Article

Artificial intelligence in acoustic ecology: Soundscape classification in the Cerrado

Bruno Daleffi da Silva, **Linilson Rodrigues Padovese**
Polytechnic School at the University of São Paulo, Brazil
E-mail: brunodaleffi@gmail.com, lrpadove@usp.br

Received 22 February 2025; *Accepted 25 March 2025*; *Published online 28 March 2025*; *Published 1 September 2025*

Abstract
This article explores the application of machine learning techniques in acoustic ecology to classify the formations of the Brazilian Cerrado (Forest, Savanna, and Grassland) based on their soundscapes. Considering the importance of the Cerrado in biodiversity and hydrology, along with the challenges faced by the biome due to agricultural expansion, the study seeks more efficient and cost-effective methods for identifying its phytophysiognomies. Five statistical models were developed and evaluated, utilizing both traditional Machine Learning and Deep Learning, with Mel Frequency Cepstral Coefficients (MFCCs) and spectrogram images as input variables. The performance comparison of these models revealed the superiority of the Convolutional Neural Network (CNN), which, although requiring higher computational costs and training time, provided high accuracy in classifications and valuable insights through the application of the LIME explainability technique. Additionally, the study proposes a multiple classification methodology by majority voting for frequently observed events, enabling reliable classifications through models with moderate performance. The conclusion is that it is possible to classify different Cerrado formations through their acoustic landscape, and the choice of the optimal model for classification should consider a balance between accuracy, operational complexity, and efficiency. The findings of this study offer relevant guidance for future research and the application of monitoring technologies in conservation and biome recovery efforts.

Keywords Cerrado; acoustic ecology; soundscapes; Artificial Intelligence; Machine Learning; Convolutional Neural Network (CNN); Gradient Boosting; Random Forest; environmental preservation; biodiversity.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The soundscape of the Cerrado, one of Brazil's richest and most threatened biomes, has been the subject of intense research due to its biodiversity and ecological importance. This biome, which occupies about 23% of the national territory and extends across 12 states (BDIA-IBGE), is characterized by a semi-arid climate and dense shrub vegetation. However, the expansion of agricultural and livestock activities has led to a significant loss of its vegetation cover, with about 150,000 square kilometers of vegetation being deforested between 2000

and 2018 (IBGE, 2020, Annex 2). The degradation of the Cerrado not only threatens its flora and fauna, including endemic species, but also affects its critical role in the country's water and climate regulation (Dias, 1991).

Given this scenario, the restoration of degraded areas and the preservation of the existing biome become imperative. To this end, it is essential to understand the region's natural formation, identify biodiversity loss, and determine the ideal formation for restoration, which may or may not correspond to the original biome, depending on the degree of degradation (Cava, 2018).

Traditionally, the process of identifying the phytophysiognomies of the Cerrado is based on qualitative analyses of flora and fauna, which requires substantial field efforts (Ribeiro Apud Embrapa, 2008).

Fig. 1 Phytophysiognomies of the Cerrado.

However, recent advances in Acoustic Ecology offer a promising alternative, using the soundscape to efficiently and economically identify and classify environmental characteristics (Tucker, 2014).

This article presents an innovative approach that integrates Artificial Intelligence in the analysis of the Cerrado soundscape, exploring the potential of machine learning models to discern its different natural formations. By using acoustic data from regions representative of the Forest, Savanna, and Grassland formations, this study aims to validate the hypothesis that similar ecological identities are reflected in similar soundscapes. The efficiency in identifying and classifying these formations through their sound profiles is practical proof of the validity of this hypothesis.

The contribution of this work is threefold: first, it highlights the application of sound classification based on Artificial Intelligence; second, it considers the relationship between the complexity of the method and its practicality; and third, it emphasizes the use of LIME in spectrogram-based CNNs, a significant advancement that allows identifying the spectral regions and frequency bands that characterize each formation. These advances not only enhance the accuracy of analyses but also provide valuable elements for the conservation and restoration of the Cerrado.

2 Literature Review

The analysis of soundscapes is an approach that has gained prominence in biodiversity assessment and ecosystem understanding. The soundscape consists of a complex array of sounds, known as biophony, geophony, and anthropophony, which reflect the interaction between living beings, natural phenomena, and human activities (Priestman, 2017). Biophony, in particular, is a vital indicator of the presence and behavior of species in an ecosystem, making it a valuable tool for biodiversity studies.

In the context of the Cerrado, a biome of extreme ecological importance with unique biodiversity, the use of soundscapes for environmental assessment is particularly pertinent. The Cerrado is a biodiversity hotspot, home to a wide variety of endemic species and crucial for maintaining important watersheds (Dias, 1991). Acoustic analysis of the environment can provide detailed information on the ecological health of these areas, enabling the detection of changes in species composition and ecosystem dynamics.

Recent studies have applied Machine Learning and Deep Learning methods to classify species and monitor biodiversity through sound recordings. For example, Nahian Ibn Hasan's (2022) study presents a hybrid of traditional signal processing and deep learning approaches to identify bird species from audio recordings, achieving an accuracy of 90.45% for a set of 10 bird classes. These advances demonstrate the potential of combining soundscape analysis techniques and Artificial Intelligence in the conservation and monitoring of biomes like the Cerrado.

