasn't been exposed to task-specific labels.

Post-processing: This stage involves analyzing the embedding vectors obtained from the inference step. We perform dimensionality reduction, clustering, and visualization analyses on these vectors. These techniques help us understand the structure and patterns in the learned feature space. This process is detailed in Section 3.1.

k-nearest neighbor (k−NN) classification: In this step, we use the embedding vectors from the inference stage to perform K-Nearest Neighbors classification for different labels. This helps evaluate how well the learned representations capture relevant features for classification tasks. The k-NN classification process is explained in Section 3.1.1.

Linear validation: This stage involves training a linear layer on top of the frozen pre-trained model for classification tasks. By keeping the pre-trained model fixed and only training a simple linear layer, we can assess how linearly separable the learned representations are for different classification tasks. This process is described in detail in Section 3.1.2.

Acoustic indices comparison: In this final stage, we compare the embedding vectors obtained from inference for a specific day with acoustic indices commonly used in ecology. This analysis helps bridge the gap between our machine learning approach and traditional ecological measures. The details of this experiment are provided in section 3.3.

For the high bird density subset (middle column in Fig. 2), we filter the original dataset to extract samples with a high probability of birdsong activity. We then train a new model from scratch on this subset and perform similar analyses as with the full dataset. This process allows us to focus on birdsong-rich samples and assess how this affects the model's performance.

With the NIPS4Bplus dataset (right column), we use the model pre-

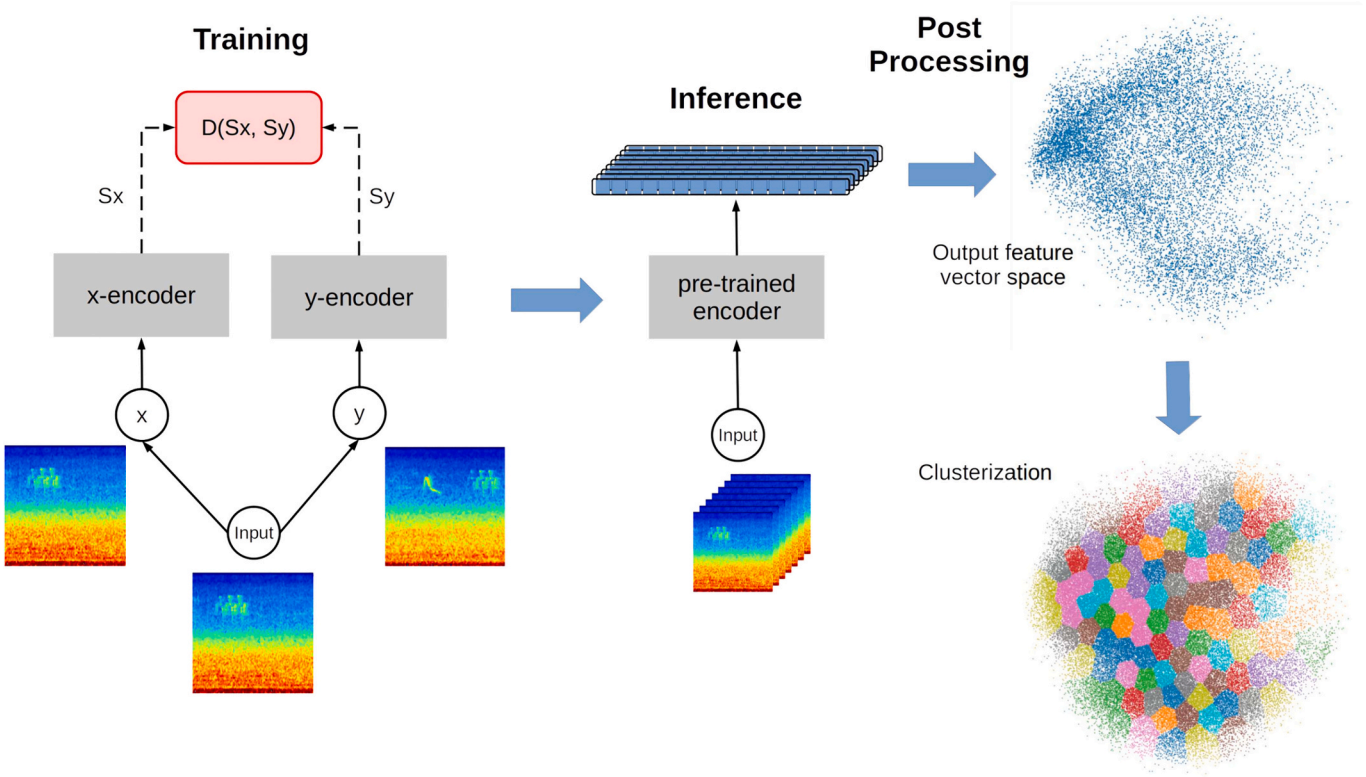


Fig. 1. Illustration of our procedure: dataset generation, pre-training of joint embedding architectures, inference to collect embedding vectors, and post-processing of the output space.

trained on the high bird density subset for inference. We then conduct k-NN and linear validation on this labeled and balanced dataset, allowing us to evaluate our method's performance on a standardized benchmark. These experiments are detailed in subsequent sections, addressing data visualization, validation, acoustic indices analysis, and comparisons across datasets. The code for reproducing these procedures is available on GitHub (Dematties and Rajani, 2024).

2.2. Recording process

We deployed nine recording devices in the Morton Arboretum which captured audio between May 24 and August 31, 2021. Fig. 3 shows a map of recorder locations, the distance between recording locations, timeline of the recordings, and mapping of recording devices to locations over the time window (Fig. 3a). The recording devices, from Frontier Labs, Australia, used in our study are shown in Fig. 3b and c. Two devices in the northwest corner were deployed in the "open grassland" habitat. The canopy recording devices, labeled Forest 1 and Forest 2, were deployed at the boundary between the forest and open grassland, and the remaining devices were canopy recorders that were deployed in the forested region. The southeast locations are closer to U. S. Interstate highways I-88 and I-355.

For canopy recorders we used the Frontier Labs BAR-LT, that has a rugged, lockable, and waterproof case made of UV resistant plastic. The case is painted in camouflage colors and offers flexible yet rugged mounting options. For open grassland we used the Frontier Labs Solar BAR, that is powered by a compact solar panel and rechargeable battery. Both devices incorporate four SD card slots and support recording in 16-bit waveform audio file format (WAV) and Free Lossless Audio Codec (FLAC) with sampling rates ranging from 8 kHz to 96 kHz. The devices have builtin GPS, which automatically time-syncs and geo-stamps the recordings. The devices feature highly sensitive, low-noise omni-directional microphones with a 20 dB class A pre-amp and cable driver, and they automatically log the microphones' serial number and manufacturing date for every recording. All devices provide text-based log-files (in comma-separated variable (CSV) format) that record specifics including the GPS location, microphone serial number, sampling frequency, audio format, and gain. We used WAV format to record audio in mono channel at 16-bit resolution and a sampling rate of 44.1n kHz. In Fig. 3b and c we also provide details of the microphones' gains.

Pseudo-replication in our recordings due to proximity of recording devices could affect the separation of the spectrograms in our analyses, the approximate distance between different recording zones is shown in Fig. 3a. Forest 1 is 85 ft. away from open grassland terrain. Forest 2 is 195 ft. away from high-voltage power lines. The organization of the recorders covers a diagonal line extending Northwest throughout a distance of 4330 ft. (1.32 km) from Forest 6 to Grassland 1 zone.

2.3. Audio data pre-processing

We pre-processed the 2.3 TB Morton Arboretum dataset by chopping the audio files into non-overlapping 9 s windows and converting the waveforms to spectrograms. We generated 1,976,583 spectrograms, which are image representations of the audio data. The decision to use a 9 s window was determined through trial and error: the duration and repetition of the bird songs we encountered were such that a 9 s window greatly improved the chances of containing a full song rather than a song being split across two or more windows. A thorough investigation, however, will be be required to analyze the effects of different windows lengths and overlap on the performance of audio classification.

2.4. Self-supervised learning and its application to audio recordings

Self-supervised learning (SSL) is a machine learning paradigm that learns useful data representations without human-labeled data, instead using proxy tasks to generate pseudo-labels at scale (Chen et al., 2020a).

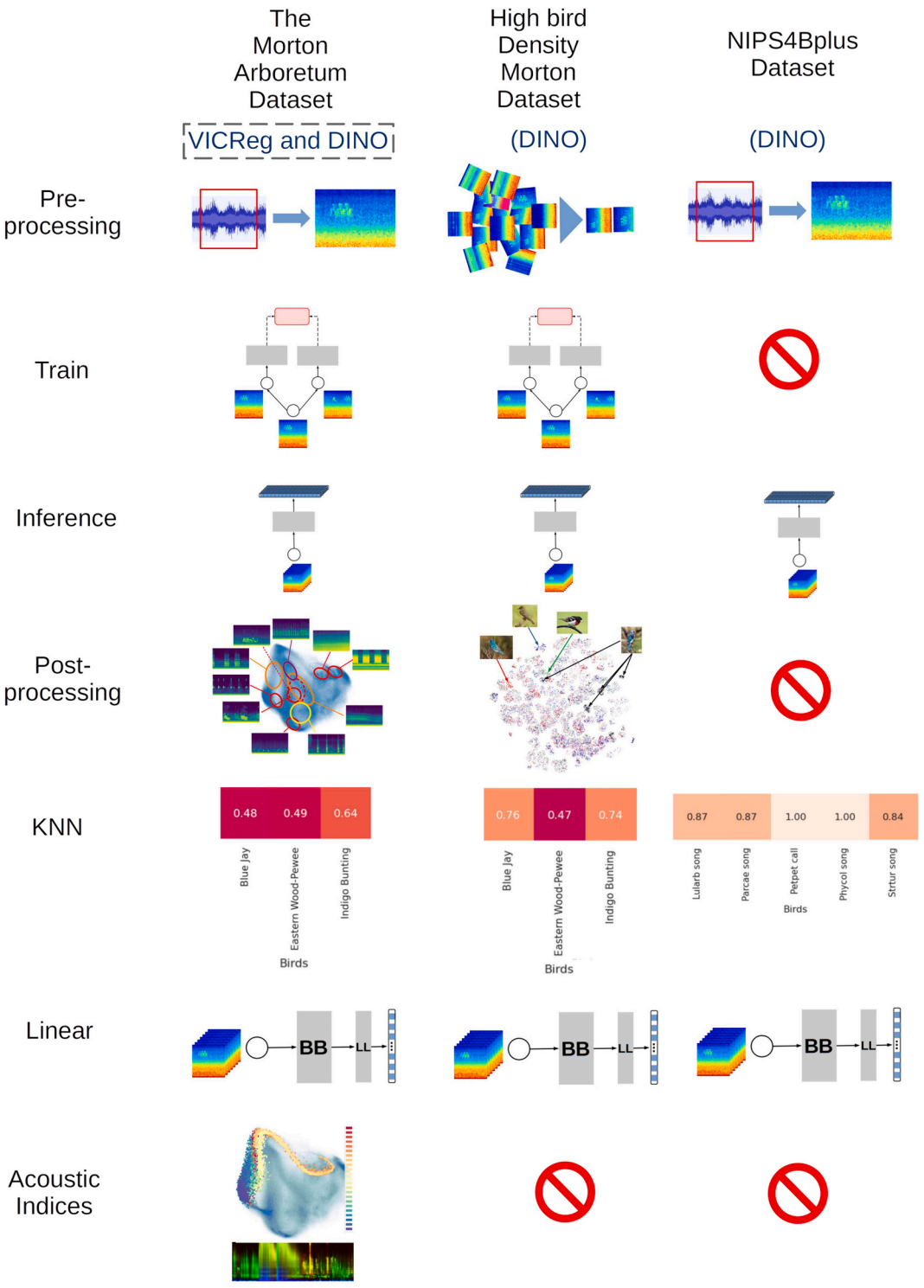
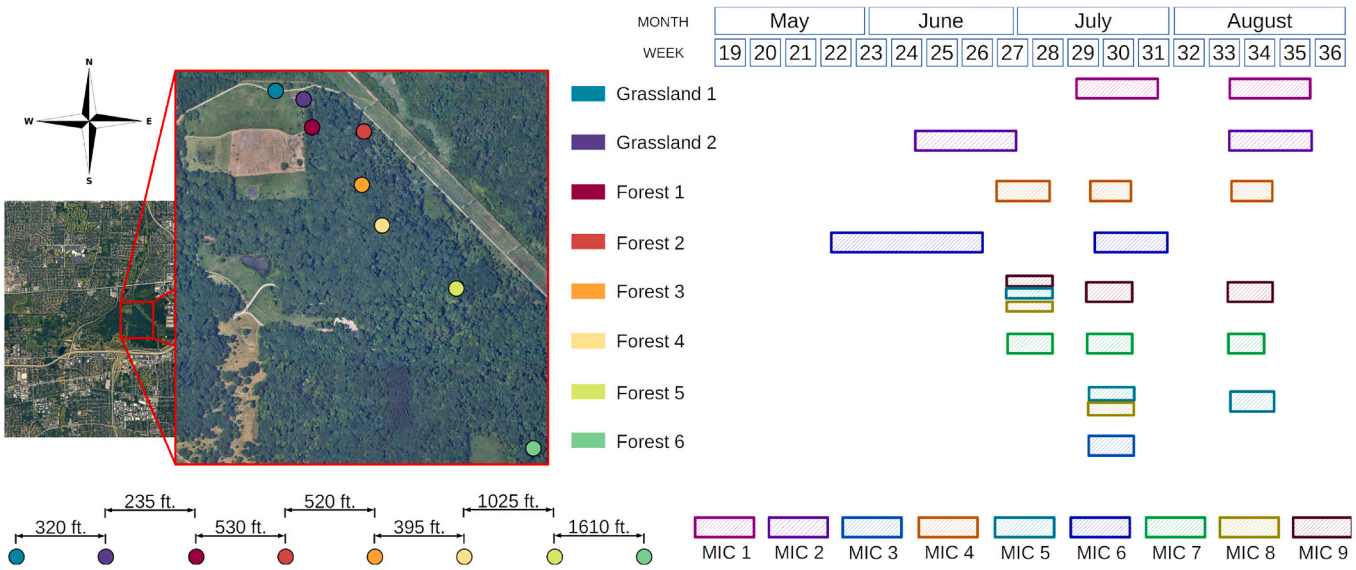


Fig. 2. Experimentation roadmap across datasets and procedures. BB: Backbone, LL: Linear Layer in linear evaluation.

