# Performance evaluation of multi-label classification models for the automated classification of anuran calls in audio recordings

Juan Sebastian De Valdenebro, Michael Hernandez Mera,
*sebasdeval@javerianacali.edu.co, maicoldead@javerianacali.edu.co*

***Abstract*—Using pre-trained convolutional neural network (CNN) models to identify three different Anuran species by sound in a time-frequency representation. The species include *Boana albopunctata, Physalaemus cuvieri, and Boana lundii*, This work also analyzes the performance of the models, to achieve multi-label classification. The methodology design for the project were divided into four stages: pre-processing, data augmentation, model training, and performance evaluation. The core of the project was developed in Python, for the data pre-processing stage in this project was designed a pipeline for raw data provided by the Humboldt Institute and involved trimming audio files, generating spectrograms, and merging them with the annotation files, to return a well-structured dataset for training. In the data augmentation stage, the techniques used were time stretching, time masking, and frequency masking techniques, finally the performance evaluation stage was performed extracting from the trained models (MobileNet, DenseNet121, Resnet50, InceptionV3), the F1-score metric using a 30 % of the not augmented dataset isolated from the training process and comparing the model's performance. Three experiments were conducted, varying hyperparameters and architecture, using different datasets. The best models were selected based on their performance, the best models achieved an average F1-Score of 81% for multi-label classification of the three different anuran calls specified earlier.**

***Palabras clave*—Transfer learning, Multilabel, Machine learning, Spectrogram, Anuran.**

## I. Introduction

Preservation of biodiversity is a crucial issue in today's world, and acoustic monitoring has become an essential tool for studying and conserving species, which health is intrinsically connected with the ecosystem they are living in. Therefore, a technological approach has surged based on implementing AI for acoustic monitoring, improving and automating the classification of these species, Consequently the use of deep learning for anuran call classification is a promising area of research, but there are still significant knowledge gaps and one of the main challenges is the limited amount of available data and its quality. Collecting and labeling anuran call data with accuracy can be difficult and time-consuming, limiting the size and diversity of datasets that can be used to train deep learning models. Additionally, environmental factors, can affect the quality of anuran call recordings, preventing deep learning models to accurately classify calls. Despite these challenges, these models have the potential to be a powerful tool for anuran call classification.

The purpose of this study is to demonstrate the applicability of AI models in the acoustic monitoring data classification process through the experimentation and analysis of pre-trained Convolutional Neural Networks (CNNs) in the context of transfer learning and multi-label classification to identify by their sound in a time-frequency representation (Spectrograms) of three specific anuran species (*Boana albopunctata, Physalaemus cuvieri, and Boana lundii*), present on the site denoted as INCT41, located in the Bifurcação locality on the Cerrado biome in Brazil. In order to accomplish this, the following methodology was divided into these four stages: pre-processing, data augmentation, model training, and performance evaluation; pre-processing techniques were used to prepare the raw audio data and annotations recorded by the Humboldt Institute on the INCT41 site. Subsequently, data augmentation techniques, Time Stretching, Time Masking, and Frequency Masking, were implemented to enhance the diversity of the training dataset. The results obtained in this

El presente documento corresponde al articulo científico del proyecto de grado del programa de Ingeniería Electrónica study will contribute to the understanding and effectiveness of using transfer learning and augmentation techniques in the context of acoustic species identification through multi-label classification. This research highlights the potential of deep learning techniques in acoustic monitoring for species classification.

## II. Methodology

The methodology followed during the project was separated into 4 stages. First, the preprocessing stage over the INCT41 audio recordings and annotations database, second the different augmentations techniques implemented over the original dataset, third the training process of the models and lastly, the final stage, the performance evaluation for each model.

As part of the deliverables, a **GitHub** repository was created, in this repository are available the created datasets used for all the experiments (Under the folder named "*Datasets*"), the coding scripts used for all de stages of the projects( Under "*SCRIPTS*" folder), the documentation for the functions developed and implemented (under "*Documentation file*" folder), all the **models** generated in this project, and a notebook of **Google Colab** designed to train the models as easy as changing a path and set the cells of the script to run.