The application of Machine Learning models, such as Gradient Boosting, has shown promising results in soundscape studies. Fonseca et al. (2017) demonstrated the effectiveness of this model, improving baseline performance by 8.2% when classifying acoustic scenes. Similarly, the use of Random Forest in audio signal classification systems achieved an overall correct classification rate of 99.25% in a study by Grama et al. (2017), highlighting the model's ability to handle imbalanced datasets and identify sounds related to wildlife intrusion detection.

Logistic Regression has also been successfully applied in the analysis of urban soundscapes, as shown by Noviyanti et al. (2019), who used Mel Frequency Cepstral Coefficients (MFCCs) (Davis et. al., 1980) to predict sound perception with Correct Classification rates of up to 88.3%.

Additionally, the Multilayer Perceptron (Rosenblatt, 1958; Zhang, 2010), a neural network model, was effective in the multinomial classification of acoustic patterns in audio clips, as explored by Zhang et al. (2016), providing detailed information on the distribution of various acoustic patterns in long-duration recordings, thus serving as another powerful tool for acoustic data analysis.

Delving further into the realm of neural networks, Convolutional Neural Networks (CNNs) have stood out in soundscape analysis, particularly in spectrogram classification (Stowell et al., 2014; Pellegrini et al., 2020). Khamparia et al. (2019) explored environmental sound classification using CNNs and achieved an accuracy of 77% on the ESC-10 dataset. This study underscores the efficiency of CNNs in recognizing and classifying spectrogram images, which are visual representations of sound frequencies over time.

Acoustic Ecology, which originated with the pioneering works of Schafer (1977) and Truax (2001), has evolved with the development of Data Science and Artificial Intelligence. The integration of these disciplines has enabled deeper and more automated analysis of soundscapes, enhancing the understanding of ecological interactions and the impacts of human activities on the environment. The World Forum for Acoustic Ecology (WFAE), established in 1993, has been a platform for the dissemination and international recognition of this research area (Wrightson, 2000).

The application of these techniques in the Cerrado can be an effective strategy for biome preservation. Identifying specific soundscapes associated with different Cerrado vegetation formations is a crucial step for implementing conservation and restoration measures. Acoustic analysis can assist in identifying degraded areas and evaluating the effectiveness of management practices, contributing to the sustainability and resilience of this ecosystem.

In summary, the literature review indicates that soundscape analysis, supported by advances in Artificial Intelligence, is a promising approach for biodiversity assessment and ecosystem conservation. In the Cerrado, this approach can not only help understand the biome's complexity but also guide preservation and restoration efforts in the face of increasing anthropogenic pressures.

3 Methodology

To investigate the central hypothesis that similar phytophysiognomic formations exhibit similar soundscapes, this study proposes to analyze the following guiding questions:

- Model Efficiency: Which machine learning models, ranging from traditional statistical methods to advanced neural networks, are most efficient and effective in classifying the types of natural formations in the Cerrado based on their soundscapes?
- Complexity and Computational Cost: How do the complexity and computational cost of the models impact the choice of the most appropriate method for classifying soundscapes?
- Relevant Sound Signal Attributes: Which attributes of the sound signals are most relevant for the classification of the Cerrado soundscapes?

To address these questions, the study analyzed the effectiveness of models using Mel Frequency Cepstral Coefficients (MFCCs) and spectrograms as input variables for the classification of soundscapes. Machine Learning and Deep Learning models such as Gradient Boosting, Random Forest, Logistic Regression, Multilayer Perceptron, and Convolutional Neural Networks were employed.

Additionally, the study examined the impact and importance of considering method simplicity, training and prediction response time, and the ability to handle a large number of observations in a short period. Finally, it sought to understand the importance of the main features of the spectrogram in classifying the Cerrado soundscapes.

The outlined objectives aim not only to validate the central hypothesis but also to contribute to the methodological advancement in the application of Artificial Intelligence for soundscape analysis in ecological contexts. By understanding the acoustic characteristics associated with each vegetation formation, this study significantly contributes to the preservation and restoration of the Cerrado biome and to the science of acoustic ecology as a whole.

Adopting a methodical and structured strategy, the study followed four crucial steps that reflect the data science cycle as proposed by Shearer (2000).

3.1 Data Collection

Field acoustic data collection was a crucial step in this study and was carried out with the help of acoustic recording equipment developed for this purpose by the Laboratory of Acoustics and Environment (LACMAM) at the Polytechnic School of the University of São Paulo (USP). For a comprehensive analysis of the Cerrado soundscapes, three Ecological Stations (EEs) were selected, each representing one of the main vegetation formations of the biome: forest, grassland, and savanna.

The chosen regions were the EE of Assis, representing the forest formation, EE of Itirapina, as an example of the grassland formation, and EE of Águas de Santa Bárbara, illustrating the savanna formation. These classifications were not arbitrarily assigned; they result from meticulous field studies conducted by specialists who examined the characteristics and densities of the vegetation at each site. This initial classification was essential to establish a reference point for subsequent acoustic analysis.