Common examples include predicting original colors from grayscale images or predicting occluded parts of an image. Many powerful language models are pre-trained using simple proxy tasks like predicting the next token in a text sequence (Radford et al., n.d.).

In computer vision, joint embedding (JE) architectures have significantly advanced SSL, reducing the need for labeled data in various tasks. Key works include SimCLR, SwAV, BYOL, SimSiam, and DINO (Caron et al., 2021a; Caron et al., 2021b; Chen et al., 2020b; Chen and He, 2020; Grill et al., 2020). These architectures aim to make networks invariant to certain input augmentations, typically using two branches processing differently augmented versions of the same image. The network learns to produce similar output vectors despite augmentation differences, acquiring robust representations.

To analyze audio spectrograms, we explored two SSL strategies: DINO (Caron et al., 2021a) and VICReg (Bardes et al., 2022). DINO uses self-distillation with a student-teacher strategy, where the student learns to mimic the teacher's outputs for different augmentations. VICReg uses a different approach without a stop gradient, where both branches learn to mimic each other, with additional regularization losses to increase information in network representations.



(a) Locations and time of the recorders at Morton Arboretum. Left: Some of recording devices were moved during the recording period, we assigned an identifier for each location (Grassland 1-2, Forest 1-6). Right: A timeline of the recording activity per week/month and location.

(b) Canopy recorder. Seven devices: MIC 5 (45dB), MIC 9 (45dB), MIC 6 (40dB), MIC 3 (45dB), MIC 8 (45dB), MIC 4 (45dB), and MIC 7 (45dB). The size of the dataset recorded is 1.99TB across 100 days.

(c) Open grassland recorder. Two devices: MIC 1 (45dB) and MIC 2 (50dB). The size of the dataset recorded is 444GB across 78 days.

Fig. 3. Data collection setup.

For both architectures, we followed a two-phase strategy to map audio spectrograms to embeddings. The first phase involved training for 45 epochs, reducing embedding dimensionality, and removing corrupted samples. In the second phase, we retrained the model from scratch for 45 epochs with extraneous samples removed. We then applied multiple clustering methods to study the resulting embeddings, including t-SNE with DBSCAN and PCA with k-means clustering.

For DINO, we performed an additional third phase focusing on bird detections. We identified clusters with the highest bird detection rates using a supervised model (BirdNET), then conducted longer training over 200 epochs on spectrograms from those clusters, followed by probing and validation experiments.

Training was parallelized across 8 NVIDIA A100 GPUs using a data parallel scheme. For VICReg, each phase of training used the same hyperparameters: the effective batch size was 2048, the base learning rate followed a linear schedule from 0 to 0.2 across 5 warmup epochs, and the LARS optimizer was used with a weight decay of scale 10^{-6}. We also used the default coefficients for the variance, invariance, and covariance terms in the loss function ($\lambda = 25, \mu = 25, \nu = 1$). For DINO, we used an effective batch size of 512, an initial learning rate of 5×10^{-5}, and a cosine learning rate schedule with 10 warmup epochs.

This approach to SSL in audio analysis demonstrates the potential for unsupervised learning techniques to extract meaningful representations from soundscape recordings, which can be particularly valuable in ecological monitoring and biodiversity assessment contexts.

3. Results

We applied SSL to the Morton Arboretum dataset and to the NIPS4Bplus dataset. In this section, we present extensive exploration of the output of the machine learning (ML) model using visualization tools and clustering methods.

3.1. The Morton arboretum dataset

After the second phase of training, we used the weights of the encoder network for each model (VICReg and DINO) to produce an embedding for each spectrogram in the dataset. To visualize the embeddings in two dimensions, we used principal component analysis (PCA) to reduce the dimensionality of the output vector to 50 (from 2048 using VICReg and 384 using DINO) and then used t-distributed stochastic neighbor embedding (t-SNE) with two components and a perplexity of 50. Next, we applied density-based spatial clustering of applications with noise (DBSCAN) with $eps = 1.2$ and a minimum number of samples per cluster of 35 for both DINO and VICReg. With the VICReg data it resulted in 919 clusters with 8 % of the points considered noise; with the DINO data we found 703 clusters with 10 % of the points considered noise. The average silhouette score in this clustering was

approximately 0.77 for VICReg and 0.71 for DINO. The DBSCAN clustering was used to support the separation of low and high bird detection rate activity.

Our data validates the attribution of separation of spectrograms to the different locations and clarifies the effects, if any, of intrinsic properties of the microphones and recorders on the clustering. In our data collection, three different microphones (MICs 5, 8, and 9) were co-deployed in location Forest 3. Later, MICs 5 and 8 were co-deployed in location Forest 5 (Fig. 3a). If the intrinsic properties of the microphones dominated substantially, the output feature vectors of location Forest 3 and Forest 5 should be considerably overlapping. However, this was not the case. Fig. 4 shows a subsample of 197,658 feature vectors representing 10 % of the complete dataset from VICReg and DINO; each point represents one spectrogram embedding. The samples are colorized by their recording location within the arboretum, and there is a visual separation between Forest 3 and 5 locations. Analysis in Section 3.1.1) further bolsters this observation.

Post data collection, we discovered that two devices (MIC 2 and MIC 6) were inadvertently configured with different gain settings, and could have affected the analysis, however, 6 devices with the same gain configuration were separated with high accuracy (see Fig. 11), suggesting inconsequential effects, if any. Deliberate data collection utilizing more microphones, longer collection periods, different configurations, and with multiple microphones in the same locations can further our understanding.

Both the time of year and time of day also significantly affect the acoustic content (Fig. 5). For both joint embedding (JE) techniques, we find that different date and time are positioned in different locations of the output feature space. For DINO, we observed a clear segmentation by month (depicted by the horizontal dimension of t-SNE in Fig. 5). In Fig. 5, June is mostly located on the left of the space, July at the center, and August on the right. Similarly, week number colorization (shown in the center of Fig. 5) follows the patterns found for the different months but adds information with finer-grained separation (e.g., for DINO, weeks 32, 33, and 34 follow a bottom-up pattern in the output space). For VICReg, different months are localized in different angular spans, with weeks presenting a finer subdivision. The diurnal distribution of the points in the output space presents certain patterns, too. We observe groups of points that correspond with continuous intervals of several hours such as 22–02 UTC (16–20 evening CDT), 5–9 UTC (23–3 night CDT), and 12–15 UTC (6–9 morning CDT) (Fig. 5, right side).

We observed that the output-space patterns captured the extent of biological activity as signified by the prevalence of birdsongs. In Fig. 6 the left-hand side of the figure shows the clusters with BirdNET detection rates less than 30 % (low activity) and the right-hand side shows clusters with detection rates greater than 70 %.

The feature embeddings appear to capture information about device location within the arboretum. We observed that open grassland devices, such as Grassland 1 and Grassland 2, as well as devices located near open grassland, such as Forest 1, have points close together in both VICReg and DINO (Fig. 4). There is a dense cluster of points associated with Forest 2, which is closer to an open grassland region under high-voltage power lines. Fig. 5 provides evidence supporting the observation that clustering of points is a function of location in addition to time. For example, the patterns associated with device location Forest 2 across all months are close in this output space. Other device locations are concentrated in diverse regions of the output space. Figs. 4 and 6 show that clusters populated with more than 70 % birdsong detections are mostly from locations Forest 2 and Grassland 1.

The t-SNE method is designed for visual analysis of the data and it performs non-linear dimension reduction while preserving the local structure (cluster) of data, but is computationally intensive. Because t-SNE does not allow the addition of new samples, it is also not suitable for statistical analysis. We hence explored only a subset of the dataset using this method, using it to guide our analysis. For in depth analysis we used a simple dimensionality reduction method, PCA, on a subsample of the data and then fit new samples to the trained dimensionality reducer. To carry out this approach, we conducted postprocessing of our data in three stages: first we scaled our data, then we reduced its dimensionality to 2 using PCA, and finally we applied k-means clustering. We fit the PCA model on 20 % of the data (395,316 vectors). Principal components 1 and 2 explain approximately 25 % (0.14, 0.11) of the original data variance. We saved the models for scaling and reducing the dimensionality, as well as the labels of the different clusters for each sample.

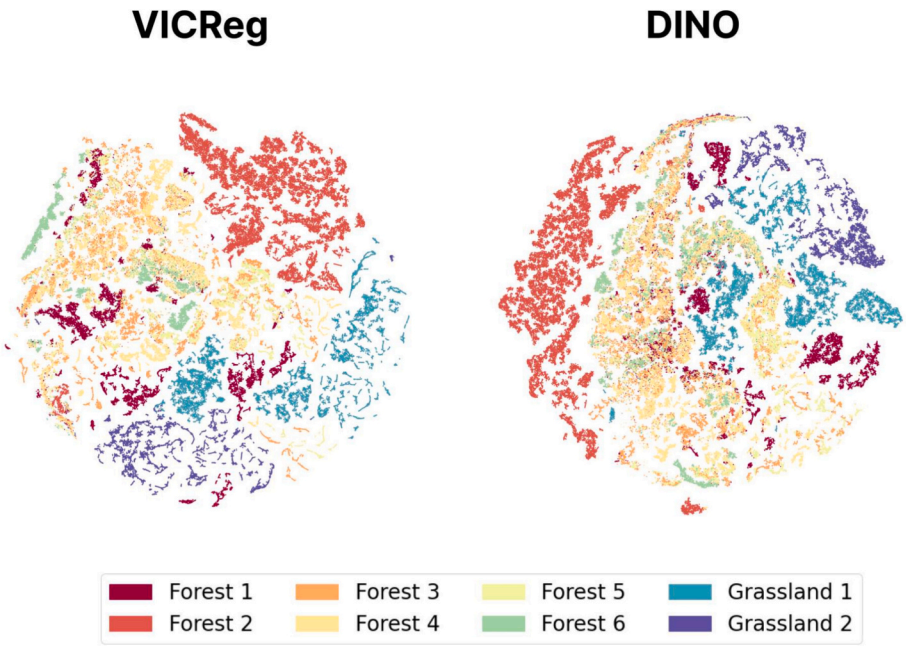


Fig. 4. Output feature vectors from VICReg (left) and DINO (right) reduced to 2 components using t-SNE. The weights of the encoder network for each model (VICReg and DINO) were used to produce an embedding for each spectrogram in the dataset. The dimensionality of the output vector was reduced to 50 using PCA, and then t-SNE was used with two components and a perplexity of 50 to make them suitable for 2D visualization. Points are colored according to their recording locations within arboretum.

