scientific reports

Check for updates

OPEN

Species-independent analysis and identification of emotional animal vocalizations

Stavros Ntalampiras[1,2]✉

Animal vocalizations can differ depending on the context in which they are produced and serve as an instant indicator of an animal's emotional state. Interestingly, from an evolutional perspective, it should be possible to directly compare different species using the same set of acoustic markers. This paper proposes a deep neural network architecture for analysing and recognizing vocalizations representing positive and negative emotional states. Understanding these vocalizations is critical for advancing animal health and welfare, a subject of growing importance due to its ethical, environmental, economic, and public health implications. To this end, a framework assessing the relationships between vocalizations was developed. Towards keeping all potentially relevant audio content, the constructed framework operates on log-Mel spectrograms. Similarities/dissimilarities are learned by a suitably designed Siamese Neural Network composed of convolutional layers. The formed latent space is appropriately clustered to identify the support set facilitating the emotion classification task. We employed a publicly available dataset and followed a thorough experimental protocol. The efficacy of such a scheme is shown after extensive experiments considering both classification and support set selection. Last but not least, by elaborating collectively the network's activations when processing positive and negative vocalizations, important differences in the time-frequency plane are evidenced across emotions and species, assisting their understanding from animal scientists.

Keywords Animal health and welfare, Audio pattern recognition, Spectral clustering, Latent representation

In recent decades, public interest in animal welfare has increased steadily, prompting farmers to meet progressively higher welfare standards[1,2]. Animal welfare encompasses various factors, including appropriate housing, nutrition, disease prevention and treatment, and responsible care. Although there is no universally accepted definition of animal welfare, the importance of affective aspects, including mood and emotions, is widely recognized[3,4]. Affective states are typically described as short-lived, yet intense responses triggered by specific objects or events[5]. These states are of paramount importance to all species as they are closely related to behavioral decisions that aim to ensure survival and reproduction by seeking rewards and avoiding punishments[6]. However, existing research lacks approaches that develop or model positive and negative emotional states of animals, let alone incorporate such tools into on-site evaluation protocols[7–10].

That said, in recent years, the use of AI-based techniques to analyze and understand animal vocalizations has gained significant attention[11–14]. These studies serve various purposes, including industrial applications such as monitoring animals on farms and supporting veterinarians in diagnosing animal diseases[15]. Numerous algorithms and tools have been developed to facilitate research in this field, such as the BIRDNET app, which specializes in identifying bird species based on their vocalizations[16]. However, inter-species analysis of emotional animal vocalizations based on machine learning technologies remains largely unexplored in the existing literature. Current research considers a limited range of species-specific vocalizations. Ruiz-Miranda et al.[17] examined the physical characteristics of domestic goat vocalizations in response to the cries of their offspring. Briefer et al.[18] investigated how the social environment and kinship influence contact calls during the early development of goats. Furthermore, Baciadonna et al.[19] demonstrated that goats can distinguish between vocalizations related to positive and negative emotions, suggesting the potential for their use in animal welfare monitoring[20]. Interestingly, the closest paper to this work is presented in[21] and proposes the use of handcrafted features along with traditional machine learning approaches to classify emotional states across various species. The respective dataset includes 7 species, while the balanced accuracy across positive and negative states reached by decision tree ensemble learning classifier (XGBoost) is 83.9%.

[1]Department of Computer Science, University of Milan, via Celoria 18, 20133 Milan, Italy. [2]Research Center on AI for Animal Health and Welfare, University of Milan, Via Festa del Perdono 7, 20122 Milan, Italy. ✉email: stavros.ntalampiras@unimi.it

Towards deepening our understanding of emotional animal vocalizations, this work aims at determining whether there are notable time-frequency differences in positive and negative states, highlight them, and use them for automatic classification. To this end, we designed a pipeline elaborating on standardized log-Mel frequency spectrograms with the aim being to reveal potential similarities/dissimilarities. As such, we designed a Siamese Neural Network learning such relationships when presented with pairs of animal vocalizations. Interestingly, such a topology is able to operate in small and imbalanced data environments, given that such datasets are scarce. Then, this work proposes to employ spectral clustering in the latent space learned by the SNN to determine the support set to be used for identifying the class of unknown vocalizations. Last but not least, the available vocalizations were collectively analyzed via the SNN's activations emphasizing the relevant regions of the employed time-frequency representations. It should be mentioned that activation maps comprise a visualization technique used to identify the regions of a spectrogram that play an important role for a neural network to make a specific prediction[22].

Importantly, the present work employs a publicly available dataset[21] and adopts a reliable experimental protocol favouring open and reproducible research practices.

The novel points of this work are

- the introduction of a relationship learning paradigm in the field able to handle imbalanced classes,
- the suitable scheme for selecting the support set, i.e. the most representative samples serving classification,
- a straightforward visualization of the time-frequency content responsible for the emotional state's prediction using activation maps,
- a collective analysis of the spectral content across species and emotional states,
- the interpretability module which is able to meaningfully assist animal scientists in understanding the model's prediction and interacting with it.

This work is organized as follows: section Results analyses the obtained results, while section Interpretation of the learned audio structurefeatures presents the conducted interpretation analysis of the learned audio structures. Subsequently, section Methods describes the proposed methodology and section Conclusion outlines the main findings of this work.

Methods

This section explains the processes regarding a) feature extraction, b) construction of the SNN, and c) the spectral clustering algorithm operating in the latent space.

Log-Mel spectrogram

To derive the features, the signal is divided into overlapping frames, and the power of the Short-Time Fourier Transform (STFT) is calculated for each frame. The resulting representation is then processed using a Mel-scale filter bank emphasizing the frequency bands that are most notable for human perception.

