Bats, Spectrograms, and Deep Learning: Unmasking the Secrets of Mining Area Caves

Arthur Gonsales[1], Vitor C. A. Santos[1], Giulliana Appel[1], Leonardo Trevelin[1], Valeria Tavares[1], Ronnie Alves[1]

[1]Instituto Tecnologico Vale, 955, 66055-090, Belém PA, BR

{arthur.gonsales, giulliana.appel}@pq.itv.org

{vitor.cirilo.santos, leonardo.trevelin}@itv.org

{valeria.tavares, ronnie.alves}@itv.org

Abstract. *Detecting and classifying bat species in mining region caves is crucial for ore extraction activities, environmental impact reduction, and worker safety. This study focuses on using bat echolocation call data collected by VALE, a mining company, to develop a deep learning-based system for species recognition. By applying transfer learning on a pre-trained MobileNetV2 model, the system achieved an accuracy of 95.21% in classifying spectrograms of bat echolocation calls from three different species. This outperformed other models tested. Implementing this system would enhance VALE's cave inspection processes, ensuring worker safety and bat population preservation in mining regions.*

1. Introduction

As members of a society that is experiencing the era of mining, it is necessary to understand how this type of operation can impact not only the lives of humans but also the lives of animals living in close proximity to mining areas, from which abundant resources are extracted. According to the secondary authors [Smith 2011, Mantz 2008], referenced by the original author [Jacka 2018], these resources range from copper, which illuminates our homes, to iron, which transforms our transportation systems, as well as various other minerals of the "digital age".

The exploitation of mineral reserves in Brazil is currently restricted due to the presence of caves in the regions. The vast majority of documented caves in the last four years (starting from 2014, totaling approximately 3,000 units) have been mapped through environmental studies conducted for mining operations [Auler and Pilo 2014].

Caves, mines, and other underground regions are critical for the survival of various species of animals, among which bats are particularly noteworthy. In hot regions like Brazil, these habitats support the survival of a rich variety of species and contribute to the maintenance of colonies that operate a substantial ecosystem of services [Voigt and Kingston 2016].

Large bat colonies inhabiting caves are particularly vulnerable to environmental disturbances, precisely because they aggregate in highly concentrated numbers in these locations. According to [Frick et al. 2020], these colonies are easily found by local communities and researchers, which results in intentional and unintentional disturbances to

these individuals. The extraction of minerals is one of the most common root causes of environmental impact on the lives of these species.

Currently, there is a high demand for robust and accurate tools for monitoring and identifying biodiversity in order to contribute to the reduction of environmental impacts on various populations [Mac Aodha et al. 2018].

Furthermore, according to [Mac Aodha et al. 2018], the monitoring of bat species is a good example of how studying their population dynamics serves not only for cataloging and managing species inhabiting caves in mining areas but also as indicators of ecosystem health as a whole, since they are sensitive to disturbances in this type of environment.

Bioacoustic monitoring is a non-invasive practice for collecting bioacoustic data on bat echolocation, both for assessing population dynamics and cataloging species [Jones et al. 2013]. Most current techniques for processing this type of data focus on classifying species based on the search-phase of echolocation calls, from which statistical attributes such as: a) call duration, b) mean frequency, and c) mean amplitude are also extracted for each species [Walters et al. 2013].

Another set of techniques focuses on learning representations directly from spectrograms of recorded calls. However, regardless of the chosen approach, event localization in audio remains a significant challenge, as in most cases, specific processing steps for each species and its habitat need to be applied to the data [Mac Aodha et al. 2018].

In [Tanveer et al. 2021], the authors demonstrate the effectiveness of methods based on spectrogram representation learning using a bat echolocation-mimicking sonar to identify tree leaf density. They use the Mel-spectrogram as an input parameter for a Convolutional Neural Network (CNN) and achieve an accuracy of 98.7%, which seems to be a promising parameter for adopting the technique.

Furthermore, according to [Tanveer et al. 2021], the Mel-spectrogram has proven to be a powerful method for representing audio signals as images, preserving information from the time and frequency domains and generating efficient input samples for CNNs.

According to [Walters et al. 2013], there are several quantitative methods available for classifying bat echolocation calls, including multivariate statistical methods, machine learning techniques, and deep learning. Many of these methods have been applied to the task of species classification through echolocation calls.