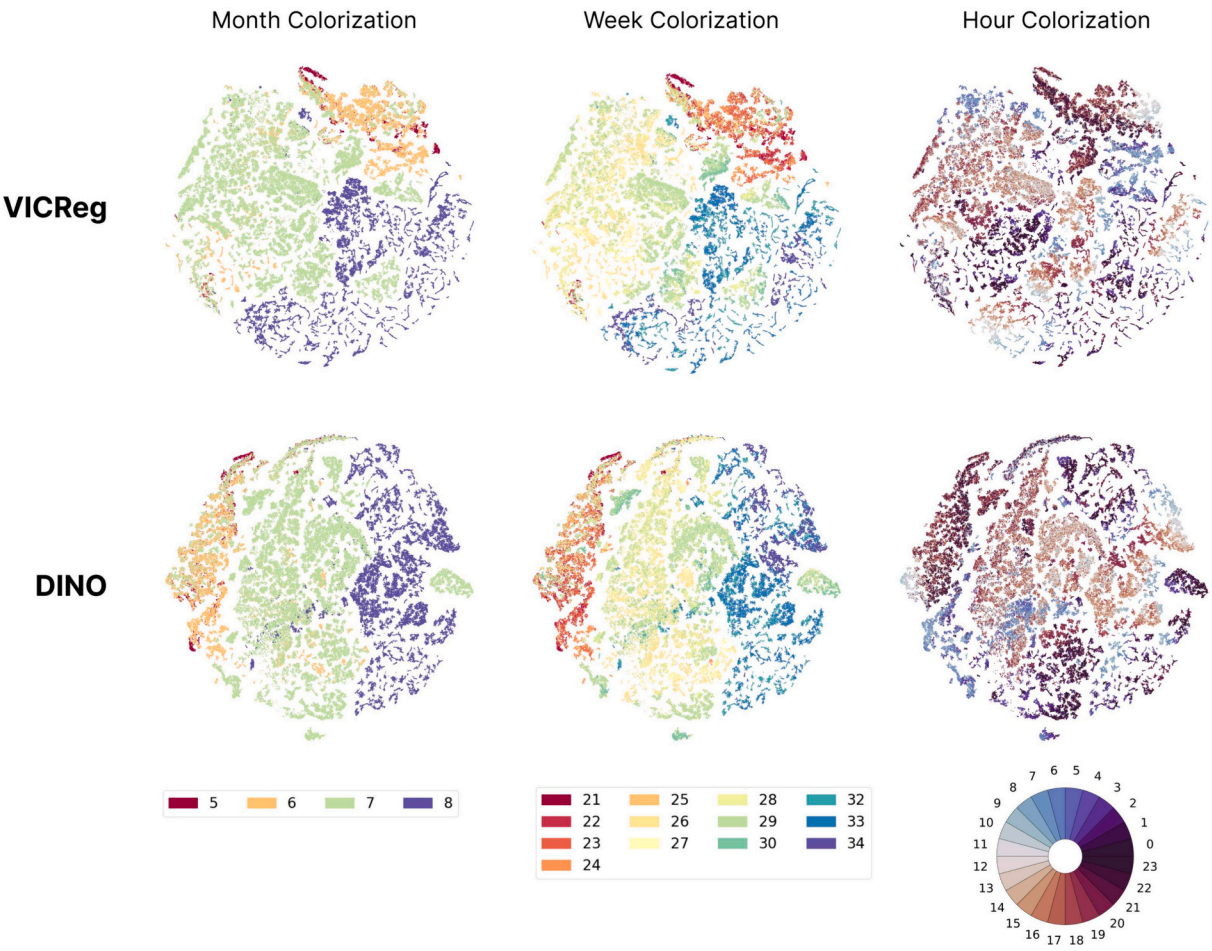


Fig. 5. Output feature vectors from VICReg (top) and DINO (bottom) reduced to 2 components using t-SNE. Left: Colorization of points corresponding to different months. Center: Colorization of points corresponding to different weeks. Right: Colorization of points corresponding to different hours in a day.

This approach allows us to process and classify new samples using a k-NN classifier. In Fig. 7 we can see more than 2 million samples transformed by the models and colorized by month, week, hour, and amount of bird activity.

Different aspects of the data, such as calendar, time, location, and bird activity, have a pronounced effect on the output features highlighted by the network, confirming the output of t-SNE (Fig. 7). In DINO, for example, points representing months May to August are encountered in order when moving gradually from left to right horizontally in the PCA output. There are also time intervals that are concentrated in different regions of the output space (previously seen in Fig. 5). t-SNE showed a better isolation of spectrograms by month, week, and hour. The *high bird detections* plot highlights clusters in which more than 60 % of spectrograms are classified with some bird detection by BirdNET. The *low bird detections* plot highlights clusters in which less than 40 % of spectrograms are classified with some bird detection by BirdNET. As can be seen in both Figs. 6 and 7, there is a clear separation in the locations of clusters with dense and sparse birdsong activity.

The network retains the semantic content of the spectrograms in the dimension-reduced feature vectors from DINO (Fig. 8). We show the results of a similarity function that finds the 400 most similar samples to a reference spectrogram. The similarity between spectrograms is measured in the original space by using Euclidean distance on 384 dimensions; then, the chosen points are reduced to our two-dimensional PCA mapping. The projected features of the similar samples are indicated by colored ellipses. The features are from non-overlapping regions in the original space but the projection to 2D results in overlapping ellipses (Fig. 8). The model identifies and isolates different kinds of environmental and biological activity in different zones of the output space, such as rain, different kinds of insects, and birds.

We observed a correspondence between the semantic content of the spectrograms with respect to their location in the output space and the classification returned by BirdNET (Fig. 8). This suggests that the network is learning important features present in the birdsong shapes, as well as spectrogram features such as background noise, insects, low-frequency noise, and accompanying background birds, in order to increase the performance on the assigned pretext task. We found two blue jay instances in different locations of the output space. Even though the most frequent birdsong classification using BirdNET was the same (blue jay) in both regions, there are different bird calls, differences in the background noise, and overlapping birdsongs. The network seemed to be capable of handling these capturing variations beyond the predominant foreground birdsong patterns.

3.1.1. *Evaluation with k-NN*

Through various experiments and visualizations we observed that the model exhibited semantic coherency, clustering certain aspects of the visual appearance of the spectrograms close together in the output space. One of the common practices adopted in SSL for validation is the use of the k-NN algorithm to objectively measure how well and under what aspects the input data is characterized by the model. Following this practice, we evaluated the clustering performance of the model analyzed using the k-NN algorithm. We explore the potential relationship between the clustering results and the data distribution.

Heatmaps of the k-NN performance in predicting the month or week in which a given sample was recorded is shown in Fig. 9. The high

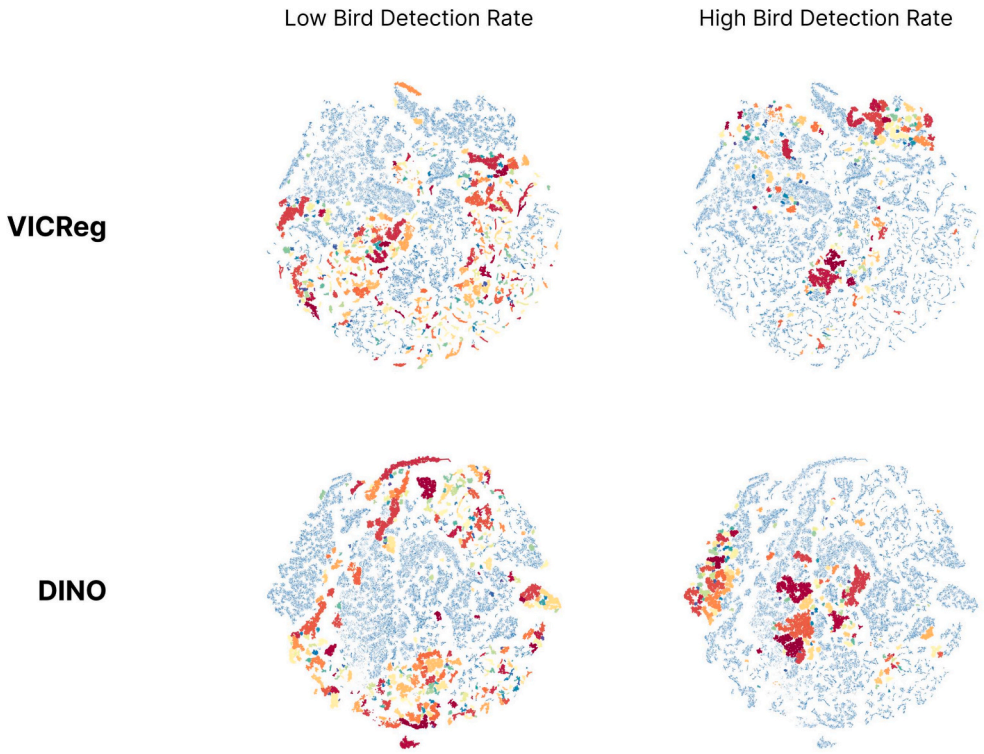


Fig. 6. Output feature vectors from VICReg (top) and DINO (bottom) reduced to 2 components using t-SNE. Left: Highlighted clusters with less than 30 % of bird detections using BirdNET. Right: Highlighted clusters with more than 70 % of bird detections using BirdNET.

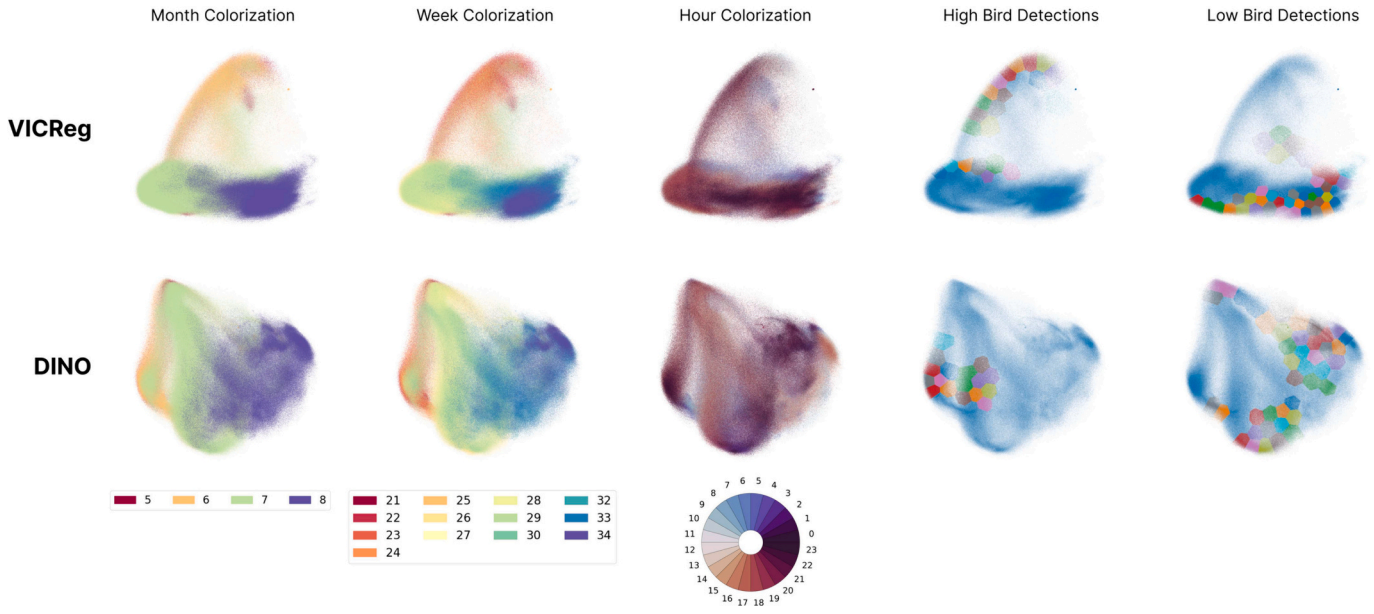


Fig. 7. PCA visualization of the complete Morton Arboretum dataset. From left to right, samples are colorized to highlight different aspects, such as month, week, and hour. In the fourth plot, k-means clusters with bird detection rates larger than 0.6 are displayed; in the fifth plot, k-means clusters with bird detection rates smaller than 0.4 are displayed.

performance of the k-NN model confirmed the phenomena observed in the clustering of the output features. The average performance in the classification decreases for larger number of neighbors, but the classification performance generally remains above 90 % for month and above 80 % for week granularities.

k-NN classification performance for different hours of the day (Fig. 10) confirmed our earlier observation (Fig. 5) that spectrograms recorded at a similar time of day contain similar acoustic content and thus form clusters in the feature space. The average k-NN classification performance for this benchmark was always 50 % and above.

k-NN classification performance when labeling the data according to the classifications assigned by BirdNET, and labeling the data according to recording location also showed high accuracy. We plot results for all recording locations and for the birds species where the model demonstrated its best performance (Fig. 11). The k-NN classification of recording location is high (generally above 90 % with all above 80 %).