For parameterization, the signals are sampled at 16 kHz and divided into windows of length of 30ms and an overlap of 20ms between consecutive frames, while 64 Mel-bands were considered. This approach preserves the full time-frequency representation, making it suitable for use by the CNN, as it maintains spatial relationships unlike Mel-Frequency Cepstral Coefficients (MFCCs), aligning well with the objectives of the present analysis.

Siamese neural network learning

Building on the findings from our earlier research[23,24], we developed a Siamese Neural Network (SNN) based representation to tackle the problem at hand. SNNs have demonstrated their effectiveness in capturing relationships, especially similarities and differences, among various audio signals. As a result, they can be employed indirectly to address classification and/or regression tasks.

An SNN consists of two symmetric neural networks, often called twins, that process distinct input signals and converge at a shared endpoint[25]. This endpoint computes a predefined distance metric using the abstract representations generated by each network. By optimizing a shared objective function, the networks develop interconnected weights, producing aligned high-level representations for similar inputs. The symmetry in their architecture ensures that reversing the networks or their inputs does not affect the output, highlighting the model's robustness and flexibility.

Training the SNN with balanced pairs of time-frequency representations, representing both similarities and dissimilarities, naturally mitigates class imbalance characterizing the present task. This approach ensures robust and equitable representation learning, enhancing the SNN's ability to handle real-world scenarios with varying degrees of similarity and dissimilarity in audio signals.

Model composition and parameterization

Given the effectiveness of Convolutional Neural Networks (CNNs) in audio pattern recognition tasks[26,27], each twin is designed with a series of layered convolutional networks, each paired with a corresponding max-pooling layer, as shown in Fig. 1. Each hidden unit in the network is associated with a specific region of the input signal, referred to as its *receptive field*. The unit's weights generate feature maps that highlight the unique characteristics of that field. This structure allows each hidden unit to concentrate on and process localized information, improving the network's ability to identify and extract relevant features from different parts of the input signal. To manage the potentially high dimensionality of these outputs, pooling layers are incorporated to retain only the maximum value within defined regions.

Following a grid search and validation set analysis to optimize the hyperparameters (number of layers, kernel and stride size, and learning rate), the SNN is configured with three convolutional layers. Each of these layers is

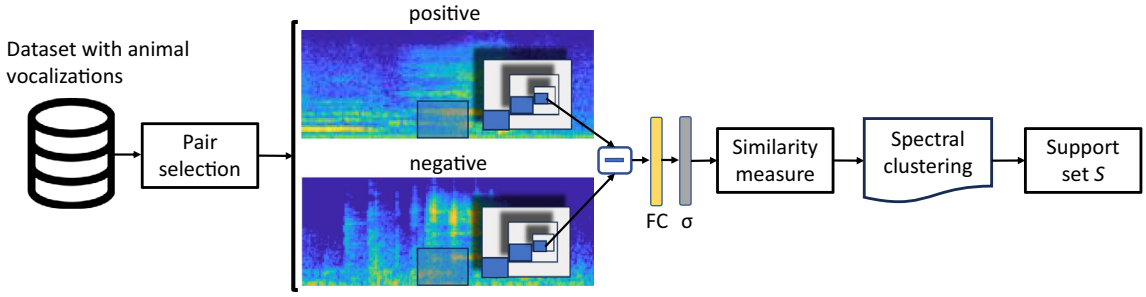

Fig. 1. The implemented pipeline starting with the log-Mel spectrogram representation of animal vocalizations, the Siamese Neural Network for similarity assessments and spectral clustering for support set selection.

followed by a rectified linear unit (ReLU) activation function ($f(x) = \max(0, x)$) and a corresponding max-pooling layer.

The final layer of the network is a fully connected (FC) one, where the SNN evaluates the distance between the outputs of the twin components using binary cross-entropy loss. This loss is normalized to the [0,1] range using a sigmoid σ function, thus characterizing the relationship between the input pair. The resulting loss value can then be thresholded to determine whether the similarity level between the pair of input vocalizations.

The convolutional layers utilize filters of varying sizes, all with a stride of one, while the max-pooling layers employ 2×2 kernels with a stride of two. The learning process follows the standard back-propagation algorithm, updating weights for each twin by summing gradients. The batch size is fixed at 50, and the learning rate is set to 6e-5. Weights and biases are initialized using narrow normal distributions with parameters $\mu = 0$ and $\sigma = 0.01$. Training is capped at a maximum of 3000 iterations, while an early-stopping criterion is applied.

Input: SNN's similarity matrix s, number of clusters;
Output: composition of each cluster in the latent space, support set $\mathcal{S}$;
1. Use s to calculate the unnormalized Laplacian matrix $\mathcal{L}$ and the normalized random-walk Laplacian matrix $\mathcal{L}_{rw}$ as in [28];
2. Create a matrix $\mathcal{V}$ containing columns $v_1, \ldots, v_k$, where the columns are the k eigenvectors that correspond to the k smallest eigenvalues of $\mathcal{L}$;
3. Treat each row of $\mathcal{V}$ as a point, cluster the n points using k-means;
4. Assign the latent space points to the same clusters as their corresponding rows in $\mathcal{V}$;
5. $\mathcal{S}$ is composed of the centers of each cluster.

Algorithm 1. The proposed latent space clustering algorithm.