Deep learning is an effective strategy for audio classification, particularly through the use of Convolutional Neural Networks (CNNs), which are specialized algorithms designed for processing images in classification tasks [Zualkernan et al. 2021]. They can be applied to images from different domains or to audios with significant noise presence, as demonstrated in [Ozer et al. 2018].

Bats represent a quarter of all mammalian species, providing key ecosystem services and serving as efficient bioindicators of environmental health, especially in underground regions such as caves. Therefore, automated methods for classifying bat echolocation calls play an important role in the identification, cataloging, preservation, and management of these species [Yoh et al. 2022], taking into account the sensitivity to disturbances exhibited by these individuals when exposed to mining activities.

2. Materials and methods

2.1. Database

A bioacoustic database consisting of 22 audio recordings was utilized, with varying durations, encompassing a total of 3 species, namely: a) Furipterus horrens; b) Lonchorhina aurita; and c) Natalus macrourus. An additional class was generated from audio segments without echolocation calls, forming the ”no calls” class.

The aforementioned species were selected based on two criteria: a) specificity of the frequency range used by the species; and b) the number of calls emitted in each audio recording. The first criterion filters species that share common frequency ranges, as this increases the processing and modeling time required to produce satisfactory results. The second criterion filters sound recordings with few or no occurrences of calls, making the class currently irrelevant for this study. More data continues to be collected and will be used to expand this study in the future.

All data were collected by the biodiversity team of Instituto Tecnologico Vale at the S11-D mine located in Carajás, State of Pará, Brazil. The recordings were made using the Wildlife Acoustics' *Song Meter SM4BAT FS* recording equipment at the cave entrance in the late afternoon, during the bats' foraging period when they exit the cave in large numbers, providing the capture of various echolocation calls simultaneously from different species. All recordings were made non-invasively at a sampling rate of 384KHz.

2.2. Data pre-processing

The raw sound recordings for the 3 species were loaded into a Jupyter Notebook, where each one was concatenated into a single large recording per species. Then, using the SciPy library, a bandpass filter was applied to remove noise, preserving only the information within the frequency range used by each species to emit calls. The species and their respective frequency ranges are listed below [Falcao et al. 2015][Gessinger et al. 2019][Arias-Aguilar et al. 2018]:

- *Furipterus horrens*: [123.41, 191.60]
- *Lonchorhina aurita*: [32.98, 59.29]
- *Natalus macrourus*: [90.80, 173.82]

After filtering within the frequency ranges listed above, the data transitioned from complete noise to a signal composed of reduced noise and multiple peaks on the x-axis, representing the calls emitted by each species. Figure 1 below shows: a) a raw segment of a recording; b) a filtered version of the same segment.

After filtering, we performed the peak detection process (representing the echolocation calls in the recording, as showed in the image above). This process was necessary as a feature extraction step for subsequent modeling.

With that done, the audio recordings of each species were windowed, with each window having a duration of 1 second and containing at least 1 peak (call) within it. Therefore, our samples became composed of 1 call per window, with 384,000 data points, corresponding to 1 second of recording.

Next, the Mel-spectrogram [Zualkernan et al. 2021] was calculated through the Python's Librosa package, for each window, thus transforming the 1-second recording

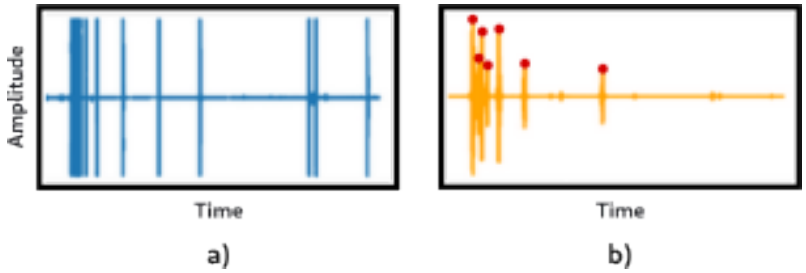

Figure 1. A figure showing: a) unfiltered segment of audio record; b) same segment filtered for the specie, with its peaks marked (Source: Authors).

samples into their respective spectrograms. This process significantly reduced the resolution of the input data, from 384,000 to [120 x 120, 1], for our neural network model while preserving information from the frequency and time domains. Figure 2 below displays a) a raw windowed sample with a call within, and b) the Mel-spectrogram extracted from a).