The following figure presents visual representations of the three analyzed Ecological Stations: Assis, Santa Bárbara, and Itirapina. The images were sourced respectively from the Catalog of Plants of the Conservation Units of Brazil[1] (Assis), the Institute of Environmental Research[2] (Santa Bárbara), and the Scielo Brazil

[1]Available at: https://catalogo-ucs-brasil.jbrj.gov.br/descr_areas.php?area=EEAssis

[2]Available at: https://www.infraestruturameioambiente.sp.gov.br/institutoflorestal/areas-protegidas/estacoes-ecologicas/santa-barbara/

platform[3] (Itirapina).

Fig. 2 Vegetation ormations of the Cerrado: Forest, Savanna, and Grassland.

The simultaneous sound recording campaigns allowed for the capture of a broad spectrum of acoustic events and variations, including sounds emitted by fauna, noises generated by abiotic elements such as wind and rain, and any anthropogenic interferences. This holistic approach ensures that the captured sound profile is representative of the complexity and richness of the Cerrado biome.

The volume of collected data was substantial, totaling approximately 3 TB of sound recordings from September to November 2019. The audio data from the EE of Itirapina were recorded in 5-minute files with a sampling rate of 8 KHz and 16-bit depth. In contrast, the soundscapes of the EE of Santa Bárbara and Assis were recorded in 3-minute files with a sampling rate of 32 KHz and 16-bit depth. Given the magnitude of the collected data, it became necessary to adopt a strategy to make the data volume more manageable while maintaining the integrity and representativeness of the information.

To this end, a random selection of 2,500 files from each of the three Ecological Stations was performed, totaling 7,500 files and amounting to 70 GB of data, constituting what will be called the Temporal Database (TDB). The randomization of the files ensured that the regions and recordings, being independent, did not bias the sample, especially considering that the data were collected over a similar recording period for the three regions. This strategic reduction of the database ensured a diverse representation of biotic conditions, climatic conditions, and recording times, encompassing different scenarios such as winds, rains, and diurnal and nocturnal variations. This approach allowed the machine learning models to be trained efficiently without compromising their ability to learn and recognize the patterns of the Cerrado soundscapes in various contexts.

[3]Available at: https://doi.org/10.1590/S1676-06032008000300019

3.2 Pre-processing

After collecting the sound data, the subsequent stage of the study involved meticulous pre-processing of the captured audio. From this phase onwards, all analyses were conducted using the R programming software (R CORE TEAM, 2024).

This pre-processing phase consists of three steps, aiming to obtain a uniform database format that will be used in the following modeling and feature extraction phase.

The first step is standardizing the duration of the recordings and the frequency band of the Temporal Database (TDB). This was achieved by restricting the analysis to the first minute of each sample. This standardization was essential to ensure the comparability between recordings, regardless of their original duration.

Next, the sampling rates were resampled using the downsample function from the tune library in R. The downsample function reduces the sampling rate of an audio file by partitioning the number of samples of the original file into n equal parts, where

$$n = Final\ Sampling\ Rate * \frac{Initial\ Number\ of\ Samples}{Initial\ Sampling\ Rate}$$

This process was necessary to synchronize the datasets, which were recorded with different sampling rates among the EEs. Resampling the data from Santa Bárbara and Assis to 8 KHz ensured that all data were in a consistent format, facilitating subsequent analyses. At the end of this step, the TDB consisted of 7,500 ".wav" files, each with a sampling rate of 8 KHz. These files represent one minute of random soundscapes collected in the EEs of Assis, Santa Bárbara, and Itirapina between September and November 2019.

The second step was the transformation of each temporal sample from the TDB to the frequency domain using the Short-Time Fourier Transform (STFT). The transformation was performed using the spectro function from the seewave package in R (Sueur et al., 2008). The seewave library, in turn, is based on the book "Animal Acoustic Communication: Sound Analysis and Research Methods" (Hopp et al., 1998).

The parameters set in the "spectro" function regulate the window size and the degree of overlap between them. In this analysis, the function's default values were used: the window size (parameter 'wl') was fixed at 512 points, while the overlap (parameter 'ovlp') was set to 0, indicating no overlap between windows.

With these parameters, each 512-point window of the audio signal corresponds to a 256-point vector in the STFT operation. Thus, the resulting matrix contains 256 rows.

Given a window size of 512 points and a sampling rate of 8kHz, each window represents 512 / 8000 = 0.064 seconds of audio. To cover 59 seconds of audio, approximately 59 / 0.064 ≈ 921 windows are required.

The multiplication of the number of rows (frequencies) by the number of columns (windows) results in the total number of cells in the matrix, which corresponds to the total number of points in the STFT representation of the audio signal. In this case, there are 256x921 = 235,776 points.

Thus, the use of STFT on the TDB resulted in the creation of a Time-Frequency Database (TFDB), a three-dimensional database that captures the temporal evolution of frequencies and their respective amplitudes, providing a detailed representation of the acoustic content of the recordings. Each audio file is transformed into a database consisting of 235,776 rows, encapsulating detailed spectral information and including relevant metadata such as the recording location and date.

Finally, the amplitudes in the TFDB were normalized to standardize the volume of the recordings and reduce distortions or variations in sound intensity caused by different recording equipment. Normalization adjusts the amplitudes to a common scale, facilitating comparison between recordings. The amplitudes at instances *j* of each database *i* are normalized using the formula:

$$z_{ij} = \frac{amplitude_{ij} - mean(amplitude_i)}{standard\ deviation(amplitude_i)}$$

ensuring that Z has a mean of 0 and a variance of 1.