Latent spectral clustering

After training the SNN, we define a set of vocalizations $\mathcal{S}$ which is going to support the decision-making process. As such, during testing, the SNN assesses the similarity between the unknown vocalization and each element of $\mathcal{S}$, and the prediction is the class associated with the highest similarity. There are several approaches in the literature for determining the composition of $\mathcal{S}$. A straightforward way is to use the full training set and compute the average similarities[29], while other approaches cluster the latent space based on k-means and identify the most centric samples of each class which then populate $\mathcal{S}$[30]. In addition, there are works clustering the latent space of deep architectures, where the aim typically is unsupervised feature learning and/or anomaly detection[31,32]. Unfortunately, such approaches are not suited for the specific classification problem given the imbalanced data across classes and arbitrarily shaped classes as represented in the latent space. In the present case, we have available knowledge regarding the number and composition of the considered classes, and we are searching for the optimal set to support the classification task.

Thus, we propose to employ spectral clustering to determine the composition of $\mathcal{S}$. Spectral clustering is a graph-based method for identifying k arbitrarily shaped clusters within a dataset[33]. The algorithm represents the data in a reduced-dimensional space where clusters are more distinctly separated, allowing for the application of conventional clustering techniques like k-means. Such a lower-dimensional representation is derived from the eigenvectors associated with the k smallest eigenvalues of a Laplacian matrix. The Laplacian matrix is a representation of a similarity graph, which encodes the local neighbourhood relationships between data points as an undirected graph.

Typically, the algorithm begins by constructing a similarity matrix from the data to form the similarity graph, computes the Laplacian matrix, and then identifies k eigenvectors from the Laplacian matrix. These eigenvectors are used to partition the graph into k clusters. Importantly, spectral clustering is particularly effective when the number of clusters is known beforehand.

The designed algorithm is based on the normalized random-walk Laplacian matrix using the method described by Shi-Malik[28] while employing the SNN's output to form the similarity graph (Fig. 2, Alg. 1, line 1). There, we see two distinct clusters even though several regions appear sparse partially due to the differences existing across the considered species. Then, matrix $\mathcal{V}$ is created, the columns of which are the k eigenvectors that correspond to the k smallest eigenvalues of $\mathcal{L}$ (Alg. 1, line 2). Each row of $\mathcal{V}$ is treated as a point and clustered using k-means (Alg. 1, line 3). Finally, the latent space points are assigned to the same clusters as their respective rows in $\mathcal{V}$ (Alg. 1, line 4) and $\mathcal{S}$ is composed of the centers of each cluster (Alg. 1, line 5). The class (positive/negative) represented by each cluster is determined based on the centers' similarity to the known training data, as assessed by the SNN. This ensures that each cluster is anchored to meaningful semantic information derived from the labelled examples. It should be mentioned that the pipeline is fully implemented in a MATLAB environment, while a block diagram of the algorithm is provided in Figure S1.

Results

This section presents the results achieved by the proposed as well as contrasted approaches in classifying positive vs. negative emotional animal vocalizations. It should be mentioned that a 5-fold cross validation experimental protocol was adopted, ensuring that calls from the same individual were exclusively allocated to either the training or testing sets. Such is set-up provides reliable results as the test set is not manually selected to maximize the achieved recognition rate[21]. Moreover, in this work, the same model is used across all considered species.

The available dataset includes contact calls of seven ungulate species, i.e. 1. cows (Bos taurus), 2. goats (Capra hircus), 3. horses (Equus caballus), 4. Przewalski's horses (Equus przewalskii), 5. pigs (Sus scrofa domesticus), 6. wild boars (Sus scrofa), and 7. sheep (Ovis aries).

Information regarding the species, sex, age, maintenance, and care of the animals whose contact calls were utilized in this study can be found in[34–38]. A total of 3,181 contact calls from seven species were analyzed. The sampling frequency is 44.1 kHz. These vocalizations conveyed emotions classified as either positive or negative valence, based on factors such as the context of production (e.g., social interactions vs. isolation), behavioral indicators (e.g., postural adjustments and movement patterns), and physiological measures in domestic species. Each call served as a distinct data point within the dataset, which was categorized by species, context, and valence for analysis. Interestingly, the dataset is made entirely of contact calls produced in negative versus positive situations, i.e. the same call type produced in situations of opposite valence. Therefore, acoustic differences between valence are expected to be rather minimal (unlike when comparing e.g. affiliative calls and alarm/distress calls). Fig. 3 demonstrates the dataset composition, where we observe that it is highly imbalanced not only across emotional states but also across species. Overall there are 2508 negative and 673 positive vocalizations. More detailed information regarding the dataset and its availability can be found in[21].

The obtained results are provided in Fig. 4. We observe that the proposed scheme based on spectral clustering in the latent space outperforms the contrasted approaches reaching an average balanced accuracy of 84.1%. This is in line with the results reported in[21] (83.9%). However, the present results are far more balanced across positive and negative emotional states, i.e. 88.2% and 80.1% vs. 94.2% and 70% for negative and positive vocalizations respectively. We argue that such promising result is a direct outcome of the proposed learning paradigm able to handle data imbalances in the training phase as it is fed with pairs of vocalizations. We also compared it with the SNN designed in[30] which reached a balanced accuracy of 78.6%. That approach considers the minimization of the intra-class distances and maximization of inter-class distances during the design phase. As such, it does not explicitly address arbitrarily-shaped clusters in the selection of $\mathcal{S}$. We argue that this is a critical point given the

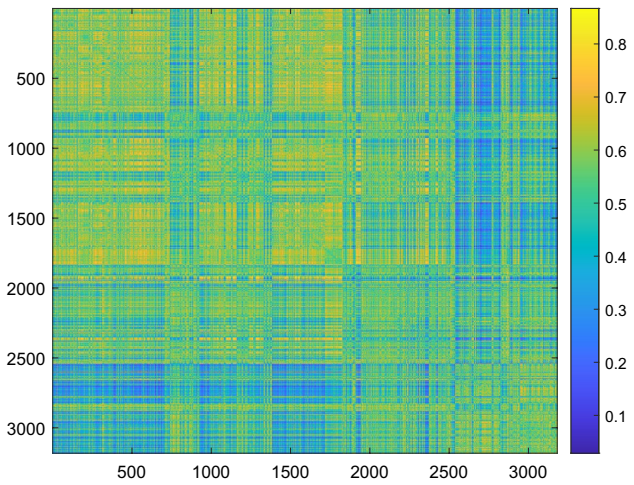

Fig. 2. The similarity matrix s considering all pairs of vocalizations as assessed by the SNN.