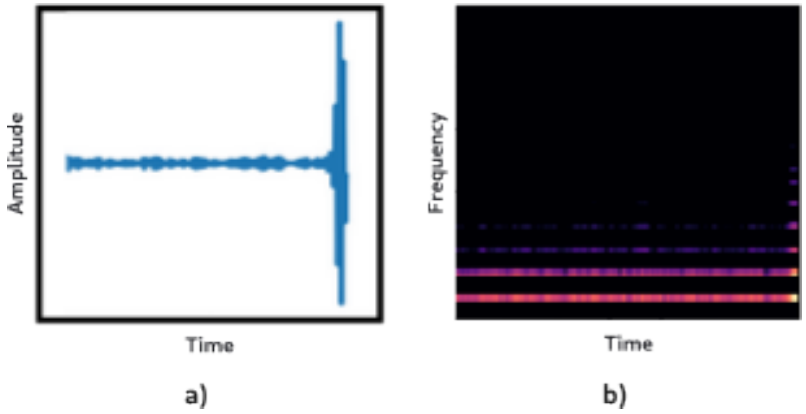

Figure 2. A figure showing: a) raw windowed sample; b) the Mel-spectrogram for the same image (Source: Authors).

Using only the samples extracted from the calls present in the raw data would not be sufficient to develop a new classification model, as we had only a few occurrences of calls per species. Therefore, we saw the need to apply a data augmentation process. Prior to the application of any techniques, we separated the training, testing, and validation sets, respecting a distribution of 70%-20%-10% to avoid any data leakage effects.

Due to the sensitivity of the sonar call recordings, we employed a data augmentation process that involves using the Python's Audiomentations package for applying frequency masks [Park et al. 2019] to specific frequency blocks in the spectrograms generated from the raw data. This process generated hundreds of new samples without altering the original call pattern. As a result, we increased the initial 698 samples, with an unbalanced distribution, to 9,415, balanced across the 4 classes of the problem, thus concluding the data preprocessing stage.

2.3. Deep learning techniques

Considering the expansion of the current project, there is the possibility of deploying the trained model on mobile recording equipment, because of that, we chose to use the Google's MobileNetV2 model [Sandler et al. 2018], which consists of a

highly efficient algorithm for mobile computer vision applications and embedded systems [Howard et al. 2017], that doesn't compromise the performance for audio classification tasks [Liu et al. 2023]. Furthermore, according to [Kornblith et al. 2019, Sowmya et al. 2022, Michele et al. 2019], there is sufficient evidence in the literature suggesting that using pretrained models, such as MobileNets can transfer their high accuracies acquired on ImageNet Large Scale Visual Recognition Competition (ILSVRC) to the new problems they're trying to solve.

For training, we loaded the pretrained weights of the MobileNetV2 model, originally trained on the ImageNet competition. After that, we froze the convolutional layers of the network to prevent them from being retrained, which would have hindered the convergence process since our augmented dataset was derived from a set with a very limited number of samples.

Subsequently, we replaced the original MobileNetV2 fully connected layer, with one created by us so that it could learn the patterns from the bioacoustic dataset.

The new classification layer consists of the following items: a) Global Average Pooling; b) Three Dense with 512, 256, and 128 units each; c) Dense with 4 units (final). All layers have ReLU activation, as well as had their optimal units and number of dense layers all searched by a Bayesian optimization process [Frazier 2018]. The figure 3 below shows the custom fully connected layer attached to the pre-trained MobileNetV2.