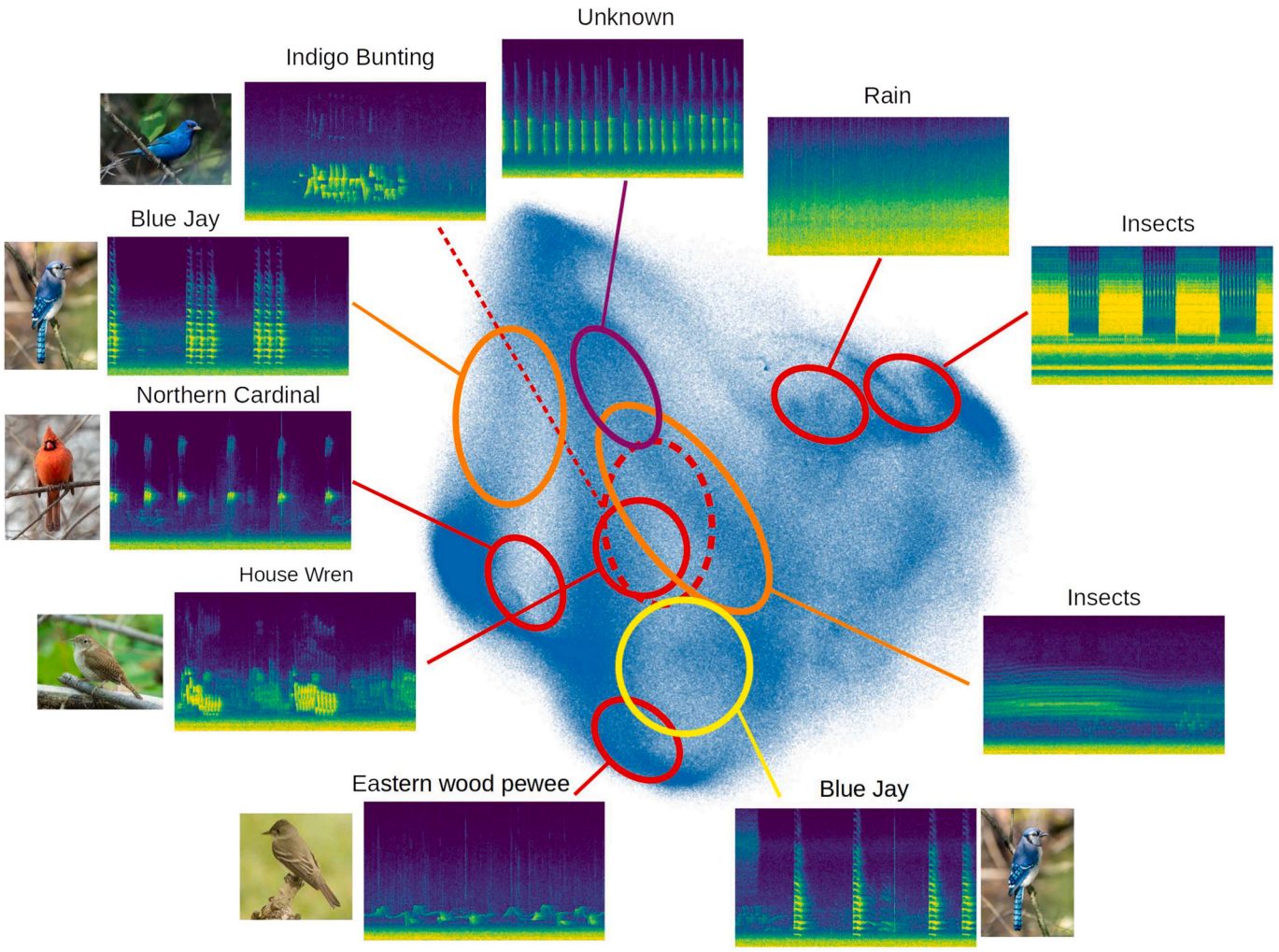


Fig. 8. Analysis of the feature space produced by DINO. Spectrograms exhibiting similar phenomena, such as birdsongs, insects, and rain, exhibit proximity in the dimension-reduced feature space.

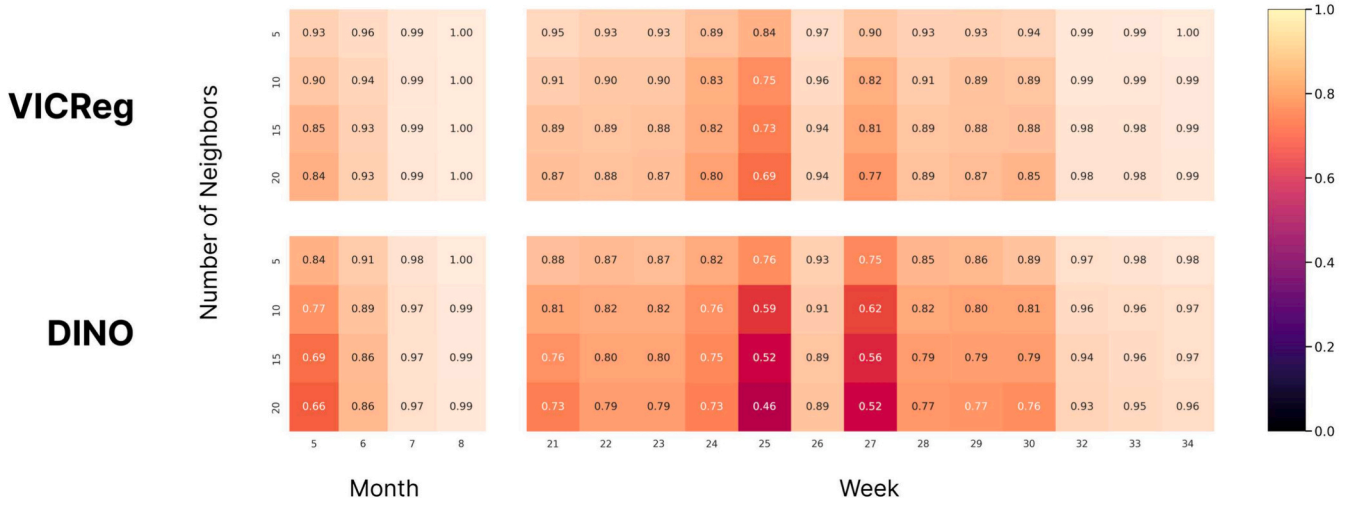


Fig. 9. Classification accuracy of k-NN for VICReg (top) and DINO (bottom) for predicting the month or week of a given sample. Experiments were performed using 5, 10, 15, and 20 neighbors.

There was a high variability in the performance of k-NN (Fig. 9–11), depending on the ground truth class of the sample. This variability can be accounted for *in part* by dataset imbalance. The distribution of our complete dataset is shown in Fig. 12. The month with the least data is May, followed by June and the accuracy of k-NN is lower for these two months. We observed relatively poor k-NN performance for weeks 25

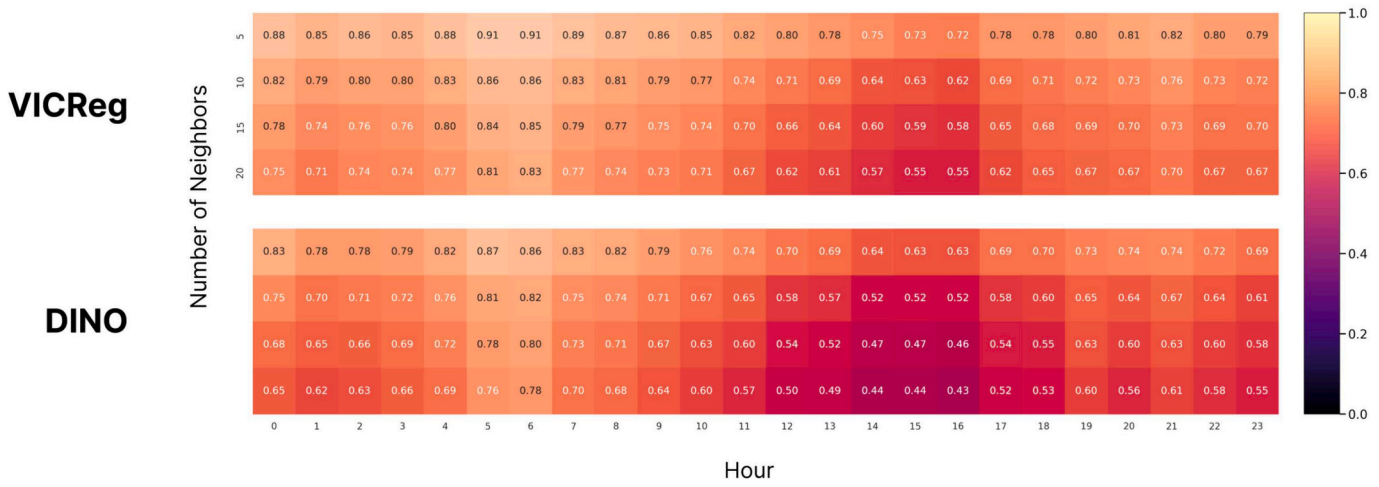


Fig. 10. Classification accuracy of k-NN for VICReg (top) and DINO (bottom) for predicting the hour of a given sample.

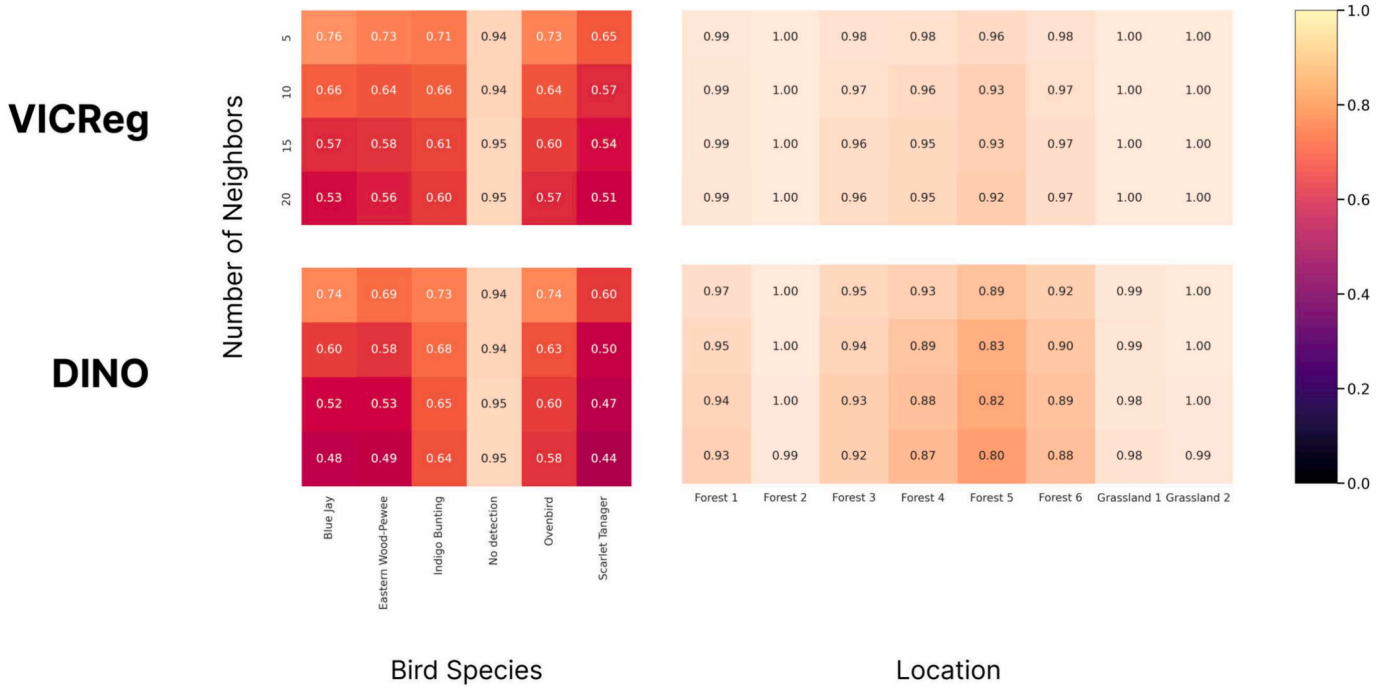


Fig. 11. Classification accuracy of k-NN for VICReg (top) and DINO (bottom) for predicting the bird species or location of a given sample.

and 27 and there is a corresponding smaller amount of data for those weeks. The data distribution also explains some of the patterns in the k-NN classification accuracy for location. Device MIC 6 has most of the spectrograms, followed by device MIC 1 (Forest 2 and Grassland 1 locations, respectively) and these two locations present the best k-NN classification performances. Both approaches (VICReg and DINO) had the most accurate classification for blue jay, eastern wood-pewee, indigo bunting, no detection, Ovenbird, and scarlet tanager. The models can predict with high accuracy whether a given spectrogram contains a birdsong (No detection) and this is likely due to the large number of training samples that are classified by BirdNET as background with no bird calls. The superior performance in classifying indigo bunting, eastern wood-pewee, blue jay, and scarlet tanager can also be explained by the data distribution (Fig. 12). Such is not the case for the Ovenbird, however, which is in the 23rd ranking position with less than 14,000 BirdNET detection events and has one of the best k-NN classification performances.

The dataset imbalance only *in part* explains the variability in classification performance. The are examples counter to this – a few examples include (1) k-NN classification for time of day was most accuarte between hours 4 and 7, and least accurate between hours 10 and 19 however, the dataset is relatively balanced with regard to the hours of recording; (2) Week 34 has high k-NN classification performance but a smaller number of samples; (3) k-NN performance on the data from device MIC 2 deployed in Grassland 2 location is one of the best, but this device collected a relatively small number of samples; and (4) the Ovenbird ranked 23rd with less than 14,000 BirdNET detection events and has one of the best k-NN classification performances (potentially the call is sufficiently distinct but requires additional study). These examples show that data imbalance contributes to, but does not fully explain k-NN classification performance.