All the pre-processing steps convert the original recordings into a matrix fundamental for future developments and the construction of classification models. At the end of this process, the audios are transformed into an Acoustic Feature Database (AFDB), which contains detailed information on time, frequency, normalized amplitude, and location for each formation of the three major natural formations of the Cerrado. The AFDB, with dimensions of 7,500 records x 4 columns (location, time, frequency, and normalized amplitude) x 235,776, is the final database, which will be used to extract features such as Mel Frequency Cepstral Coefficients (MFCCs) and create spectrograms, which will be the input variables for the supervised models constructed.

MFCC

With the AFDB in hand, the melfcc function from the "tune" package was used to calculate the Mel-frequency Cepstral Coefficients (MFCCs). This function is inspired by the melfcc.m function developed in the rastamat library of Matlab (Hermansky, 1990; Hermansky et al., 1994).

MFCCs are obtained by emulating human auditory perception (Davis et al., 1980). The process begins with a pre-emphasis filter to amplify the high frequencies, followed by the Short-Time Fourier Transform (STFT) with a Hamming window to obtain the short-term spectrum. This spectrum is then mapped onto a Mel scale, approximating human perception of frequencies. Next, the logarithm of this spectrum is applied, and the Discrete Cosine Transform (DCT) is used to obtain the cepstral coefficients. Finally, the first 12 cepstral coefficients are extracted and represent the MFCCs. After obtaining these coefficients, normalization was performed to eliminate possible device noise and ensure the robustness of the method.

Spectrograms

Spectrograms are graphical representations that illustrate how the frequencies of a signal vary over time, offering an intuitive and informative view of the studied sound environment. These images were used as input variables in the Convolutional Neural Network models.

The first step in the spectrogram construction process involved reading the AFDB. Then, with the help of the ggplot2 library in R (Wickham, 2016), the data were transformed into visual spectra.

Fig. 3 shows examples of these spectrograms, illustrating the sound diversity of the Cerrado formations.

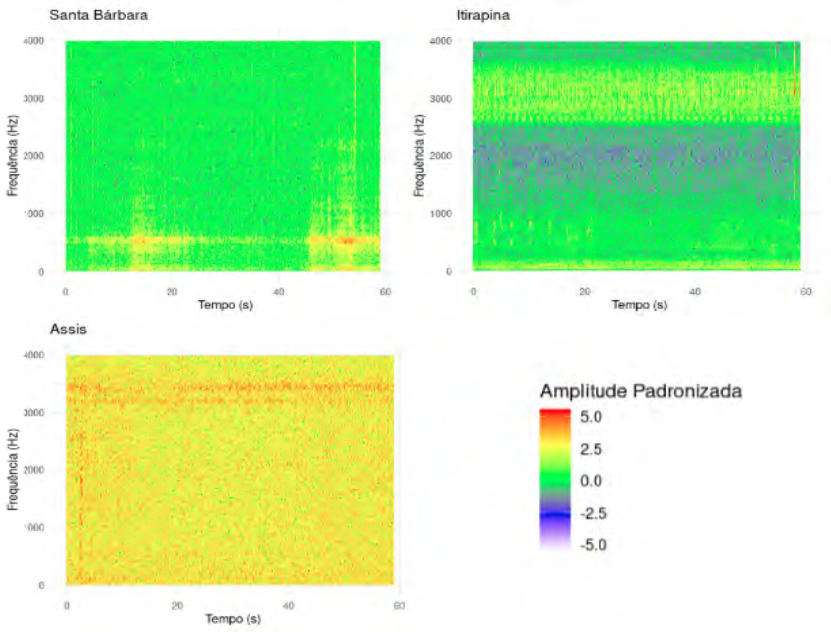

Fig. 3 Example of spectrograms from the three regions/formations with standardized amplitudes.

3.3 Modeling

The modeling phase was a crucial step in this study, where the information contained in the AFDB was processed through a range of advanced machine learning techniques. The selection of models was based on a literature review and previous technological advances, encompassing traditional statistical methods to deep neural networks. Among the traditional models, Gradient Boosting, Random Forest, Logistic Regression, and Multilayer Perceptron were selected, using Mel Frequency Cepstral Coefficients (MFCCs) (Davis et al., 1980) as input variables. On the other hand, for the Convolutional Neural Network (CNN), spectrogram images were used as the basis for classification.

Except for the MLP, the models were implemented with the tidymodels library (Kuhn et al., 2020; Kuhn et al., 2022) in R, which provides an integrated and consistent approach to statistical modeling and machine learning, following the principles of the tidyverse (Wickham et al., 2019).

Among the packages offered by tidymodels are "tune" (Kuhn, 2024) and "parsnip" (Kuhn, Vaughan, 2024). The "tune" package plays a fundamental role in optimizing the hyperparameters of machine learning models. Hyperparameters are model settings that need to be specified before training and are not learned from the data. They can influence the model's effectiveness, and their optimization allows for the evaluation and comparison of different models, improving their performance.

On the other hand, the "parsnip" package is used for model specification. It provides a consistent interface for defining and configuring machine learning models, regardless of the underlying computational engine implementing them. This package was extensively used in this study to define and adjust the hyperparameters of the Gradient Boosting, Random Forest, and Logistic Regression models.