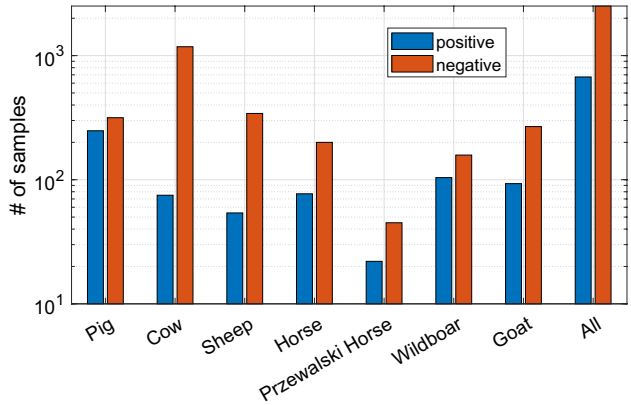

Fig. 3. Dataset composition across the considered species.

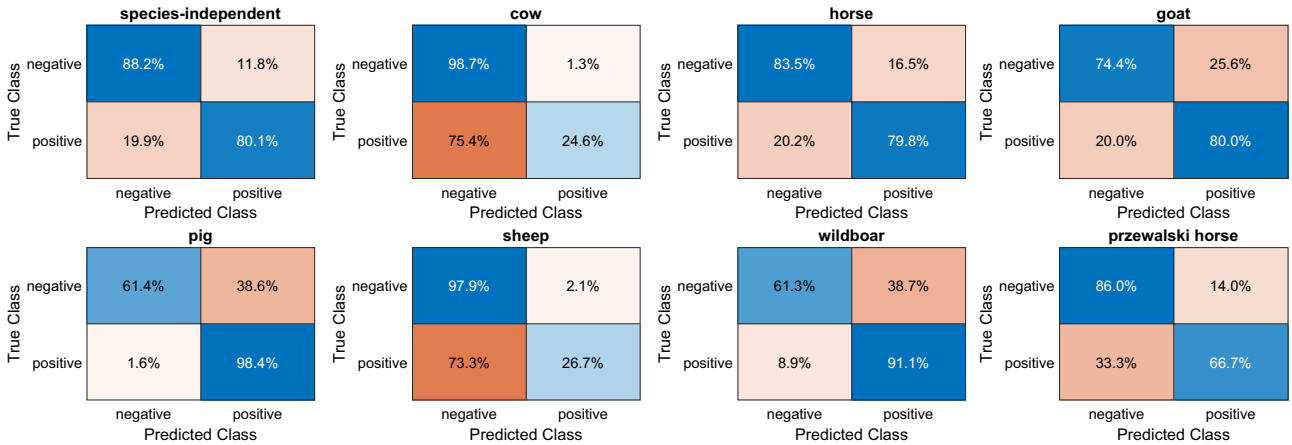

Fig. 4. Species-independent and species-depended confusion matrices achieved by the proposed approach.

imbalanced classes which are further characterized by different densities in the problem at hand. Utilizing the full training set for support during the classification phase does not provide satisfactory accuracy (68.5%) given the characteristics of the available dataset.

Even though approaches based on transfer learning may provide satisfactory classification performance in related tasks[39–41], they are not fully suitable in the task at hand due to the poor explainability of the constructed model, which may be biased by the original training dataset[14]. This limitation can lead to misleading predictions and reduced trustworthiness, especially in the present application scenario where interpretability and data representation are essential. As such, they were not considered in this work given that the aim is a systematic analysis of the frequency content characterizing emotional animal vocalizations. Including such methods could have introduced confounding factors unrelated to the acoustic properties of the involved vocalizations, thereby compromising the clarity of the analysis.

In general, negative vocalizations are recognized with higher accuracies with respect to positive ones. This may be due to the sparse distribution of the respective vocalizations in the log-Mel and latent spaces, which burdens the modeling phase. In addition, there is a higher availability of negative calls, meaning that the proposed spectral clustering approach is better equipped to suitably select the most representative samples to populate the support set facilitating the classification process.

Figure 4 includes the species-specific confusion matrices as well while Table 1 the corresponding precision, recall and F1-scores. Overall, the results are promising; horses, pigs, goats, Przewalski's horses , and wild boars have a balanced accuracy above 75% demonstrating a clearer distinction between valence categories in these species. Cows and sheep showed worse performances (62%) indicating a considerable overlap in time-frequency patterns, burdening valence classification. These results are in line with[21], where species dependent models have been constructed.

While certain time-frequency regions include features able to reliable predict emotional valence across different species, these findings indicate that there is also notable variability at the species level in how emotional states are conveyed through vocalizations. This suggests that while some vocal markers of emotion may be broadly shared, the specific ways in which different species express their emotions through sound can vary evidently.