3.1.2. *Linear evaluation*

Another commonly used evaluation protocol in SSL is linear

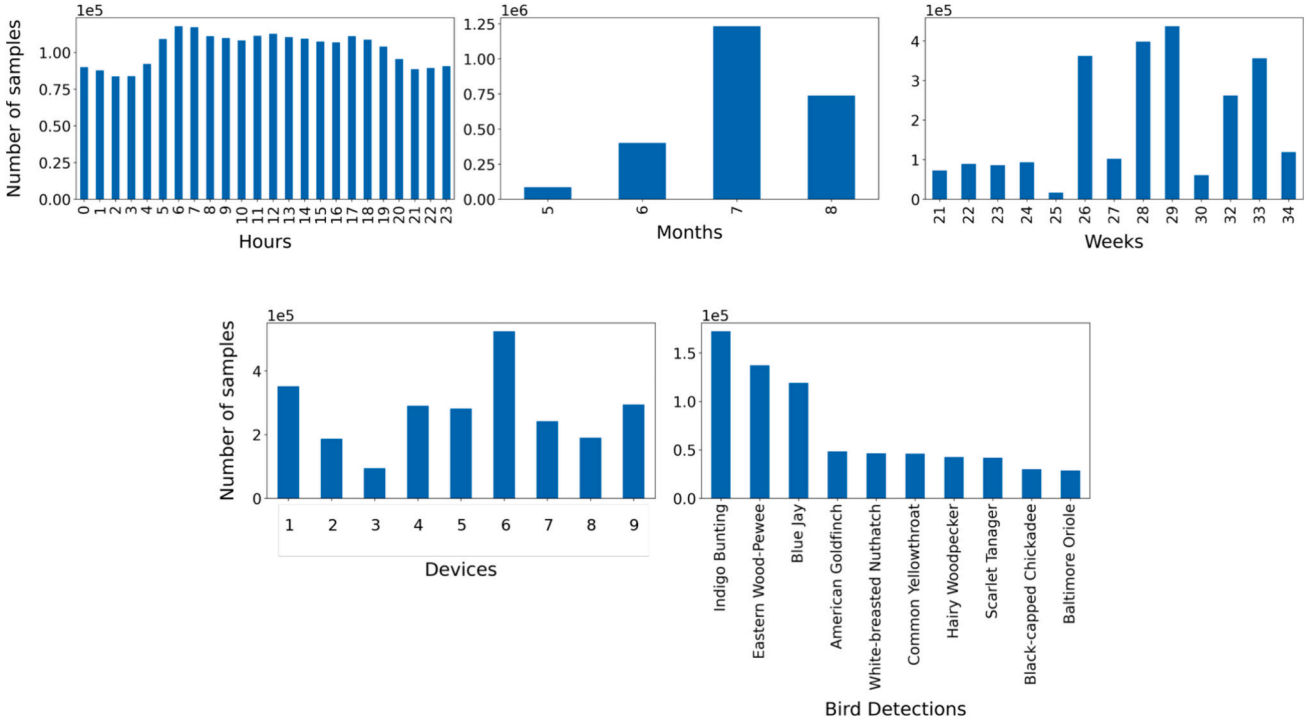


Fig. 12. Distribution by hour, month, week, device, and bird detections, of the complete dataset collected at the Morton Arboretum natural reserve.

evaluation. Linear evaluation consists of classifying samples by fixing the weights of the pretrained model and training an additional linear layer responsible for classification. We pursue this evaluation for DINO, using the sklearn multi-layer perceptron (MLP) classifier with a 20-unit hidden layer. We trained the model on BirdNET detection, month, week, hour, and location classification. We ran experiments using three different optimizers and obtained similar results with each (adam, lbfgs, and sgd); here, we present results using the lbfgs optimizer.

We took a sub-sample of 100,000 spectrograms from the dataset and conducted 10 k-folds cross-validation steps. We calculated the average performance computed on the 10 different test portions. The results are compiled in Table 1. For BirdNET detection classification performance in Table 1, only the classes shown in Fig. 11 were evaluated.

3.1.3. *High bird density dataset*

To study birdsongs detection in more detail, we performed a third phase of training using DINO. We extracted data from three clusters with the densest bird activity (from our PCA analysis, Fig. 7), and trained a new model using it. The dataset contained more than 100,000 spectrograms, mostly populated with birdsong activity. We trained the new model for 200 epochs. We used DBSCAN to cluster the outputs and used BirdNET for classification.

When visualizing the results we saw high correlation between dense patches of birdsong classification with the 97 clusters produced by DBSCAN. Fig. 14 shows some example spectrograms from clusters highlighted with colored circles in Fig. 13. Using birdNET we found: cluster number 77 contains spectrograms classified as indigo bunting (over 70 %); clusters 64 and 29 are mostly populated with spectrograms classified as blue jay; and clusters 28 and 94 are populated with eastern wood-pewee and rose-breasted grosbeak, respectively. Clusters of spectrograms reflected clear differences in birdsong classification and in properties such as the background noise, the frequency of the birdsong repetition, and the distance to the microphone (signal-to-noise ratio (SNR)) (Fig. 14).

Some regions that did not overlap with DBSCAN clusters contained outputs with similar BirdNET classification for the samples in the region (we have indicated some of these regions with colored squares in Fig. 13, right).

One region showed a highly concentrated patch with indigo bunting spectrograms (red square), another tiny cluster of densely packed spectrograms that are mostly classified as northern cardinal by BirdNET (light purple square), and another region had spectrograms with a biological pattern similar in morphology and frequency range to a birdsong but is not detected by BirdNET (purple square) (Fig. 15).

As above (§3.1.1) we evaluated the embedding using k-NN. We isolated the spectrograms with the highest classification accuracy (blue jay, eastern wood-pewee, indigo bunting, and no detection and analyzed these using k-NN (Fig. 16). With this new dataset, we observed an asymptotic performance (Fig. 16 left) whose minimum value is below the one seen when using the full dataset (Fig. 11 left). The decline in k-NN performance arises from poor classification of the *no detection* class, most likely because this dataset had a significant decrease in the relative number of samples with *no detection* classification. The k-NN classification performance of Blue jay increased by 30 %, while indigo bunting classification showed a 10 % improvement.

A linear evaluation was also conducted on the dataset with high bird activity density. We obtained a performance score of about 71 % for all the solvers and for only classes shown in Fig. 16.

Table 1
Linear evaluation performance for different classification tasks.

Linear Validation		Architecture	
		VICReg	DINO
Task	BirdNET detection	0.86	0.83
	Month	0.94	0.92
	Week	0.82	0.73
	Hour	0.41	0.33
	Location	0.97	0.92

3.2. *NIPS4Bplus dataset*

To validate our approach, an additional evaluation was performed using a balanced labeled dataset of birdsong identified by experts called

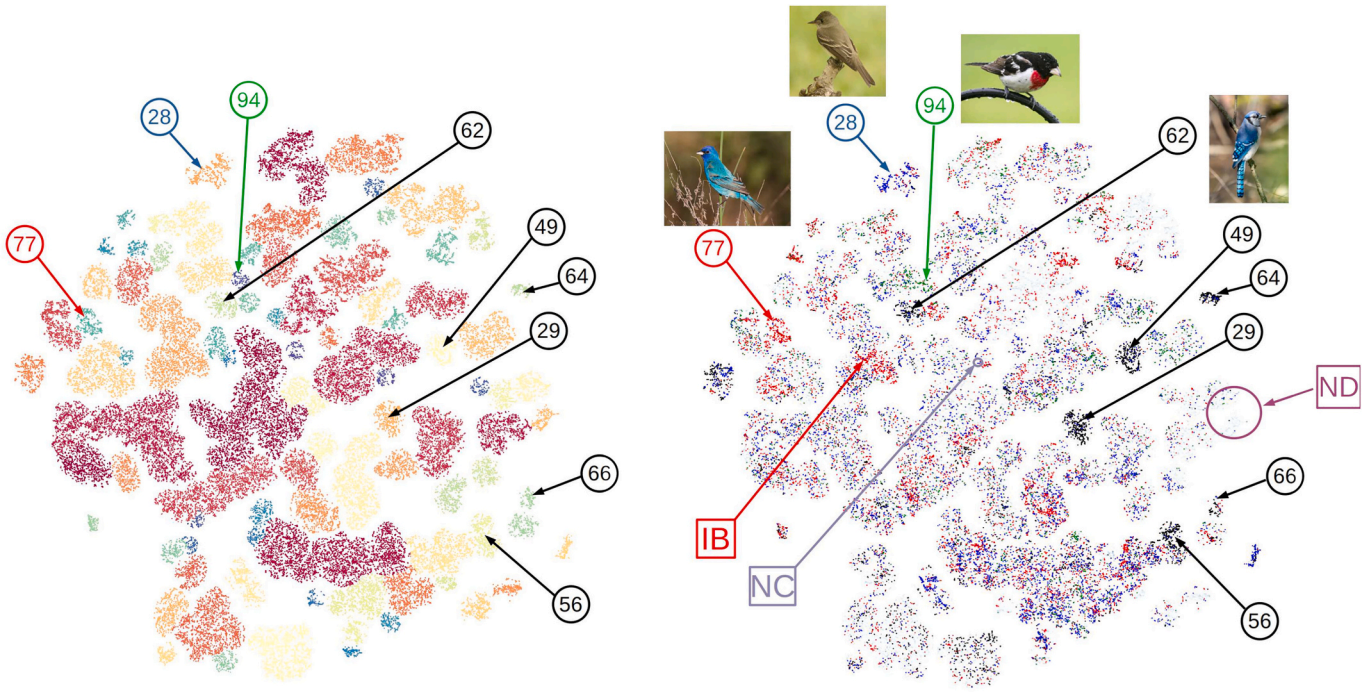


Fig. 13. Output from a model trained on dense birdsong activity spectrograms. Left: t-SNE plus DBSCAN clustering. Right: Colorization by BirdNET classification. Circles indicate clusters highly correlated with certain birdsong classified by BirdNET for blue jay (black circle), indigo bunting (red circle), eastern wood-pewee (blue circle), and rose-breasted grosbeak (green circle). Squares highlight other highly dense phenomena that did not overlap well with DBSCAN clusters on the left. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

NIPS4Bplus (Morfi et al., 2018). NIPS4Bplus is composed of 687 five-second recordings containing 87 classes with species tags and annotations. We produced spectrograms from one-second segments. We computed embeddings by conducting inference using the third-phase model, trained on the filtered Morton Arboretum dataset § 3.1.3).

The k-NN classification performance obtained for 11 of the 87 birds in the dataset is shown in Fig. 17. We chose the birds for which the network presented the highest classification performance to demonstrate the upper bounds of the model's capabilities. Keeping in mind its SSL nature, the model tested here was never trained on labeled data for birdsong classification, furthermore, this model was never exposed to this data during training. That being said, the global k-NN performance of the model on the 87 labels was around 37 %. The k-NN classification performance is essentially perfect for 4 of the 11 classes and is above 90 % for 7 of the 11 classes. The average k-NN performance is well above 80 % for up to 20 neighbors (Fig, 17 left). We conducted linear validation using the embeddings returned by the pre-trained network on the classes (Fig. 17 right) and obtained a classification performance of about 88 %. This classification performance, using a pre-trained model from a different dataset, indicates that our SSL approach combined with expert-assigned labels can identify birdsong with performance similar to birdNET (which reported an average precision of 0.791 (Kahl et al., 2021)).

3.3. Exploring the relationship with acoustic indices

SSL embedding vectors extract features from the audio data, however, biologists often study soundscapes using acoustic indices. Acoustic indices characterize and quantify spectral and temporal features that are correlated with certain human-interpretable aspects of the audio data. Soundscapes are divided into three main components: the biophony (biologically produced sounds), the geophony (geophysically produced sounds), and the anthrophony (human produced sounds) (Pijanowski et al., 2011a; Pijanowski et al., 2011b; Sueur et al., 2014). A simplified approach divides the spectral profile of the soundscape into two main regions where the anthrophony occupies the frequency band between 0.2 and 2 kHz and the frequency band between 2 and 8 kHz is generally occupied by animal sounds (biophony). Sounds produced by wind or rain mostly cover the entire spectrum, with more energy concentrated in the lower frequencies (Qi et al., 2008).

Other acoustic indices include the acoustic entropy index (AEI), acoustic complexity index (ACI), acoustic gradient index (AGI), and acoustic diversity index (ADI). AEI shows a logarithmic correlation with the number of species within the acoustic community (Sueur et al., 2009), and unharmed forests of Tanzania had significantly higher AEI values than degraded forests (Sueur et al., 2009). The ACI returns the quantification of the complex vocalizations or sounds produced by living organisms and, in particular, animals, by computing the variability of the intensities registered in audio recordings, despite the presence of constant human-generated noise (Pieretti et al., 2011). ADI measures the energy level in each of the 1 kHz bands in a spectrogram, indicating the extent to which different acoustic niches are occupied in the recording (Pekin et al., 2012). AGI is the real derivative of the spectrogram in time, normalized by the median derivative, which should correspond to the background (noise) derivative (Ulloa et al., 2021). For our analysis, we use scikit-maad, an open source Python package devoted to the analysis of environmental audio recordings (Ulloa et al., 2021). We used the Morton Arboretum dataset. To eliminate background noise, only sounds above −50 decibels relative to full scale were used.