For the Gradient Boosting model, using the boost_tree function from the parsnip package, the following hyperparameters were adjusted: mtry, min_n, tree_depth, sample_size, learn_rate, and loss_reduction. Here, mtry refers to the number of variables available for splitting at each tree node, min_n is the minimum number of observations in the nodes, tree_depth is the maximum depth of any tree, sample_size is the fraction of the data used to build each tree, learn_rate is the learning rate, and loss_reduction refers to the minimum loss reduction required to make a new split in the tree.

The Random Forest model, implemented through the rand_forest function from the parsnip package, had the hyperparameters mtry and min_n adjusted. Mtry, as in the previous model, is the number of variables available for splitting at each tree node, while min_n is the minimum number of observations in the nodes.

In Logistic Regression, using the multinom_reg function from the parsnip package, the hyperparameters penalty and mixture were adjusted. The penalty refers to the amount of regularization applied, which helps avoid overfitting, while mixture determines the type of regularization applied (L1, L2, or a mix of both).

The implementation of the CNN and MLP was carried out using the TensorFlow (GOOGLE, 2023) and Keras (Chollet et al., 2023) libraries, which are open-source tools for machine learning and deep neural networks. These libraries offer a flexible and powerful environment for building, training, and validating complex models, such as CNNs, which are particularly suitable for processing visual and acoustic data.

The MLP architecture featured six intermediate layers, including two dense layers with 30 and 18 neurons, respectively, along with a dropout layer and a normalization layer, placed between the dense layers. The total number of adjustable parameters in the MLP model was 1,101.

The CNN architecture comprised several layers, including convolutional layers, batch normalization layers, max pooling layers, and spatial dropout layers. The network starts with an input layer with dimensions 250x250x3, representing the dimensions of the input images.

Specifically, the model has three convolutional blocks, each with two convolutional layers, followed by a batch normalization layer, a max pooling layer, and a spatial dropout layer. The number of filters in the

convolutional layers progressively increases with each block, starting with 32, then 64, and finally 128. All convolutional layers use ReLU activation functions and have a kernel size of 5x5, with 'same' padding to ensure the output has the same dimension as the input.

After the three convolutional blocks, the network has a global average pooling layer, followed by a flatten layer and a dense layer with 128 units and ReLU activation function. A dropout layer with a rate of 0.4 is applied before the output layer, which has three units and a softmax activation function, corresponding to the three classification categories.

The model has a total of 815,139 adjustable parameters, of which 814,243 are trainable and 896 are not trainable. During training, the model was compiled using the Adam optimizer with a learning rate of 0.001, categorical_crossentropy loss function, and accuracy as the evaluation metric. Two callback functions were used: one to reduce the learning rate when the validation loss stopped improving and another to stop training when the validation accuracy exceeded 98% after 10 consecutive epochs above 97%.

The model was trained for up to 300 epochs with a batch size of 64. Validation data were used to monitor the model's performance and adjust the learning rate as needed. The use of callbacks allowed for more efficient training, avoiding overfitting and saving computational time.

The modeling methodology also included a strategic data split, separating the total dataset into three distinct parts: training, validation, and blind testing. The training set was used to adjust the models and teach them the data patterns. The validation set played a role in optimizing hyperparameters and preventing overfitting, ensuring model generalization. The blind test set, composed of data not exposed to the models during training, was crucial for testing the models' effectiveness in new and unknown conditions, providing a reliable performance evaluation.

Each model underwent a careful training and validation process, aiming to maximize its accuracy and generalization capacity. Metrics such as accuracy, precision, and recall were used to evaluate each algorithm's performance, providing a comprehensive view of the models' effectiveness. Additionally, explainability techniques, such as LIME, were planned to be applied to the CNN to interpret the model's decisions and identify the most relevant acoustic features for classifying the Cerrado soundscapes.

In summary, the modeling was conducted with methodological rigor, enabling not only efficient classification of soundscapes but also a detailed analysis of the models, contributing to the advancement of knowledge at the intersection of Acoustic Ecology and Data Science.

3.4 Explainability

Understanding the decisions made by machine learning models, especially in complex neural networks such as CNNs, is essential for validating the reliability and applicability of the results obtained. In this context, the explainability analysis focused on the predictions generated by the CNN, employing the LIME (Local Interpretable Model-agnostic Explanations) technique as proposed by Ribeiro et al. (2016). LIME is a methodology that facilitates the interpretation of complex models by introducing perturbations in the input data and observing the variations in the model's predictions. This technique induces the CNN to formulate local linear models that are inherently more understandable, revealing the impact of different areas of the spectrogram image on the classification of soundscapes.

The "lime" library (Hvitfeldt et al., 2022) in R was used to perform the analyses. This library provides an efficient implementation of the LIME technique, allowing seamless integration with other R packages used for data analysis and visualization. The choice of this library is due to its flexibility and ability to provide precise local explanations, aligned with the needs of the explainability analysis applied in this study.