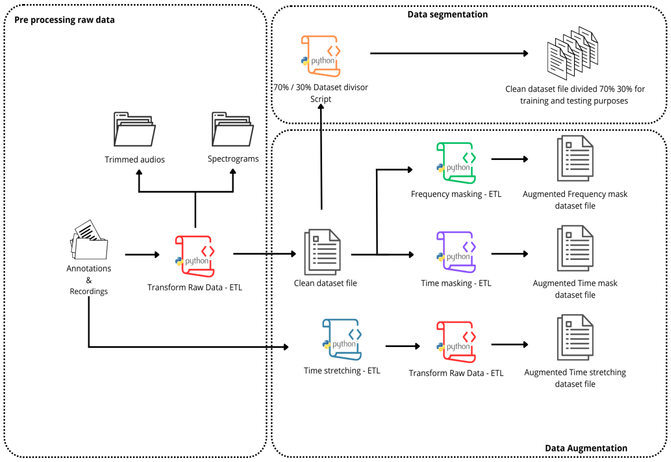

Figure 1. Pipeline of preprocessing, augmentation, and segmentation processes

## III. Theoretical framework

### A. Convolutional Neural Networks:

Convolutional neural networks (CNNs) are a type of deep learning algorithm that has been used in a variety of real-world applications. CNNs can be trained to classify images, detect objects in an image, and even predict the next word in a sentence with incredible accuracy. CNNs can also be applied to more complex tasks such as natural language processing (NLP).[1]

### B. Sound Theory:

The sound theory is the study of the properties of sound waves and how they are detected, recorded, and analyzed. Sound waves are produced by a vibrating object, which causes pressure waves to propagate through a medium, such as air or water. The properties of sound waves include amplitude, wavelength, and frequency. Amplitude is the measure of the strength of the sound wave, and it determines the loudness of the sound. Wavelength is the distance between two consecutive peaks or troughs of the sound wave, and it determines the pitch of the sound. Frequency is the number of cycles per second, and it is measured in hertz (Hz).[2]

### C. Transfer Learning:

Transfer learning, used in machine learning, is the reuse of a pre-trained model on a new problem. In transfer learning, a machine exploits the knowledge gained from a previous task to improve generalization about another.[3]

### D. F1-score:

The F1-score metric uses a combination of precision and recall. In fact, the F1 score is the harmonic mean of the two. Now, a high F1 score symbolizes high precision as well as high recall. It presents a good balance between precision and recall and gives good results on imbalanced classification problems.[4].

## IV. Results

### A. Models training stage 1

This experiment aimed to explore the impact of varying hyperparameters on the performance of the convolutional neural network (CNN) models. Four CNN architectures were selected based on previously mentioned criteria, by modifying several key hyperparameters, including the learning rate, batch size, number of epochs, and regularization techniques (L1 and L2). Selected Models: DenseNet121, ResNet50, MobileNet, and Inception V3. To begin with, the models were trained using the original dataset and different combinations of hyperparameters to assess their effects on the performance of the selected models. The learning rate was varied from 0.0001 to 0.1, incrementing in steps of 10, this was repeated for every model adding in L2 regularization. In total 22 models were trained variating the epochs, lr, L2, and L1 regularization techniques. Here is a table with the top results obtained from these 22 trained models

Table I
PERFORMANCE METRICS TABLE OF FIRST EXPERIMENT EVALUATION, TO DETERMINE THE BEST HYPERPARAMETERS.

| Model Name | F1 Score |
|---|---|
| MobileNet_Reg_L2_lr_00001_Batch_32 | 0.680 |
| Resnet50_Reg_L2__lr_0001_Batch_32 | 0.567 |
| InceptionV3_Reg_L2_lr_00001_Batch_32 | 0.579 |
| DenseNet_Reg_L2_lr_0001_Batch_32 | 0.598 |

After conducting the experiments, it was observed that the performance of the models varied significantly with different hyperparameter settings. Notably, the choice of learning rate had a substantial impact on the convergence and overall F1-Score of the models. Lower learning rates, such as 0.0001 and 0.001, resulted in slower convergence but often yielded a higher F1-Score. On the other hand, higher learning rates, such as 0.01 and 0.1, led to faster convergence but occasionally suffered from suboptimal F1-Score. Regarding the batch size, it was found that if a batch size larger than 32 is intended to be used, it is needed more than 16 GB of RAM memory due to the nature of the dataset composed of images that were transformed into Numpy arrays, therefore a batch size of 32 was standardized during the study.

### *B. Models training Stage 2*

Based on the previous results, a series of hypotheses were made to orientate the strategies for this training stage:
Hypothesis:

- The addition of n fully connected layers to a pre-trained model can improve the model's ability to learn complex non-linear relationships in the data, resulting in improved performance for the task.
- The addition of $n$ fully connected layers may result in overfitting the model to the training data, leading to poor generalization performance on unseen data.
- Models with a lower number of parameters can perform better for this multi-label classification problem.