We used the acoustic indices for characterizing the output feature vectors from the spectrograms obtained from two dates July 2, 2021 (Fig. 18a) and May 26, 2021 (Fig. 18b). Panels A show false color spectrograms with plots of ACI, AGI, and the temporal mean of the energy in the soundscape. Panels B plot anthropic energy, biological energy, AEI, ACI, AGI, and ADI. Panels C show all the output feature vectors from DINO colored by the hour of the day. On July 2, the mean energy of the spectrogram was concentrated in the lowest frequencies; on May 26, between hours 5 and 11, the mean energy tends toward higher frequencies. This energy dispersion is produced by rain in the soundscapes, which was confirmed by listening to the corresponding

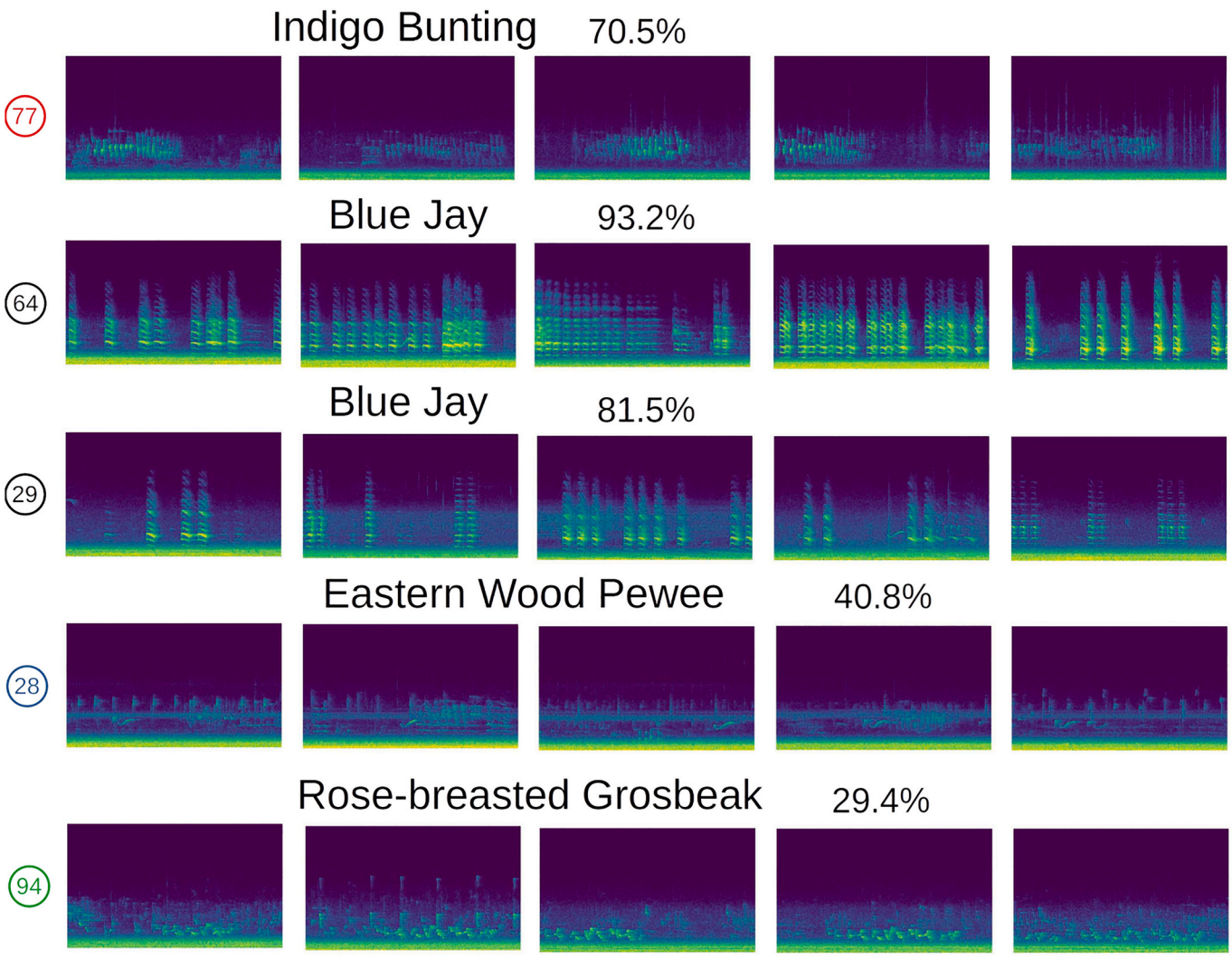


Fig. 14. Sample Birdsong spectrograms from different clusters from Fig. 13.

audio samples. In Fig. 18b C, the orange and yellow points in the plot correspond with feature vectors associated with hours 5 to 11; these points form a distinct "tail" in the series of embeddings and correspond to rain.

On July 2, the indices do not change much between the hours of 0 to 3 and 17 to 23, but they vary during the hours 3 to 17. Mapping the hours from 0 to 3 (red to orange) and from 17 to 23 (green to blue to purple) to the output space we observed that the features were more concentrated in the lower sections of the space; in contrast, the points corresponding with changing acoustic indices, from 3 to 17 (orange to yellow to green), tended to be distributed toward the top of the output space (Fig. 18a, C). From our spectrogram visualization, one can clearly see an increase in biological activity between hours 3 and 17, specifically arising from birdsongs.

On May 26 the acoustic indices are relatively flat between hours 5 and 11. Anthropic and biological energy attain their maximum values between these hours. After hour 19, and between hours 0 and 4, the acoustic indices present lower variability. In the output space visualization these points are mostly located on the left of the output space and form vertical bands. Low-variability intervals are from hours 0 to 4 (colors red to orange) and from hours 19 to 23 (colors light blue to purple) (Fig. 18b C).

Acoustic indices deliver hand-crafted features based on human expertise which summarize biologically accountable aspects of soundscapes. We explored correspondence between acoustic indices and the embedding vectors returned by our pre-trained models and envision two future investigations. First, one might explore directly mapping embedding vectors to acoustic indices. This may be possible but is beyond the scope of our current efforts. Second, the SSL embedding vectors encode properties of the soundscape and one might explore determining the relationship between the output feature space and ecological health or activity. There is potential for additional insights beyond the hand-crafted features when using automated feature extraction.

4. Discussion

Our study explores the application of SSL techniques in audio classification, particularly for identifying birdsongs and other environmental sounds in ecological contexts. While numerous studies have employed supervised learning for birdsong identification and audio classification (Kahl et al., 2021), the exploration of SSL in this domain has been limited. Our work contributes to this emerging field by demonstrating the efficacy of two SSL techniques—DINO and VICReg—that do not require large annotated datasets, making them particularly suitable for ecosystem studies.

Recent advancements in ML and deep learning (DL), specifically in SSL techniques, have shown promise in separating soundscape components using various clustering methods. For instance, Morales et al. (2022) employed UMAP and a deep neural network for passive acoustic

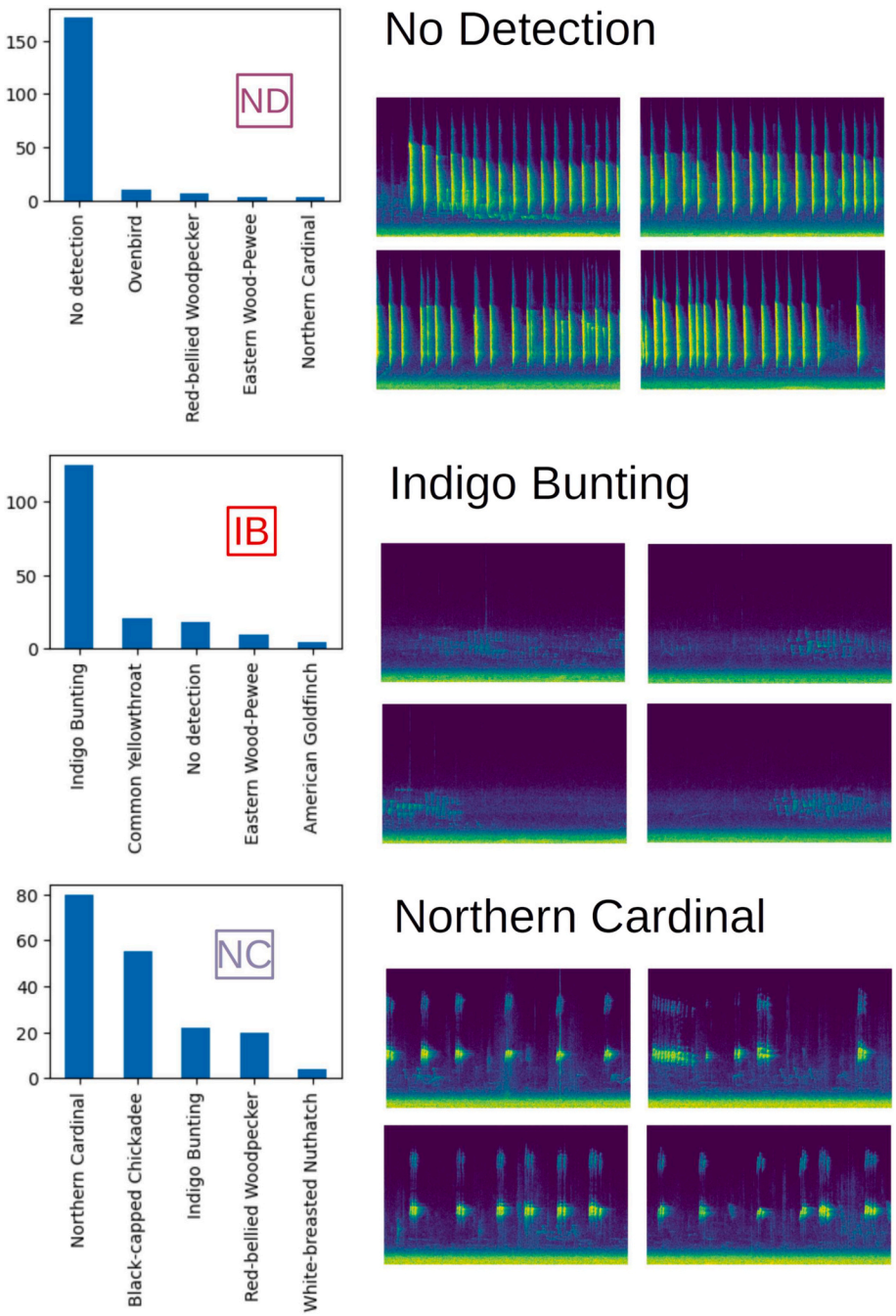


Fig. 15. High-density clusters in Fig. 13 right, which do not overlap well with single DBSCAN clusters on the left. Left: BirdNet classifications histogram. Right: Some example spectrograms.

monitoring of bird communities (Morales et al., 2022), while Michaud et al. (2023) used unsupervised classification to improve the quality of bird song recording datasets (Michaud et al., 2023). Our approach aligns with these studies but focuses on the application of JE frameworks, namely DINO and VICReg, to generate embedding vectors that represent audio spectrogram data.

Similar to the work of Rowe et al. (2021), who used acoustic autoencoders for biodiversity assessment (Rowe et al., 2021), our SSL approach demonstrates the ability to classify various sounds (such as birdsong, rain, insects) from acoustic signals collected in natural settings. The clustering techniques applied to our embedding outputs separated the data in meaningful ways, comparable to the results achieved by Thomas et al. (2021) using spectrogram-based latent space representations (Thomas et al., 2021).

A key advantage of our SSL approach, shared by studies like Sun et al. (2022) (Sun et al., 2022), is its potential to discover new classes in the data based on what was sensed, rather than being limited to pre-programmed classes as in supervised learning. This feature is particularly valuable in ecological studies where unexpected or rare sounds may be present.

Our findings regarding data imbalance and its effect on classification performance echo those of Sun et al. (2022), who concluded that the combination of transfer learning and data augmentation could be essential for classifying species' vocalizations in tropical forests (Sun et al., 2022). Our suggestion of a multi-step process (applying SSL, extracting samples of interest, and retraining the model) aligns with the approach proposed by Wisdom et al. (2020) for unsupervised sound separation (Wisdom et al., 2020).

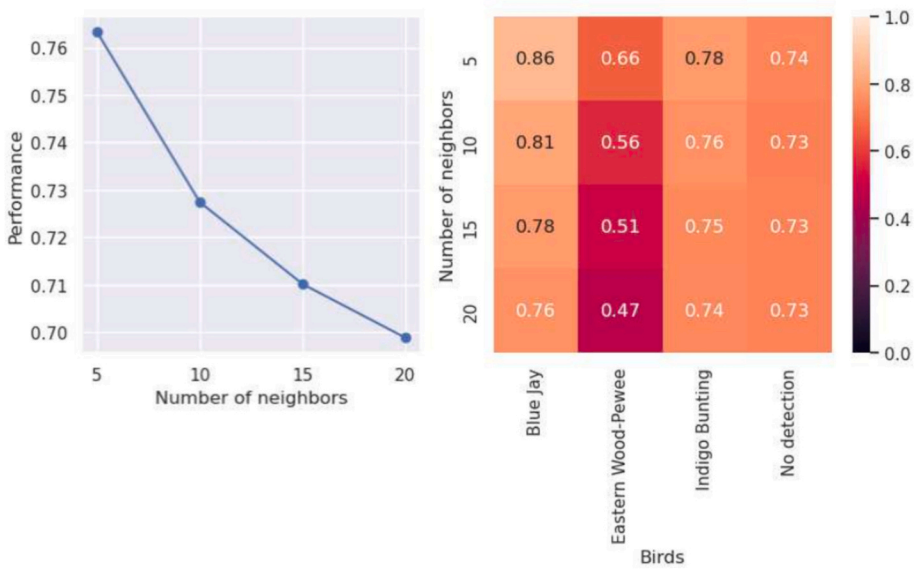


Fig. 16. k-NN classification performance on birdsongs on filtered spectrograms with high birdsong detection density. Left: Performance evolution as the number of neighbors increase from 1 to 20. Right: Heat map desegregated by birdsong.