The methodology implemented by LIME involves generating modified versions of the input data (spectrograms) and subsequently analyzing the changes in the CNN's predictions. This analysis allows

identifying which features of the images are crucial for the model's decisions. By applying LIME to specific spectrograms, it is possible to identify which frequency segments or temporal patterns are considered most relevant by the CNN in classifying a soundscape as belonging to a specific formation in the Cerrado.

The adopted explainability approach not only demonstrated the relevance of the acoustic attributes but also promoted greater transparency in the model's predictive decisions.

In summary, the methodology employed in this study allowed the development of effective statistical models and provided a detailed understanding of the predictive decisions. The use of LIME, in particular, ensured a consistent and reliable approach, enabling the validation of the CNN's predictions and a deeper interpretation of the characteristic acoustic patterns of the various formations in the Cerrado. The clarity in the models' decisions is a crucial aspect for advancing the application of Artificial Intelligence in ecological and environmental research.

4 Results

The application of machine learning models for the classification of Cerrado soundscapes revealed significant results. The performance of each model was evaluated based on metrics of accuracy, precision, and recall, as presented in Table 1.

Table 1 Performance of the developed models according to the analyzed metrics.

Model	**Accuracy**	**Precision**	**Recall**	**Training Time**
Gradient Boosting	93%	93%	93%	5,86 minutes
Random Forest	92%	92%	92%	5,31 minutes
Logistic Regression	82%	83%	82%	9,89 seconds
Multilayer Perceptron	83%	79%	69%	2,33 minutes
CNN	98%	97%	97%	~ 8 days

The Convolutional Neural Network (CNN) demonstrated superiority, achieving 98% accuracy, 97% precision, and 97% recall. This remarkable performance indicates that the CNN was able to accurately identify the different formations of the Cerrado from the soundscapes. However, the substantially longer training time (~8 days) compared to the other models must be considered in the context of practical application.

On the other hand, the strategy of "repetitive classification" proved to be effective when using moderately performing models for frequently occurring events. This strategy involves repeatedly applying the same model to several small samples of the same sound event. The final decision is made based on the most frequent response among the classifications, using the majority voting method.

This approach is effective due to the consistency and convergence observed in the accurate identification of the Cerrado's natural formation in various observations. This reflects the principles of the Law of Large Numbers (Bernoulli, 1713), inferring that less complex models, with an acceptable accuracy rate, tend to achieve an overall accuracy rate of 100% as the number of observations of the same sound event increases infinitely.

Nonetheless, although the CNN requires significant investment in terms of time and computational

resources, it brings an additional advantage: the ability to explain its decisions. With the help of LIME (Ribeiro et al., 2016), a model interpretation technique, it was possible to identify which sound features are most important for classifying the different formations of the Cerrado. This is clearly demonstrated in Figure 4, where LIME helps to clarify which aspects of the audio recordings are most relevant to the CNN, providing not only accurate classification but also a deeper understanding of the acoustic data we are analyzing.

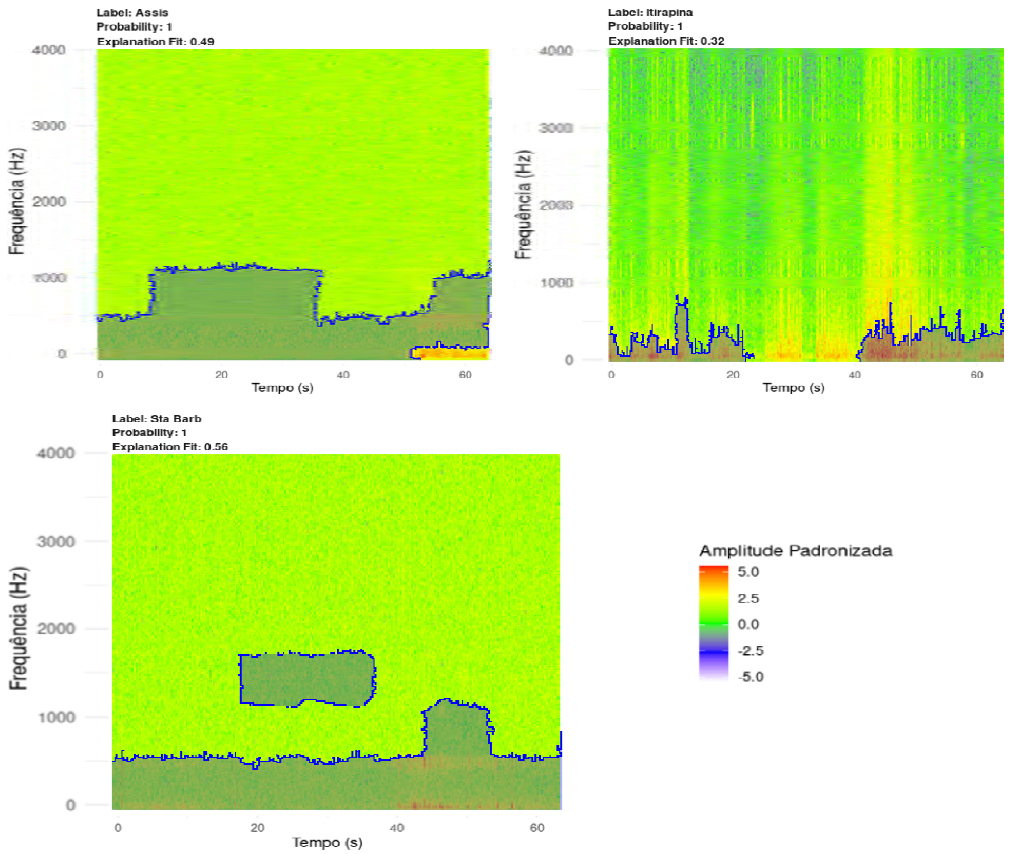

Fig. 4 Application of LIME to explain the CNN in examples of spectrograms.