Species	Precision	Recall	F1-score
Cow	56.7	98.7	72
Horse	80.5	83.6	82
Goat	78.8	74.5	76.6
Pig	97.4	61.4	75.3
Sheep	57.2	97.9	72.2
Wildboar	87.3	61.3	72
Przewalskihorse	72.1	86.1	78.5
Species-independent	**81.6**	**88.2**	**84.8**

Table 1. Precision, Recall and F1-scores (in %) for each species and in the species-independent case

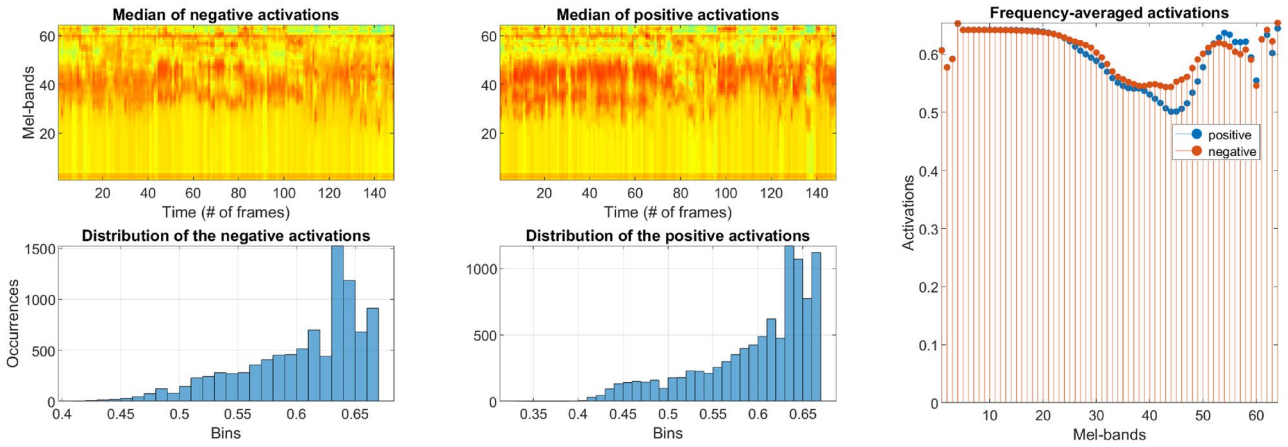

Fig. 5. The median of SNN's activation maps per vocalization class along with the distribution of values when grouping positive and negative classes.

Overall, the obtained performances are more than encouraging given the complexity and the perceptual nature of the specific problem, while employing a time-frequency representation free of domain knowledge. As such, we further elaborated the SNN's latent space to investigate the audio content which distinguishes each class of vocalizations across species.

Interpretation of the learned audio structure features

Following the encouraging results obtained when considering the latent space, we analyzed the available vocalizations when processed by the SNN via the corresponding activations. More precisely, we calculated the median activation maps per class as well the distribution of the values across positive and negative vocalizations. Fig. 5 illustrates the average SNN's activation maps per vocalization class along with the distribution of the respective values when grouping positive and negative classes.

There, we observe that when processing negative emotional states, the network places emphasis on specific mid to high frequency regions evidenced in the activation maps. Furthermore, the negative class includes notably more regions associated with high values as shown in the respective histograms and examining the number of occurrences. This finding suggests a relevant consideration of the selected time-frequency patterns during modeling similar/dissimilar relationships.

As regards to positive vocalizations, there is a wider focus in the respective activation maps as the spectrogram is considered in a relatively more uniform way, while mid to high frequencies are highlighted here as well. At the same time, high frequencies are associated with lower activation values with respect to the high frequencies of the negative class. Such differences become particularity clear when examining the frequency averaged activations.

As we observed in the emotion classification results, the model processes log-Mel spectrogram in a species-specific way. Fig. 6 illustrates the frequency-averaged activations with respect to the considered species when processing positive and negative vocalizations. There, we do not observe evident differences in the initial Mel-bands; the situation changes drastically in mid to high frequency bands where each species' vocalizations are considered in a very different way by the SNN. For example, high frequency regions are relevant when processing pig vocalizations, while mid frequency regions appear to play a relevant role in sheep and horses. In addition, the SNN focuses on high frequencies when processing wild boar and Przewalski horse vocalizations. Overall, we observe that the SNN processes the input log-Mel spectrograms in a species-specific manner varying the emphasis it places upon the time-frequency plane. The presented time-frequency regions are the ones facilitating