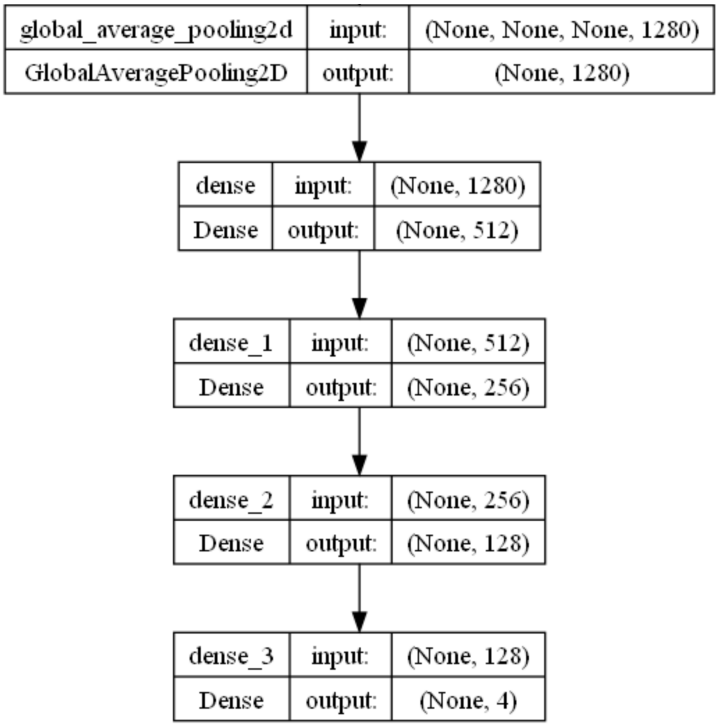

Figure 3. A figure showing the custom fully connected layer attached to the pre-trained MobileNetV2. (Source: Authors)

The model compilation was done with the following parameters for convergence and performance evaluation: a) categorical cross entropy for loss calculation; b) Adam with learning rate of 0.0001 for optimization (also searched through a Bayesian optimiza-

tion process); c) 30 epochs with callback for early stopping; and d) validation accuracy as the measurement of performance.

For comparison purposes, we performed transfer learning on two other CNN models, with increasing complexity compared to MobileNetV2, namely: a) MobileNet; b) NasNetMobile; c) EfficientNetB0. Each of them has 4.3, 5.3, and 5.3 million adjustable parameters, respectively, and were trained under the same conditions as the MobileNetV2 model.

2.4. Performance metrics

We used accuracy as the main performance metric since we have a well-balanced dataset.

3. Results

The results of this work demonstrate that the utilization of echolocation calls, filtered by the species-specific frequency range, converted into Mel-spectrograms, and used to generate synthetic samples through the application of frequency masks, is promising.

The MobileNetV2 model proved to be more efficient in classifying a large number of synthetic samples generated from a limited dataset, and its reduced number of adjustable parameters demonstrated to expedite the path towards convergence when compared to the other more complex models trained in this study.

Limitations still exist, mainly due to the limited number of samples used to generate synthetic data, which slightly restricts the convergence of the main model. However, this will be discussed further in the conclusion.

The figure 4 below shows a plot of Train and Test accuracy over the epochs, as well as the loss in the background. The Table 1 that comes after shows the comparative performance of the trained models.

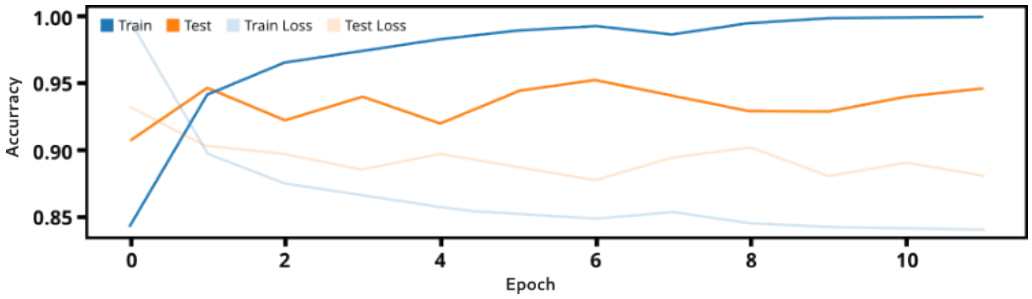

Figure 4. The changes of accuracy and loss (not sharing y-axis) over the epochs for both train and test sets. (Source: Authors)

Based on the table above, we can clearly observe that the more complex the model in terms of adjustable parameters, the worse the result. Our hypothesis, yet to be answered, is that this occurs due to the limited amount of samples used to generate the synthetic dataset, and as more data is collected, this tendency may change.

Despite the limitations in the results, the MobileNetV2 model emerges as the way to go, allowing, in the future, the deployment of the algorithm in mobile collection equipment, which efficiently fulfills one of the long-term objectives of this project.

Table 1. Comparative performance table on the validation set of the four trained models.

Modelo	Number of Adjustable Parameters	Accuracy	Loss
MobileNetV2	3.5M	0.9521	0.1029
MobileNet	4.3M	0.9150	0.1935
EfficientNetB0	5.3M	0.9107	0.2038
NasNetMobile	5.3M	0.7146	0.5507

4. Conclusion

In the case of an initial work, with proprietary and non-public data, still in the data collection phase in the caves of the S11-D mine, in Carajás, State of Pará, this study was able to show the direction to follow for the construction of a model capable of generalizing various bat species, whether they share the same frequency range for echolocation or not.