5 Discussion

The advances achieved in this study underline the synergy between Acoustic Ecology and Data Science in the effective classification of soundscapes, providing valuable analytical tools to advance the understanding and preservation of the Cerrado. The explored models exhibited performance variability, but the applicability and efficiency of their implementation proved to be equally important factors for their selection in environmental analysis projects.

The remarkable performance of the CNN in quantitative results reflects its aptitude for extracting and learning complex features from soundscapes. However, it should be noted that the high accuracy of this model can be offset by the simplicity and agility of less sophisticated methods, such as Gradient Boosting or Random Forest, which are advantageous in various practical scenarios, especially when frequent assessments are necessary.

The methodological approach for classifying high-exposure events had a crucial influence, suggesting that reasonably performing models, when applied iteratively, can provide accurate classification and converge towards the true identification of the natural formation. This technique relies on the consistency of results

generated by dynamically analyzing multiple samples, establishing a reliable precedent for environmental categorizations.

The use of LIME in the CNN analysis provided additional insights, allowing not only the identification of Cerrado formations with high accuracy but also the understanding of the specific acoustic properties that determine each classification. This elucidation highlights the importance of combining different techniques and models to extract the maximum value from the analyzed data.

In conclusion, the research emphasized the relevance of balancing various factors in the choice of models for acoustic landscape classification. The multiparametric perspective of this article highlights the effectiveness of interdisciplinarity in applying artificial intelligence for monitoring and conserving natural ecosystems.

6 Conclusions

The investigation conducted in this study provided robust evidence supporting the central hypothesis that similar natural formations have similar soundscapes. Through the application of machine learning and deep learning models, it was possible to accurately classify the formations of the Cerrado based on their acoustic landscapes, validating the hypothesis and demonstrating the potential of these technologies in ecological analysis.

Addressing the first guiding question, it was found that both traditional statistical methods and advanced neural networks are efficient in classifying soundscapes. However, the Convolutional Neural Network (CNN) stood out as the most effective model, although it requires significant training time and computational resources.

Regarding the second question, the complexity and computational cost of the models influenced the choice of the most appropriate method. While the CNN offered the best performance, models such as Gradient Boosting and Random Forest proved to be viable and efficient alternatives, especially when considering the repetitive classification strategy and the majority voting principle for high-exposure events.

In relation to the third question, the sound signal attributes that were most relevant for the classification of Cerrado soundscapes were the specific frequencies identified by the models, especially the CNN, whose interpretation was enhanced by the use of LIME.

This study concludes that the integration of Artificial Intelligence in soundscape analysis is a promising approach for biodiversity assessment and ecosystem conservation. The developed and tested models offer valuable tools for identifying natural formations, contributing to the preservation and recovery of threatened biomes such as the Cerrado. The choice of the appropriate model should be guided by a balance between accuracy, efficiency, and practicality, considering the specific needs of each application. The explainability of the models, particularly the CNN, reinforces confidence in the classification decisions and provides valuable insights for future research and practical applications.

References

Bernoulli J. 1713. Ars Conjectandi: Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis. Thurneysen Brothers, Basel, Switzerland

Breiman L. 2001. Random Forests. Machine Learning, 45: 5-32. https://doi.org/10.1023/A:1010933404324

Cava MGB, Pilon NAL, Ribeiro MC, Durigan G. 2018. Abandoned pastures cannot spontaneously recover the attributes of old-growth savannas. Journal of Applied Ecology, 55: 1164-1172. https://doi.org/10.1111/1365-2664.13046

Chollet F. Keras: The Python Deep Learning library. https://keras.io/. Accessed June 6, 2023

Cox DR. 1958. The Regression Analysis of Binary Sequences (with Discussion). Journal of the Royal Statistical Society, Series B, 20(2): 215-242

Davis SB, Malmberg P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4): 357-366

Dias BFS. 1991. Alternativas de Desenvolvimento do Cerrado: Manejo e Conservação dos Recursos Naturais Renováveis. Brasília: Fundação Pró-Natureza (FUNATURA), 9, 21

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27(8): 861-874

Fonseca E, Gong R, Bogdanov D, Slizovskaia O, Gomez E, Serra X. 2017. Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks. In: Detection and Classification of Acoustic Scenes and Events Workshop DCASE2017, Munich, Germany

Friedman JH. 2001. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5): 1189-1232

Google Brain Team. TensorFlow: An end-to-end open source machine learning platform. https://www.tensorflow.org/. Accessed June 6, 2023

Grama L, Rusu C. 2017. Audio signal classification using Linear Predictive Coding and Random Forests. In: International Conference on Speech Technology and Human-Computer Dialogue (SpeD), IEEE, Bucharest, Romania

Hasan NI. 2022. Bird Species Classification And Acoustic Features Selection Based on Distributed Neural Network with Two Stage Windowing of Short-Term Features. https://doi.org/10.48550/arXiv.2201.00124. Accessed March 15, 2024

Hermansky H. 1990. Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of the Acoustical Society of America, 87(4): 1738-1752

Hermansky H, Morgan N. 1994. Rasta Processing Of Speech. IEEE Transactions on Speech and Audio Processing, 2(4): 578-589

Hopp SL, Owren MJ, Evans CS. 1998. Animal Acoustic Communication: Sound Analysis and Research Methods. Berlin: Springer International.