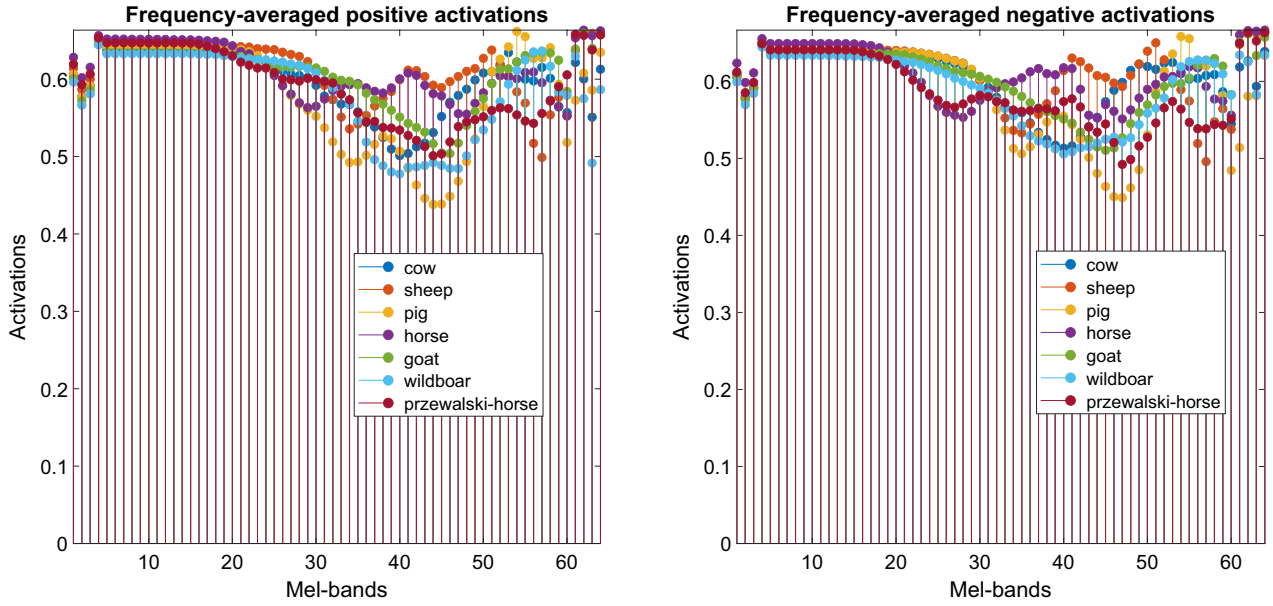

Fig. 6. The species-specific frequency-averaged activations for positive and negative vocalizations.

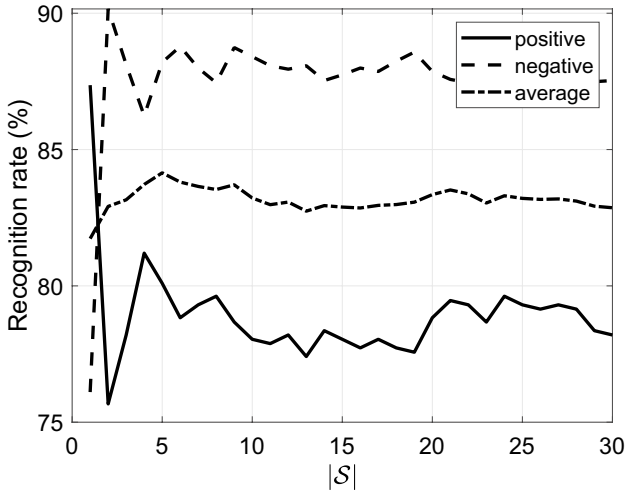

Fig. 7. The recognition rates as a function of the cardinality of the support set.

the network in learning and quantifying similar and dissimilar relationships in emotional animal vocalizations, thus obtaining satisfactory recognition rates.

Composition of $\mathcal{S}$

This part of the analysis of the obtained results is dedicating in the support set $\mathcal{S}$ serving the classification purposes. Those are the calls which are considered to be the most representative ones of each class according to the proposed spectral clustering approach. Figure 7 shows the recognition rates per emotional class as well as the average one as the size of $\mathcal{S}$ increases. We observe that very small values provide higher rate for positive calls while this observation changes as the size increases. The highest average performance is observed when $|\mathcal{S}| = 5$, i.e. when 5 samples of each class (positive and negative) are considered. The positive samples in $\mathcal{S}$ come from *Przewalski Horse*, *pig* and *sheep* species, while the negative from *goat*, *cow* and *sheep*.

Interestingly, those are the species' calls which can effectively represent the considered emotional classes and offer high recognition rates in a species-independent setting. We can further observe that different species represent the different emotional classes except for sheep which is the only species existing in both. Moreover, 3 out of 5 positive calls come from pigs meaning that the specific vocalizations can well represent the given emotional state. Similarly, cow calls compose 3/5 of the negative part showing that they play an important role in that state according to the presented spectral clustering scheme operating in the latent space formed by the SNN.

Facilitating the analysis made by animal experts

The interpretation process outlined in Section Interpretation of the learned audio structurefeatures can offer valuable insights into the functioning of the proposed model and the execution of its predictions. The next step involves utilizing this information by animal scientists, thereby completing the usability loop. Notably, the implemented framework enables animal scientists to engage with the model, providing them with insights into the reasoning behind each prediction. This interactive feature helps experts familiarize themselves with the AI-based tool, addressing a notable gap in the literature[42–44] related to building trust.

During this phase, the proposed framework addresses four key types of inquiries:

- highlighting the spectrogram regions that have the greatest influence on a given decision,
- assessing similarities between a-priori known animal vocalizations and novel ones, so as to detect related/unrelated ones,
- sonifying such similarities/differences in the most relevant time-frequency content between two vocalizations.

To answer the first question, the framework delivers a detailed response by following these steps: a) generating and displaying the activation map associated with the specific prediction, b) emphasizing regions of increased significance within the map, c) identifying the corresponding content along the time-frequency axes, and d) sonifying the identified content, enabling human experts to listen exclusively to that specific segment.

For the second and third questions, the framework simply feeds the network with the calls of interest and quantifies it similarity. Interestingly, such a similarity score may facilitate the analysis carried out by animal scientists, offer insights and boost the understanding of similar and dissimilar vocalizations.

We conclude that this interactive system has the potential to significantly support animal scientists in effectively leveraging an AI-driven solution. Its primary objective is to offer clear and easily interpretable assistance to human experts. By adopting this transparent approach, the framework ensures a deep understanding of the factors influencing the final prediction. Lastly, it is important to highlight that this module aligns with the recently introduced regulations for AI-based systems outlined in the EU's AI Act (https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence).

Conclusion

This article addressed the problem of classifying positive vs. negative emotional animal vocalizations as well as revealing differences in the respective time-frequency contents. The method is based on spectral clustering which is able to suitably group the available classes in the latent space while considering pair-wise similarities as assessed by means of a Siamese Neural Network elaborating on standardized audio representation. Interestingly, the proposed method outperforms the contrasted approaches and sheds light in the spectral components which play an important role during such a classification task. Importantly, the obtained balanced recognition rate is quite high in a species independent setting. It was shown that specific time-frequency contents are emphasized when processing different emotional vocalizations across the considered species.