As demonstrated earlier, there are still limitations that hinder a better convergence of the MobileNetV2 model. Among these limitations, the most significant one is the limited occurrence of calls for each species, which were used to generate the synthetic database through the application of frequency masks. This limits the variation in the specific patterns of each species that the model needs to learn in order to effectively generalize each class.

This is an ongoing work that will expand in the future, but currently serves to validate initial hypotheses regarding the methodology for processing sound data, as well as which CNN model to use, in order to meet not only the objective of efficiently classifying bat species but also enabling the deployment of the algorithm on mobile collection equipment, which is one of the main objectives of this project in the future. We hope that in a future version of this study, we will have more data for the species used in this work, as well as new species covered by the model.

References

Arias-Aguilar, A. et al. (2018). Who's calling? acoustic identification of brazilian bats. *Mammal Research*, 63:231–253.

Auler, A. S. and Pilo, L. B. (2014). Caves and mining in brazil: the dilemma of cave preservation within a mining context. In *Hydrogeological and environmental investigations in karst systems*, pages 487–496. Springer Berlin Heidelberg.

Falcao, F. et al. (2015). Unravelling the calls of discrete hunters: acoustic structure of echolocation calls of furipterid bats (chiroptera, furipteridae). *Bioacoustics*, 24(2):175–183.

Frazier, P. I. (2018). A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Frick, W. F., Kingston, T., and Flanders, J. (2020). A review of the major threats and challenges to global bat conservation. *Annals of the New York Academy of Sciences*, 1469(1):5–25.

Gessinger, G. et al. (2019). Unusual echolocation behaviour of the common sword-nosed bat lonchorhina aurita: an adaptation to aerial insectivory in a phyllostomid bat? *Royal Society open science*, 6(7):182165.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Jacka, J. K. (2018). The anthropology of mining: the social and environmental impacts of resource extraction in the mineral age. *Annual Review of Anthropology*, 47:61–77.

Jones, K. E. et al. (2013). Indicator bats program: a system for the global acoustic monitoring of bats. In *Biodiversity monitoring and conservation: bridging the gap between global commitment and local action*, pages 211–247.

Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671.

Liu, X. et al. (2023). Simple pooling front-ends for efficient audio classification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mac Aodha, O. et al. (2018). Bat detective—deep learning tools for bat acoustic signal detection. *PLoS computational biology*, 14(3):e1005995.

Mantz, J. W. (2008). Improvisational economies: Coltan production in the eastern congo. *Social Anthropology/Anthropologie Sociale*, 16(1):34–50.

Michele, A., Colin, V., and Santika, D. D. (2019). Mobilenet convolutional neural networks and support vector machines for palmprint recognition. *Procedia Computer Science*, 157:110–117.

Ozer, I., Ozer, Z., and Findik, O. (2018). Noise robust sound event classification with convolutional neural network. *Neurocomputing*, 272:505–512.

Park, D. S. et al. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Smith, J. H. (2011). Tantalus in the digital age: Coltan ore, temporal dispossession, and "movement" in the eastern democratic republic of the congo. *American Ethnologist*, 38(1):17–35.

Sowmya, M., Balasubramanian, M., and Vaidehi, K. (2022). Classification of animals using mobilenet with svm classifier. In *Computational Methods and Data Engineering: Proceedings of ICCMDE 2021*, pages 347–358. Springer.

Tanveer, M. H. et al. (2021). Mel-spectrogram and deep cnn based representation learning from bio-sonar implementation on uavs. In *2021 International Conference on Computer, Control and Robotics (ICCCR)*, pages 220–224. IEEE.

Voigt, C. C. and Kingston, T. (2016). *Bats in the Anthropocene: conservation of bats in a changing world*. Springer Nature.

Walters, C., Collen, A., Lucas, T., Mroz, K., Sayer, C., and Jones, K. (2013). Challenges of using bioacoustics to globally monitor bats. In *Bat Evolution, Ecology, and Conservation*, page 479–99. Springer New York.

Yoh, N. et al. (2022). A machine learning framework to classify southeast asian echolocating bats. *Ecological Indicators*, 136:108696.

Zualkernan, I. et al. (2021). An aiot system for bat species classification. In *2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 155–160. IEEE.