Hvitfeldt E, Pedersen TL, Benesty M. 2022. Lime: Local Interpretable Model-Agnostic Explanations. https://CRAN.R-project.org/package=lime. Accessed July 15, 2024

Instituto Brasileiro de Geografia e Estatística (IBGE). Banco de Dados de Informações Ambientais. https://bdiaweb.ibge.gov.br/#/consulta/pesquisa. Accessed August 23, 2022

Instituto Brasileiro de Geografia e Estatística (IBGE). 2020. Contas De Ecossistemas - O Uso da Terra nos Biomas Brasileiros. https://biblioteca.ibge.gov.br/visualizacao/livros/liv101753.pdf. Accessed August 23, 2022

Khamporia A, Gupta D, Nguyen NG, Khanna A, Pandey B, Tiwari P. 2019. Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. IEEE Access, 7: 7717-7727. https://ieeexplore.ieee.org/abstract/document/8605515

Kuhn M. 2024. tune: Tidy Tuning Tools. R package version 1.2.1. https://CRAN.R-project.org/package=tune. Accessed June 30, 2024

Kuhn M, Silge J. 2022. Tidy Modeling with R. O'Reilly Media, California, USA

Kuhn M, Vaughan D. 2024. parsnip: A Common API to Modeling and Analysis Functions. R package version 1.2.1. https://CRAN.R-project.org/package=parsnip. Accessed June 30, 2024

Kuhn M, Wickham H. 2020. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org. Accessed June 30, 2024

LeCun Y. 1998. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11): 2278-2324

Mehyadin AE, Abdulazeez AM, Hasan DA, Saeed JN. 2021. Birds Sound Classification Based on Machine Learning Algorithms. Asian Journal of Research in Computer Science. https://www.researchgate.net/profile/Dathar-Abas-Hasan/publication/352523116. Accessed August 23, 2022

Noviyanti A, Sudarsono AS, Kusumaningrum D. 2019. Urban soundscape prediction based on acoustic ecology and MFCC parameters. AIP Conference Proceedings, 2187. https://doi.org/10.1063/1.5138335. Accessed February 3, 2023

Pellegrini AFA et al. 2020. Birdsong and anthropogenic noise: implications for conservation. Landscape Ecology, 35(5): 1161-1179

Piczak KJ. 2015. Environmental Sound Classification with Convolutional Neural Networks. In: Proceedings of the International Conference on Computational Intelligence in Music, Sound, Art and Design (CIM'15).

Priestman K. 2017. The Science of Soundscapes. Inside Ecology. https://insideecology.com/2017/09/28/the-science-of-soundscapes/. Accessed August 23, 2022

R Core Team. 2024. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/. Accessed June 30, 2024

Ribeiro JF, Walter BMT. 2008. As Principais Fitofisionomias do Bioma Cerrado. In: Cerrado: Ecologia E Flora Vol. 2 (Sano SM, Almeida SP, Ribeiro JF, eds). EMBRAPA-CERRADOS, Brasília

Ribeiro MT, Singh S, Guestrin C. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 97-101, San Diego, California. USA. https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf. Accessed April 22, 2023

Rosenblatt F. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65(6): 386-408

Schafer RM. 1977. The soundscape: Our sonic environment and the tuning of the world. Destiny Books, Rochester, VT

Shearer C. 2000. The CRISP-DM model: The new blueprint for data mining. Journal of Data Mining, 5(4). https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf. Accessed August 26, 2023

Stowell D et al. 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ, 2: e488

Sueur J, Aubin T, Simonis C. 2008. Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics, 18: 213-226

Truax B. 2001. Acoustic Communication (2nd ed). Ablex Publishing, Westport, CT

Tucker D, Gage SH, Williamson I et al. 2014. Linking ecological condition and the soundscape in fragmented Australian forests. Landscape Ecology, 29: 745-758. https://doi.org/10.1007/s10980-014-0015-1

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag. https://ggplot2.tidyverse.org. Accessed July 7, 2024

Wickham H et al. 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43): 1686. https://doi.org/10.21105/joss.01686. Accessed June 30, 2024

Wrightson K. 2000. An Introduction to Acoustic Ecology. Soundscape: The Journal of Acoustic Ecology, 1: 10–13. http://www.econtact.ca/5_3/wrightson_acousticecology.html. Accessed August 23, 2022

Zhang L, Towsey M, Xie J, Zhang J, Roe P. 2016. Using multi-label classification for acoustic pattern detection and assisting bird species surveys. Applied Acoustics, 110: 91-98. https://doi.org/10.1016/j.apacoust.2016.03.027

Zhang WJ. 2010. Computational Ecology: Artificial Neural Networks and Their Applications. World Scientific, Singapore