The results of this study may be useful to a) build physical models representing animal vocalizations, b) develop interpretable-by-design classification schemes, c) create didactic material for young animal scientists, students, veterinarians, technicians and farmers, d) compare the present model to how humans judged the valence in the calls of these same species[45], e) monitor animals' stress levels under various farm management scenarios, and f) serve the analysis and modelling of additional species and emotional states as long as those data become available, potentially also by employing data augmentation techniques.

Data availability

The dataset used in this work is publicly available at https://www.sciencedirect.com/science/article/pii/S2589004222500094X

Received: 28 March 2025; Accepted: 30 July 2025
Published online: 06 August 2025

References

1. Laurijs, K. A., Briefer, E. F., Reimert, I. & Webb, L. E. *Vocalisations in farm animals: A step towards positive welfare assessment* **236**, 105264. https://doi.org/10.1016/j.applanim.2021.105264 (2021).
2. Cornish, A., Raubenheimer, D. & McGreevy, P. *What we know about the public's level of concern for farm animal welfare in food production in developed countries* **6**(11), 74. https://doi.org/10.3390/ani6110074 (2016).
3. Green, T. & Mellor, D. Extending ideas about animal welfare assessment to include 'quality of life' and related concepts. *Nature* **59**(6), 263–271. https://doi.org/10.1080/00480169.2011.610283 (2011).
4. Webb, L. E., Veenhoven, R., Harfeld, J. L. & Jensen, M. B. What is animal happiness?. **1438**(1), 62–76. https://doi.org/10.1111/nyas.13983 (2018).
5. Paul, E. S. & Mendl, M. T. *Animal emotion: Descriptive and prescriptive definitions and their implications for a comparative perspective* **205**, 202–209. https://doi.org/10.1016/j.applanim.2018.01.008 (2018).
6. Kremer, L., Holkenborg, S. E. J. K., Reimert, I., Bolhuis, J. E. & Webb, L. E. *The nuts and bolts of animal emotion* **113**, 273–286. https://doi.org/10.1016/j.neubiorev.2020.01.028 (2020).
7. Andonovic, I., Michie, C., Cousin, P., Janati, A., Pham, C. & Diop, M. Precision livestock farming technologies. In: 2018 Global Internet of Things Summit (GIoTS), 1–6 (2018). https://doi.org/10.1109/GIOTS.2018.8534572
8. Ntalampiras, S., Pezzuolo, A., Mattiello, S., Battini, M. & Brščic, M. Automatic detection of cow/calf vocalizations in free-stall barn. In: 2020 43rd TSP, 41–45 (2020). https://doi.org/10.1109/TSP49548.2020.9163522
9. Ntalampiras, S. On acoustic monitoring of farm environments. In: Communications in Computer and Information Science, 53–63. Springer, ??? (2019). https://doi.org/10.1007/978-981-13-5758-9_5