The workflow for this stage was similar to the first one, but this time the architecture of the models is modified. On the output layer, this time one fully connected layer or two can be added depending on the experiment. The specifications of these fully connected layers are the following:

- A Dense layer with 256 units, ReLU activation, and L2 regularization (with a coefficient of 0.01) follows the Flatten layer.
- A Dense layer with 128 units, ReLU activation, and L2 regularization (with a coefficient of 0.01) follows the Flatten layer.
- To mitigate overfitting, a Dropout layer with a dropout rate of 0.5 is included after each new layer is added.

Another important difference between this stage and the first one is that now each experiment that involves a hyperparameter change for a new model training process is going to be done over 3 different datasets: Time Masking Dataset, Frequency Masking Dataset, and Original Dataset.

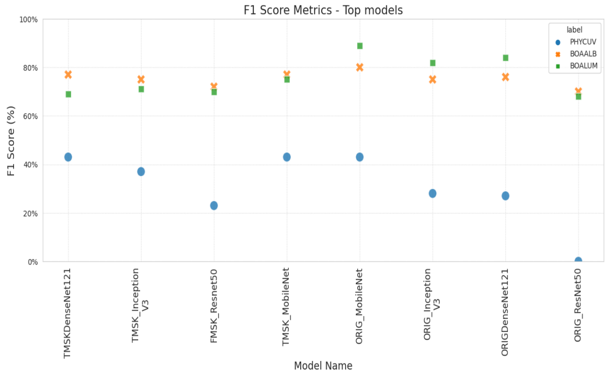

Figure 2. F1 – Score of best models of the stage, in this performance evaluations the models trained in the augmentation techniques known as time masking (TMSK), frequency masking(FMSK) and the ones trained with no augmentation techniques labeled as ORIG. There's a clear tendency over the BOAALB and BOALUM labels achieving metric values between 65 and 90 percent of effectiveness in the classification process.

**Findings:**

- MobileNet trained models have the best performance overall trained models in stages 1 and 2.
- Adding 2 fully connected layers achieves a better performance in the models over other architectures while using a not-augmented dataset and time mask-augmented dataset.
- The combination of Adam optimizer, binary cross-entropy, and early-stopping has been demonstrated to be a useful combination to evaluate and improve the real performance of the trained models.

### C. *Models training stage 3*

In this stage, the objective was to evaluate the performance of pre-trained models (Inception V3, DenseNet121, MobileNet, and ResNet50) on two new datasets following almost the same experimental structure made in Stage 2. One dataset was built with the time-stretching augmentation technique, and the second dataset comprised the original data along with all the previous augmented datasets (time-mask, frequency-mask, and time-stretching techniques). Then determine if this new augmentation technique will improve the performance metrics and if the mix of all datasets in the training can also improve the metrics of the models.

Here are the best F1-score performing models in this stage, trained on the Combined dataset :

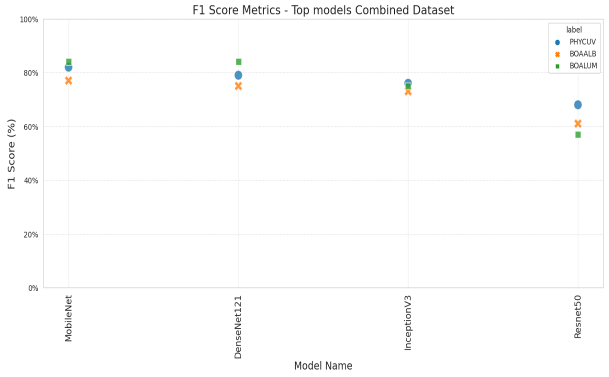

Figure 3. F1 – Score of the best models trained using the Combined dataset, here an overall improvement on all the labels to predict is shown, demonstrating that the combination of all the augmentation techniques represents an improvement for the trained models.

These results indicate that the MobileNet model with one fully connected layer exhibited the highest F1-Score values for the PHYCUV, BOAALB, and BOALUN species multi-label classification task, followed closely by DenseNet121 with one layer fold. Inception V3 with two layers and k fold = 0 and ResNet50 with one layer displayed relatively lower F1-Score values. Here are two tables to compare the standard deviation and the mean of all metrics extracted of the models trained with the combined dataset using cross-validation table 7.7 and using normal training and cross-validation table 7.8, the table highlights the significance of cross-validation as a valuable tool for assessing the stability and robustness of model predictions. By calculating the standard deviation of performance metrics across multiple folds, important insights can be obtained. Notably, a smaller variability indicates greater reliability and consistency in the predictions made by the models. In this particular cases, can be noticed that Resnet50 performs worst over all models and according to the standard deviation MobileNet and Densenet121 good candidate to be used for real applications. Based on Figure 3, an analysis was performed according to the comparison