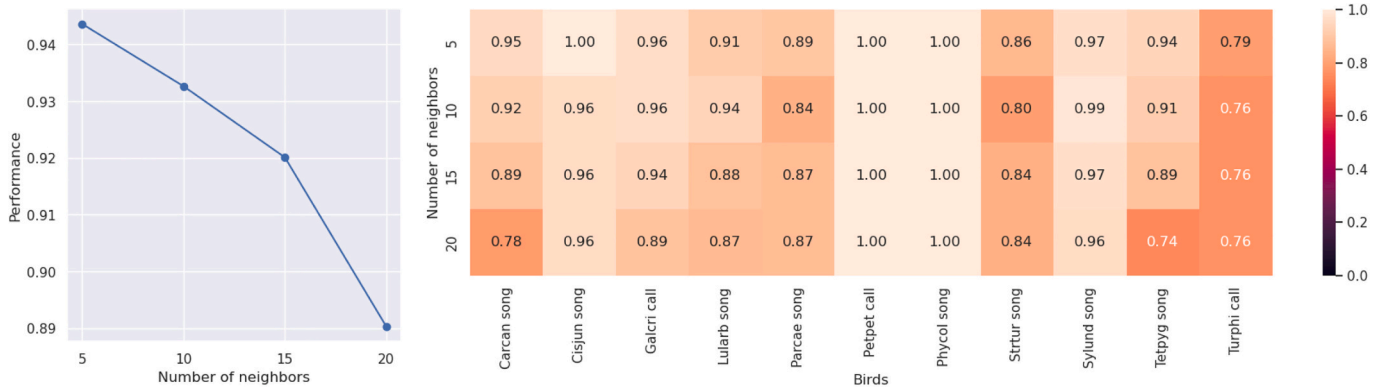


Fig. 17. k-NN classification performance on NIPS4Bplus dataset. Left: Average classification performance evolution as the number of neighbors increase. Right: Disaggregated classification performance per class.

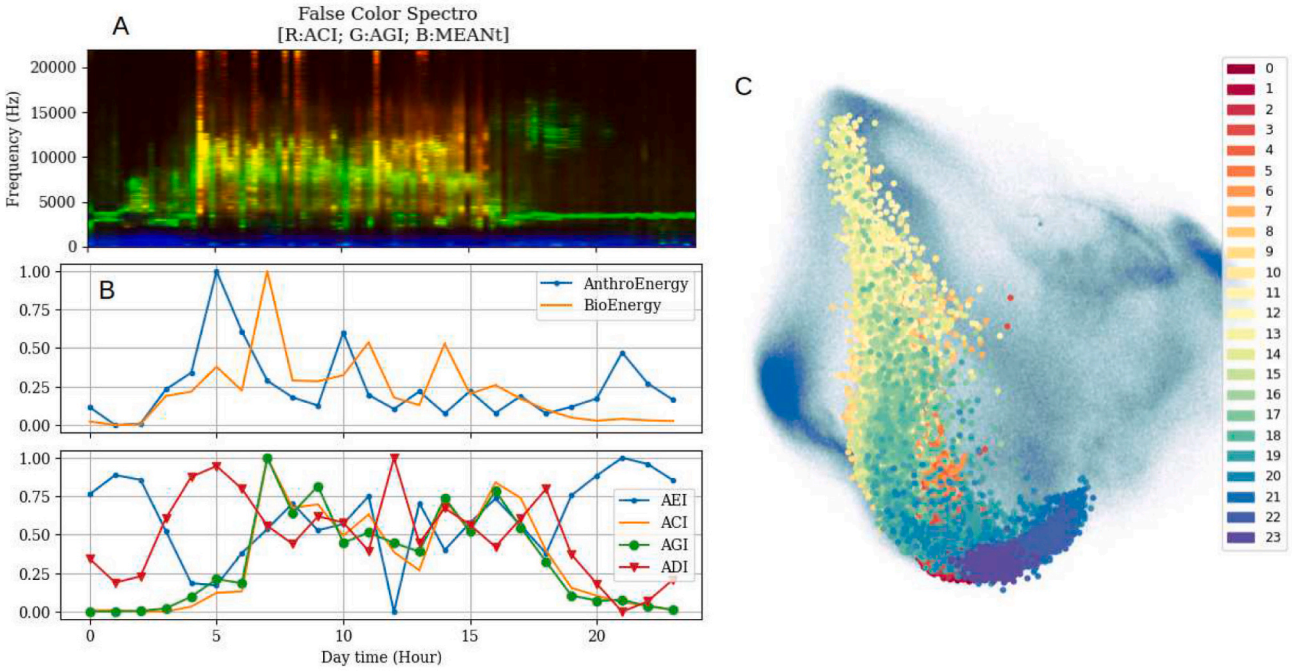
While we used PCA and t-SNE for dimensionality reduction and DBSCAN and k-NN for clustering, future work could explore other algorithms, such as the uniform manifold approximation and projection (UMAP) used by Morales et al. (2022) (Morales et al., 2022). The potential of our approach for within-species classifications, as demonstrated by McGinn et al. (2023) using BirdNET algorithm embeddings (McGinn et al., 2023), warrants further investigation.

The relationship between embedding vectors and acoustic indices, which we propose as an avenue for future studies, could build upon work like that of Morita et al. (2022), who used artificial neural networks to measure context-dependency in birdsong (Morita et al., 2022). This could provide deeper insights into the ecological significance of the sounds we classify.

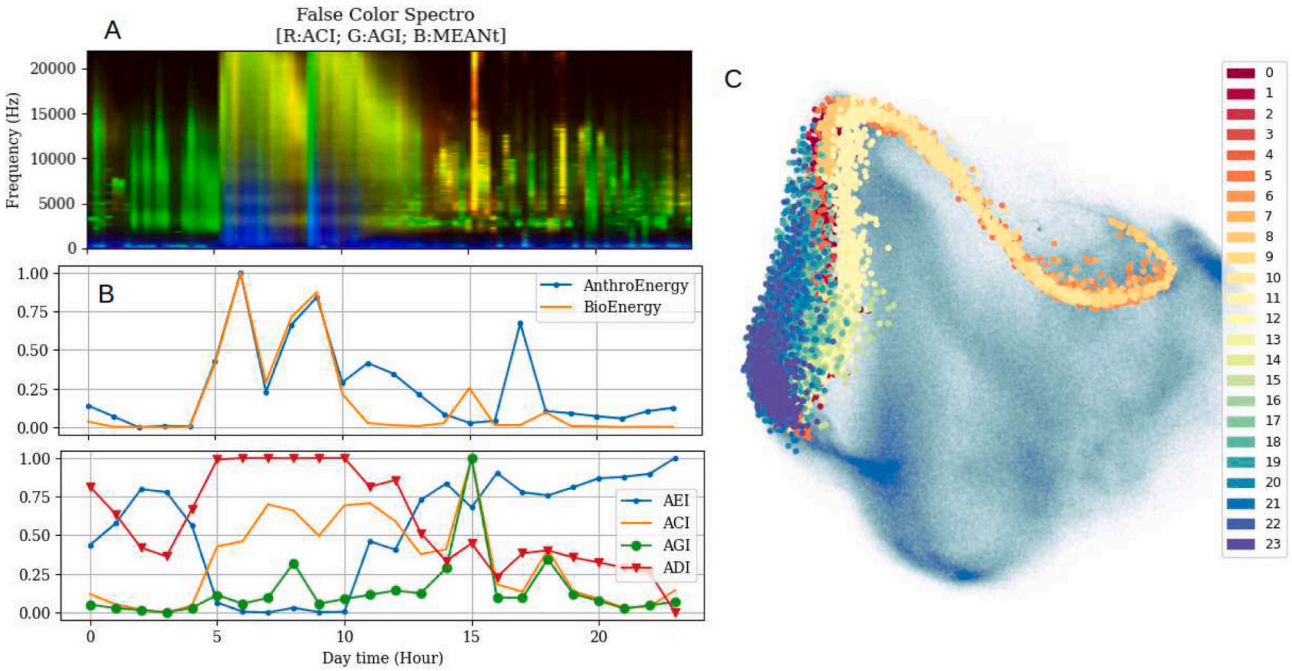
Our work contributes to the growing body of research applying advanced ML techniques to ecological data (Ghani et al., 2023). Like the study by Dematties et al. (2023) on cloud image analysis (Dematties et al., 2023), our approach demonstrates the potential of self-supervised methods in environmental monitoring, showing the growing relevance and transferability of these techniques that the community is adopting in different environments (such as analyzing the auditory settlement behavior of coral reef fishes (Google | SurfPerch | Kaggle, n.d.; Gordon et al., 2018)). As edge computing continues to grow (Edge Computing Market Size, Share, and Growth Report, 2030, n.d.), these techniques could become increasingly valuable for real-time, in-situ ecological monitoring and analysis.

It is important to note that this research, while demonstrating SSL's potential as a key component in edge-based ecosystem monitoring, cannot be directly applied without further development. Our study employs batch processing to validate the method, acknowledging that real-world applications require real-time analysis. Transitioning to a deployable edge infrastructure necessitates additional research, particularly in continual learning for adapting to dynamic ecosystems (Parisi et al., 2019). Future work should focus on adapting SSL models for continuous processing, integrating lifelong learning, and developing strategies for edge-based pretraining and inference. These advancements are essential for realizing flexible, label-efficient monitoring systems capable of real-time adaptation, bridging the gap between our current research and practical applications.

In conclusion, our study demonstrates the potential of SSL techniques in audio classification for ecological applications, offering a promising alternative to supervised learning approaches that require large annotated datasets. By building on and contributing to the work of researchers across the field of ecological informatics, we hope to advance the use of ML in understanding and monitoring ecosystems. Our future work will focus on optimizing these techniques, exploring their application to a wider range of ecological sounds, and investigating their



(a) Spectrograms from device MIC 5 on July 2, 2021



(b) Spectrograms from device MIC 6 on May 26, 2021.

Fig. 18. Comparison between acoustic features and embedding vectors. (A) False color spectrograms, (B) acoustic indices, and (C) embedding vectors for two different days of audio in the output feature space. Colored legend on the right represents daily hours.

potential for real-time, edge-based environmental monitoring systems.

CRediT authorship contribution statement

Dario Dematties: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Samir Rajani:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Rajesh Sankaran:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sean Shahkarami:** Software, Investigation, Conceptualization. **Bhupendra Raut:** Methodology, Investigation, Formal analysis, Conceptualization. **Scott Collis:** Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Pete Beckman:** Supervision, Resources, Project administration, Investigation, Funding acquisition,

Conceptualization. **Nicola Ferrier:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

5. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the first author used ChatGPT in order to find alternative words or rephrase sentences. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

None.

Data availability

Data and code will be shared upon manuscript acceptance

Acknowledgments

The Sage project is funded through the U.S. National Science Foundation's Mid-Scale Research Infrastructure program, NSF-OAC-1935984. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. The field data collected for analysis was done in cooperation with the Center for Tree Science at the Morton Arboretum, a multidisciplinary tree research center in an internationally recognized outdoor tree museum located in Lisle, Illinois.

Appendix A. Research data for this article

The complete dataset for reproducing the experiments conducted in this research are available by means of Zenodo (Dematties, 2024).

References

AbdelRahman, M.A.E., 2023. An overview of land degradation, desertification and sustainable land management using GIS and remote sensing applications. Rendiconti Lincei. Scienze Fisiche e Naturali 34 (3), 767–808. https://doi.org/10.1007/s12210-023-01155-3.

Abeßer, J., 2020. A review of deep learning based methods for acoustic scene classification. Appl. Sci. 10 (6). https://doi.org/10.3390/app10062020, 2020, number: 6 Publisher: Multidisciplinary Digital Publishing Institute. URL https://www.mdpi.com/2076-3417/10/6/2020.

Ackerman, S., Farchi, E., Raz, O., Zalmanovici, M., Dube, P., 2022. Detection of data drift and outliers affecting machine learning model performance over time. arXiv:2012.09258.

Al-Atat, G., Fresa, A., Behera, A.P., Moothedath, V.N., Gross, J., Champati, J.P., 2023. The case for hierarchical deep learning inference at the network edge. arXiv:2304.11763.

Ashford, O.S., Guan, S., Capone, D., Rigney, K., Rowley, K., Cordes, E.E., Cortés, J., Rouse, G.W., Mendoza, G.F., Sweetman, A.K., Levin, L.A., 2021. Relationships between biodiversity and ecosystem functioning proxies strengthen when approaching chemosynthetic deep-sea methane seeps. Proc. R. Soc. B Biol. Sci. 288 (1957). https://doi.org/10.1098/rspb.2021.0950, 20210950, publisher: Royal Society. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2021.0950.

Bardes, A., Ponce, J., LeCun, Y., Jan. 2022. VICReg: variance-invariance-covariance regularization for self-supervised learning. arXiv:2105.04906 [cs]ArXiv: 2105.04906. URL http://arxiv.org/abs/2105.04906.

Beckman, P., Catlett, C., Altintas, I., Kelly, E., Collis, S., Ferrier, N., Papka, M., Olds, J., Reed, D., Sankaran, R., October 2019. Sage: Cyberinfrastructure for AI at the edge. https://sagecontinuum.org/, (accessed: 08/28/2020).

Biodiversity and Ecosystem Stability | Learn Science at Scitable, cg_cat: Biodiversity and Ecosystem Stability Cg_level: ESY Cg_topic: Biodiversity and Ecosystem Stability. URL https://www.nature.com/scitable/knowledge/library/biodiversity-and-ecosystem-stability-17059965/.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021a. Emerging properties in self-supervised vision transformers. arXiv:2104.14294 [cs]ArXiv: 2104.14294. URL http://arxiv.org/abs/2104.14294.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2021b. Unsupervised learning of visual features by contrasting cluster assignments. arXiv:2006.09882 [cs]ArXiv: 2006.09882. URL http://arxiv.org/abs/2006.09882.

Catlett, C., Beckman, P., Sankaran, R., Ferrier, N., Park, S., Kim, Y., 2019. Software-defined sensors: using edge computing to revolutionize sensing. In: AGUFM 2019. IN34A–01.

Catlett, C., Beckman, P., Ferrier, N., Solin, J., Taylor, V., Pancoast, D., Reed, D., 2022. Hands-on computer science: the Array of things experimental urban instrument. IEEE Comp. Sci. Eng. https://doi.org/10.1109/MCSE.2021.3139405.

Chen, X., He, K., Nov. 2020. Exploring simple Siamese representation learning. arXiv:2011.10566 [cs]ArXiv: 2011.10566. URL http://arxiv.org/abs/2011.10566.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. arXiv:2002.05709.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations. arXiv:2002.05709 [cs, stat]ArXiv: 2002.05709. URL http://arxiv.org/abs/2002.05709.

Cianfagna, M., Bolon, I., Babo Martins, S., Mumford, E., Romanelli, C., Deem, S.L., Pettan-Brewer, C., Figueroa, D., Velásquez, J.C.C., Stroud, C., Lueddeke, G., Stoll, B., Ruiz de Castañeda, R., 2021. Biodiversity and education offerings: a first global overview. Front. Public Health 9. URL https://www.frontiersin.org/articles/10.3389/fpubh.2021.637901.

Dematties, D., Jan. 2024. Complete Dataset links and GPS timestamps and locations of our recordings at the Morton Arboretum. https://doi.org/10.5281/zenodo.10573487.

Dematties, D., Rajani, S., Jan. 2024. Soundscape analysis: A self-supervised learning approach for ecosystem activity monitoring. https://doi.org/10.5281/zenodo.10459763.

Dematties, D., Raut, B.A., Park, S., Jackson, R.C., Shahkarami, S., Kim, Y., Sankaran, R., Beckman, P., Collis, S.M., Ferrier, N., 2023. Let's unleash the network judgment: a self-supervised approach for cloud image analysis. Artif. Intell. Earth Syst. 2 (2), 220063. https://doi.org/10.1175/AIES-D-22-0063.1. URL https://journals.ametsoc.org/view/journals/aies/2/2/AIES-D-22-0063.1.xml.

Edge Computing Market Size, Share & Growth Report, 2030. URL https://www.grandviewresearch.com/industry-analysis/edge-computing-market.

Fan, W., Chen, Z., Hao, Z., Wu, F., Liu, Y., 2023. Joint task offloading and resource allocation for quality-aware edge-assisted machine learning task inference. IEEE Trans. Veh. Technol. 72 (5), 6739–6752. https://doi.org/10.1109/TVT.2023.3235520.

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

Gomiero, T., 2016. Soil degradation, land scarcity and food security: reviewing a complex challenge. Sustainability 8 (3), 281 number: 3 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/su8030281. URL https://www.mdpi.com/2071-1050/8/3/281.

Google | SurfPerch | Kaggle. URL https://www.kaggle.com/models/google/surfperch.

Gordon, T.A.C., Harding, H.R., Wong, K.E., Merchant, N.D., Meekan, M.G., McCormick, M.I., Radford, A.N., Simpson, S.D., 2018. Habitat degradation negatively affects auditory settlement behavior of coral reef fishes. Proc. Natl. Acad. Sci. 115 (20), 5193–5198 publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1719291115. URL https://www.pnas.org/doi/10.1073/pnas.1719291115.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., Sep. 2020. Bootstrap your own latent: a new approach to self-supervised learning. arXiv:2006.07733 [cs, stat]ArXiv: 2006.07733. URL http://arxiv.org/abs/2006.07733.

Hao, J., Subedi, P., Ramaswamy, L., Kim, I.K., Feb 2023. Reaching for the sky: maximizing deep learning inference throughput on edge devices with ai multi-tenancy. ACM Trans. Internet Technol. 23 (1). https://doi.org/10.1145/3546192.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: a deep learning solution for avian diversity monitoring. Eco. Inform. 61, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236. URL https://www.sciencedirect.com/science/article/pii/S1574954121000273.

Kato, S., Huang, W., 2021. Land use management recommendations for reducing the risk of downstream flooding based on a land use change analysis and the concept of ecosystem-based disaster risk reduction. J. Environ. Manag. 287, 112341. https://doi.org/10.1016/j.jenvman.2021.112341. URL https://www.sciencedirect.com/science/article/pii/S0301479721004035.

Lin, J., Zhu, L., Chen, W.-M., Wang, W.-C., Gan, C., Han, S., 2022. On-device training under 256kb memory. arXiv:2206.15472.

Marselle, M.R., Hartig, T., Cox, D.T., de Bell, S., Knapp, S., Lindley, S., Triguero-Mas, M., Böhning-Gaese, K., Braubach, M., Cook, P.A., de Vries, S., Heintz-Buschart, A., Hofmann, M., Irvine, K.N., Kabisch, N., Kolek, F., Kraemer, R., Markevych, I., Martens, D., Müller, R., Nieuwenhuijsen, M., Potts, J.M., Stadler, J., Walton, S., Warber, S.L., Bonn, A., 2021. Pathways linking biodiversity to human health: a conceptual framework. Environ. Int. 150, 106420. https://doi.org/10.1016/j.envint.2021.106420. URL https://www.sciencedirect.com/science/article/pii/S0160412021000441.

McGinn, K., Kahl, S., Peery, M.Z., Klinck, H., Wood, C.M., 2023. Feature embeddings from the BirdNET algorithm provide insights into avian ecology. Eco. Inform. 74, 101995. https://doi.org/10.1016/j.ecoinf.2023.101995. URL https://www.sciencedirect.com/science/article/pii/S1574954123000249.

Michaud, F., Sueur, J., Le Cesne, M., Haupert, S., 2023. Unsupervised classification to improve the quality of a bird song recording dataset. Eco. Inform. 74, 101952.

https://doi.org/10.1016/j.ecoinf.2022.101952. URL https://www.sciencedirect.com/science/article/pii/S1574954122004022.
Morales, G., Vargas, V., Espejo, D., Poblete, V., Tomasevic, J.A., Otondo, F., Navedo, J. G., 2022. Method for passive acoustic monitoring of bird communities using UMAP and a deep neural network. Eco. Inform. 72, 101909. https://doi.org/10.1016/j.ecoinf.2022.101909. URL https://www.sciencedirect.com/science/article/pii/S1574954122003594.
Morfi, V., Bas, Y., Pamuła, H., Glotin, H., Stowell, D., 2018. Nips4bplus: a richly annotated birdsong audio dataset. arXiv:1811.02275.
Morita, T., Koda, H., Okanoya, K., Tachibana, R.O., 2022. Measuring context dependency in birdsong using artificial neural networks. PLoS Comput. Biol. 17 (12), 1–24. https://doi.org/10.1371/journal.pcbi.1009707.
Murshed, M.G.S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., Hussain, F., 2021. Machine learning at the network edge: a survey. ACM Comput. Surv. 54 (8), 1–37. https://doi.org/10.1145/3469029.
Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S., 2019. Continual lifelong learning with neural networks: a review. Neural Netw. 113, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.
Paz, D.B., Henderson, K., Loreau, M., 2020. Agricultural land use and the sustainability of social-ecological systems. Ecol. Model. 437, 109312. https://doi.org/10.1016/j.ecolmodel.2020.109312. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7116488/.
Pekin, B.K., Jung, J., Villanueva-Rivera, L.J., Pijanowski, B.C., Ahumada, J.A., 2012. Modeling acoustic diversity using soundscape recordings and LIDAR-derived metrics of vertical forest structure in a neotropical rainforest. Landsc. Ecol. 27 (10), 1513–1522. https://doi.org/10.1007/s10980-012-9806-4.
Pieretti, N., Farina, A., Morri, D., 2011. A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). Ecol. Indic. 11 (3), 868–873. https://doi.org/10.1016/j.ecolind.2010.11.005. URL https://www.sciencedirect.com/science/article/pii/S1470160X10002037.
Pijanowski, B.C., Farina, A., Gage, S.H., Dumyahn, S.L., Krause, B.L., 2011a. What is soundscape ecology? An introduction and overview of an emerging new science. Landsc. Ecol. 26, 1213–1232.
Pijanowski, B.C., Villanueva-Rivera, L.J., Dumyahn, S.L., Farina, A., Krause, B.L., Napoletano, B.M., Gage, S.H., Pieretti, N., 2011b. Soundscape ecology: the science of sound in the landscape. BioScience 61 (3), 203–216 arXiv:https://academic.oup.com/bioscience/article-pdf/61/3/203/19404645/61-3-203.pdf.
Qi, J., Gage, S.H., Joo, W., Napoletano, B.M., Biswas, S., 2008. Soundscape Characteristics of an Environment: A New Ecological indicator of Ecosystem Health.
A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training.
Rowe, B., Eichinski, P., Zhang, J., Roe, P., 2021. Acoustic auto-encoders for biodiversity assessment. Eco. Inform. 62, 101237. https://doi.org/10.1016/j.ecoinf.2021.101237. URL https://www.sciencedirect.com/science/article/pii/S1574954121000285.
Stephenson, P.J., 2020. Technological advances in biodiversity monitoring: applicability, opportunities and challenges. Curr. Opin. Environ. Sustain. 45, 36–41. https://doi.org/10.1016/j.cosust.2020.08.005. URL https://www.sciencedirect.com/science/article/pii/S1877343520300592.
Sueur, J., Pavoine, S., Hamerlynck, O., Duvail, S., 2009. Rapid acoustic survey for biodiversity appraisal. PLoS One 3 (12), 1–9. https://doi.org/10.1371/journal.pone.0004065.
Sueur, J., Farina, A., Gasc, A., Pieretti, N., Pavoine, S., 2014. Acoustic indices for biodiversity assessment and landscape investigation. Acta Acust. Acust. 100, 772–781.
Sun, Y., Midori Maeda, T., Solís-Lemus, C., Pimentel-Alarcón, D., Buřivalová, Z., 2022. Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation. Ecol. Indic. 145, 109621. https://doi.org/10.1016/j.ecolind.2022.109621. URL https://www.sciencedirect.com/science/article/pii/S1470160X22010949.
Thomas, M., Jensen, F.H., Averly, B., Demartsev, V., Manser, M.B., Sainburg, T., Roch, M. A., Strandburg-Peshkin, A., 2021. A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations, bioRxiv. https://doi.org/10.1101/2021.12.16.472881 arXiv:https://www.biorxiv.org/content/early/2021/12/17/2021.12.16.472881.full.pdf. URL https://www.biorxiv.org/content/early/2021/12/17/2021.12.16.472881.
Ulloa, J.S., Haupert, S., Latorre, J.F., Aubin, T., J., 2021. Sueur, scikit-maad: an open-source and modular toolbox for quantitative soundscape analysis in Python. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13711, 2041–210X.13711. URL https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13711.
Walz, Y., Janzen, S., Narvaez, L., Ortiz-Vargas, A., Woelki, J., Doswald, N., Sebesvari, Z., 2021. Disaster-related losses of ecosystems and their services. Why and how do losses matter for disaster risk reduction? Int. J. Disast. Risk Reduct. 63, 102425. https://doi.org/10.1016/j.ijdrr.2021.102425. URL https://www.sciencedirect.com/science/article/pii/S2212420921003861.
Wickramasinghe, D., 2021. Ecosystem-based disaster risk reduction. In: Oxford Research Encyclopedia of Natural Hazard Science. https://doi.org/10.1093/acrefore/9780199389407.013.360.
Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R.J., Wilson, K., Hershey, J.R., 2020. Unsupervised Sound Separation using Mixture Invariant Training. arXiv:2006.12701.
Wu, C.-J., Brooks, D., Chen, K., Chen, D., Choudhury, S., Dukhan, M., Hazelwood, K., Isaac, E., Jia, Y., Jia, B., Leyvand, T., Lu, H., Lu, Y., Qiao, L., Reagen, B., Spisak, J., Sun, F., Tulloch, A., Vajda, P., Wang, X., Wang, Y., Wasti, B., Wu, Y., Xian, R., Yoo, S., Zhang, P., 2019. Machine Learning at Facebook: Understanding Inference at the Edge. In: 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), IEEE, Washington, DC, USA, pp. 331–344. https://doi.org/10.1109/HPCA.2019.00048. URL https://ieeexplore.ieee.org/document/8675201/.