10. Ahmed, N., De, D. & Hussain, I. Internet of things (iot) for smart precision agriculture and farming in rural areas. *IEEE Internet of Things Journal* **5**(6), 4890–4899. https://doi.org/10.1109/JIOT.2018.2879579 (2018).
11. Ntalampiras, S. et al. Automatic classification of cat vocalizations emitted in different contexts. *Animals* **9**(8), 543. https://doi.org/10.3390/ani9080543 (2019).
12. Kukushkin, M. & Ntalampiras, S. Automatic acoustic classification of feline sex. In: Proceedings of the 16th International Audio Mostly Conference. AM '21, 156–160. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3478384.3478385 .
13. Acconcjaioco, M. & Ntalampiras, S. One-shot learning for acoustic identification of bird species in non-stationary environments. In: 2020 25th International Conference on Pattern Recognition (ICPR), 755–762 (2021). https://doi.org/10.1109/ICPR48806.2021.9412005
14. Ntalampiras, S. & Pesando Gamacchio, G. Explainable classification of goat vocalizations using convolutional neural networks. *PLOS ONE* **20**(4), 0318543. https://doi.org/10.1371/journal.pone.0318543 (2025).
15. Carroll, B. T., Anderson, D. V., Daley, W., Harbert, S., Britton, D. F. & Jackwood, M. W. Detecting symptoms of diseases in poultry through audio signal processing. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 1132–1135 (2014). IEEE
16. Denton, T., Wisdom, S. & Hershey, J. R. Improving bird classification with unsupervised sound separation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 636–640 (2022). IEEE
17. Ruiz-Miranda, C., Szymanski, M. D. & Ingals, J. W. Physical characteristics of the vocalization of domestic goat does capra hircus in response to their offspring's cries. *Bioacoustics* **5**(1–2), 99–116. https://doi.org/10.1080/09524622.1993.9753232 (1993).
18. Briefer, E. F. Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology* **288**(1), 1–20. https://doi.org/10.1111/j.1469-7998.2012.00920.x (2012).
19. Baciadonna, L., Briefer, E. F., Favaro, L. & McElligott, A. G. Goats distinguish between positive and negative emotion-linked vocalisations **16**(1) (2019) https://doi.org/10.1186/s12983-019-0323-z
20. Ntalampiras, S. et al. An integrated system for the acoustic monitoring of goat farms. *Ecological Informatics* **75**, 102043. https://doi.org/10.1016/j.ecoinf.2023.102043 (2023).
21. Lefèvre, R. A., Sypherd, C. C. R. & Briefer, îF. Machine learning algorithms can predict emotional valence across ungulate vocalizations. *iScience* **28**(2), 111834. https://doi.org/10.1016/j.isci.2025.111834 (2025).
22. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
23. Ntalampiras, S. One-shot learning for acoustic diagnosis of industrial machines. *Expert Systems with Applications* **178**, 114984. https://doi.org/10.1016/j.eswa.2021.114984 (2021).
24. Ntalampiras, S. Speech emotion recognition via learning analogies. *Pattern Recognition Letters* **144**, 21–26. https://doi.org/10.1016/j.patrec.2021.01.018 (2021).
25. Ntalampiras, S. Explainable siamese neural network for classifying pediatric respiratory sounds. *IEEE Journal of Biomedical and Health Informatics* **27**(10), 4728–4735. https://doi.org/10.1109/JBHI.2023.3299341 (2023).
26. Purwins, H. et al. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing* **13**(2), 206–219. https://doi.org/10.1109/JSTSP.2019.2908700 (2019).
27. Srivastava, S., Wu, H.-H., Rulff, J., Fuentes, M., Cartwright, M., Silva, C., Arora, A. & Bello, J. P. A study on robustness to perturbations for representations of environmental sound. In: 2022 30th European Signal Processing Conference (EUSIPCO), 125–129 (2022). https://doi.org/10.23919/EUSIPCO55093.2022.9909557
28. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(8), 888–905. https://doi.org/10.1109/34.868688 (2000).
29. Ntalampiras, S. One-shot learning for acoustic diagnosis of industrial machines. *Expert Systems with Applications* **178**, 114984. https://doi.org/10.1016/j.eswa.2021.114984 (2021).
30. Ntalampiras, S. & Qi, W. Siamese neural network for speech-based depression classification and severity assessment. *Journal of Healthcare Informatics Research* **8**(4), 577–593. https://doi.org/10.1007/s41666-024-00175-4 (2024).
31. Cao, W. et al. Unsupervised discriminative feature learning via finding a clustering-friendly embedding space. *Pattern Recognition* **129**, 108768. https://doi.org/10.1016/j.patcog.2022.108768 (2022).
32. Pinon, N., Trombetta, R. & Lartizien, C. One-class svm on siamese neural network latent space for unsupervised anomaly detection on brain mri white matter hyperintensities. In: Oguz, I., Noble, J., Li, X., Styner, M., Baumgartner, C., Rusu, M., Heinmann, T., Kontos, D., Landman, B., Dawant, B. (eds.) Medical Imaging with Deep Learning. *Proceedings of Machine Learning Research* **227**, 1783–1797 (2024).
33. Fowlkes, C., Belongie, S., Chung, F. & Malik, J. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(2), 214–225. https://doi.org/10.1109/TPAMI.2004.1262185 (2004).
34. Briefer, E. F., Vizier, E., Gygax, L. & Hillmann, E. Expression of emotional valence in pig closed-mouth grunts: Involvement of both source- and filter-related parameters. *The Journal of the Acoustical Society of America* **145**(5), 2895–2908. https://doi.org/10.1121/1.5100612 (2019).
35. Maigrot, A.-L., Hillmann, E. & Briefer, E. F. Encoding of emotional valence in wild boar (sus scrofa) calls. *Animals* **8**(6), 85. https://doi.org/10.3390/ani8060085 (2018).
36. Briefer, E. F., Maigrot, A.-L., Mandel, R., Freymond, S. B., Bachmann, I. & Hillmann, E. Segregation of information about emotional arousal and valence in horse whinnies. Scientific Reports **5**(1) (2015) https://doi.org/10.1038/srep09989
37. Briefer, E. F., Tettamanti, F. & McElligott, A. G. Emotions in goats: mapping physiological, behavioural and vocal profiles. *Animal Behaviour* **99**, 131–143. https://doi.org/10.1016/j.anbehav.2014.11.002 (2015).
38. Maigrot, A.-L., Hillmann, E., Anne, C. & Briefer, E. F. Vocal expression of emotional valence in przewalski's horses (equus przewalskii). Scientific Reports **7**(1) (2017) https://doi.org/10.1038/s41598-017-09437-1
39. Ghani, B., Denton, T., Kahl, S. & Klinck, H. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Scientific Reports **13**(1) (2023) https://doi.org/10.1038/s41598-023-49989-z
40. Dufourq, E., Batist, C., Foquet, R. & Durbach, I. Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics* **70**, 101688. https://doi.org/10.1016/j.ecoinf.2022.101688 (2022).
41. Manriquez, P. R., Kotz, S. A., Ravignani, A. & Boer, B. Bioacoustic classification of a small dataset of mammalian vocalisations using deep learning. *Bioacoustics* **33**(4), 354–371. https://doi.org/10.1080/09524622.2024.2354468 (2024).
42. Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **32**(11), 4793–4813. https://doi.org/10.1109/tnnls.2020.3027314 (2021).
43. Zhang, Y., Tino, P., Leonardis, A. & Tang, K. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* **5**(5), 726–742. https://doi.org/10.1109/tetci.2021.3100641 (2021).
44. Bhaskhar, N., Rubin, D. L. & Lee-Messer, C. An explainable and actionable mistrust scoring framework for model monitoring. IEEE Transactions on Artificial Intelligence, 1–12 (2023) https://doi.org/10.1109/TAI.2023.3272876
45. Greenall, J. S., Cornu, L., Maigrot, A.-L., Torre, M. P. & Briefer, E. F. Age, empathy, familiarity, domestication and call features enhance human perception of animal emotion expressions. Royal Society Open Science **9**(12) (2022) https://doi.org/10.1098/rsos.221138

Acknowledgements
This work was supported by the project European Partnership for Animal Health and Welfare funded by the European Commission. We would like to thank the authors of[21] for making available the dataset used in this work.

Author contributions
S.N. carried out all the work associated with this article.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-14323-2.

Correspondence and requests for materials should be addressed to S.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.