Table II
STANDARD DEVIATION AND MEDIAN FROM ALL THE MODELS TRAINED WITH THE COMBINED DATASET USING CROSS-VALIDATION

| Cross Validation | | |
|---|---|---|
| Model | Median | Standard Deviation |
| MobileNet | 66% | 0.149756441 |
| DenseNet 121 | 65% | 0.147942536 |
| Inception V3 | 67% | 0.15302091 |
| Resnet50 | 54% | 0.219188461 |

between the model's weights and their average F1 score result, this analysis was made over this graph because these were the best results achieved.

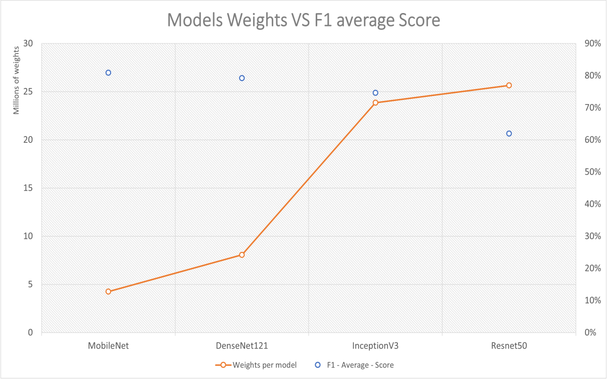

Figure 4. Model Weights VS F1 – Average Score – combined dataset, in this graph the relationship between each model weights and the model's trained with the combined dataset average performance, showing the best performance was achieved by the model with fewer weights.

As can be noticed, while the model weights increase, the average performance starts to decrease. An explanation is that the model with fewer weights is less likely to overfit the data, this can happen when a model has too many parameters, allowing it to learn the noise in the data as well as the signal intended to classify, so a model with fewer weights is less likely to overfit because it has fewer parameters to learn. Therefore, the model with a higher number of parameters requires more data samples to decrease overfitting.

## V. Conclusions

The study's purpose was to demonstrate the applicability of AI models in the acoustic monitoring data classification process through the experimentation and analysis of pre-trained Convolutional Neural Networks (CNNs) in the context of transfer learning and multi-label classification to identify by their sound in a time-frequency representation (Spectrograms) of three specific anuran species (Boana albopunctata, Physalaemus cuvieri, and Boana lundii), present on the site denoted as INCT41, located in the Bifurcação locality on the Cerrado biome in Brazil, the obtained results show that is possible and appropriate to implement pre-trained CNN models to classify anuran calls into the acoustic monitoring field, in this case the best performance achieved had an average F1-Score of 81% with a maximum standard deviation of 0.039, indicating that a proper architecture and hyperparameters were selected to achieve this task, there was identified a relation between the models weights and its performance. Therefore, it is important to conclude that due to the imbalance on the amount of samples provided and the size of the dataset generated, augmentation were used and the combination of these techniques to build a dataset result in more robust models and better performance over all. Throughout this project, several valuable insights were gained, offering guidance for future works within the same thesis theme. Firstly, among the four pre-trained Convolutional Neural Networks (CNNs) employed, MobileNet and DenseNet121 consistently exhibited superior performance. Notably, these CNNs also possessed fewer weight parameters, suggesting their efficiency. Additionally, it was observed that combining multiple augmentation techniques into one dataset can significantly enhance model performance and robustness.

## VI. Bibliography

[1] F. meAjitesh KumarI have been recently working in the area of Data analytics including Data Science and K. A. Machine Learning / Deep Learning. I am also passionate about different technologies including programming languages such as Java/JEE, *Real-world applications of convolutional neural networks*, Nov. 2021. [Online]. Available: https://vitalflux.com/real-world-applications-of-convolutional-neural-networks/.

[2] E. Browning, R. Gibb, P. Glover-Kapfer, and K. Jones, *Passive acoustic monitoring in ecology and conservation*. Oct. 2017. DOI: 10.13140/RG.2.2.18158.46409.

[3] J. Blanch, *What is transfer learning? exploring the popular deep learning approach.* Mar. 2020. [Online]. Available: https://builtin.com/data-science/transfer-learning.

[4] A. Bajaj, *Performance metrics in machine learning [complete guide]*, Jul. 2022. [Online]. Available: https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide.