International Journal of Applied Science and Engineering

Traditional and modern processing of digital signals and images for the classification of birds from singing

Camargo Luis [1*], Gasca Maira [2], Linero Rafa [1]

[1] *Faculty of Engineering, Universidad del Magdalena, Santa Marta, Colombia*
[2] *School of Basic Sciences, Technology and Engineering, Universidad Nacional Abierta y a Distancia, Santa Marta, Colombia*

ABSTRACT

Avitourism depends on the understanding of birds and the intention of birdwatchers to see or hear a specific species. Increase of this activity strengthens the economy of communities and helps finance biodiversity conservation projects. Technological products that incorporate traditional and modern tools for signal and image processing facilitate the tracking, classification and observation of birds. This article has two approaches. The first one proposes and evaluates a lightweight classification model that uses feature vector extracted from the bird's song spectrum and is based on comparison of Euclidean distance between sample features and a set vector by species. The second approach adapts and evaluates convolutional neural network architectures for bird classification using the spectrogram of the bird's song. The methodology applied in both approaches consists of: pre-processing, feature extraction, classification and evaluation metrics. The main results are the feasibility of the proposed lightweight classification model with an accuracy of 0.8 and a loss of 2.32, and the feasibility of using convolutional neural networks with an accuracy above 0.9 and a loss of less than 1, in the ResNet50, VGG19, and InceptionV3 architectures, this using as a minimum 30 spectrograms per species during training. It is concluded that the model that best meets the needs of fewer samples and less computational resources required for training is ResNet50. Additionally, it is discussed to combine the two approaches in a hierarchical and hybrid classification model that allows to introduce in the reduction and classification layers of the neural network features of another type and of other sources.

Keywords: CNN, ResNet50, VGG19, InceptionV3, DFT, spectrogram, Euclidean distance.

OPEN ACCESS

Received: June 14, 2023
Revised: September 8, 2023
Accepted: September 27, 2023

Corresponding Author:
Camargo Luis
lcamargoa@unimagdalena.edu.co

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

1. INTRODUCTION

Trips with recreational activities that focus on nature are recognized as nature-based tourism. These activities must not negatively impact the environment (Quintana, 2017). This type of tourism is motivated by the observation and appreciation of biodiversity and the culture of the populations. This tourism sector is divided into three sub-products: ecotourism that includes bird watching, whale watching, and visits to pristine landscapes; rural tourism that contemplates the cultural landscape of the regions and their traditional activities; and adventure tourism that involves exploration, risk and extreme physical activity.

Ecotourism is one of the most sustainable tourism activities, because Ecotourists are sensitive to nature conservation (Hvenegaard and Dearden, 1998). Birdwatching and avitourism is growing worldwide (Afanasiev, 2022). The viability of this tourism depends on the understanding of birds and the intention of birdwatchers to see or hear a particular species (Steven et al., 2021). Facilitating the understanding and location of some birds drives the growth of this activity. The growth of this activity strengthens the

economy of the communities and helps finance conservation projects (Steven et al., 2015).

The country of Colombia is a possible destination for nature-based tourism. In Colombia there are about 1900 species of birds, approximately 20% of the known bird species in the world (Donegan et al., 2016). Due to the above, Colombia is seen as a globally representative destination in bird watching tourism, for this it must consolidate differentiated, competitive and sustainable tourist offers. This type of tourism could be the motor of development of remote regions affected by the Colombian armed conflict (Ocampo-Peñuela and Winton, 2017).

On the other hand, information and communication technologies can facilitate the tracking, identification and observation of birds. These technological products can improve: the results of birdwatchers, the sighting experience of inexperienced tourists, and encourage and activate the desire for bird watching among sun and beach tourists in countries like Colombia.

Taking the above into account, the following question arises: What type of signal processing is more precise and efficient in a technological product designed for the classification of birds, with the purpose of promoting the conservation of species and fostering nature-based tourism?

Various investigations consider song as a criterion for the classification and identification of birds. Birds can make sounds to claim, alert, petition, court, among others. The song of the birds is mostly due to the courtship of the males in their reproductive season. The song presents long silences, and ordered and coherent repetitive sounds that are pleasant to the human ear (Chápuli, 2018). These sounds have sudden changes in intensity, different pitches and harmonics, and complex envelopes. The males of some species emit a very simple song and others have long, complex songs with great variation. Complex songs are difficult to identify, so they must be divided into phrases, syllables, or elements (Catchpole and Slater, 2003). Additionally, bird species are not isolated, birds interact with other birds and other living beings, this causes the following to be present in the audio recording: several repetitions of the song, noise from other living beings and the environment, and interference produced by songs of another kind. Advances in digital signal and image processing, techniques, and computations are a solution for obtaining acoustic features and bird classification from these extracted features.

Song analysis can be performed in the time domain, the frequency domain, or both. The characterization in time shows the evolution of the signal throughout its presence, and the characterization in frequency identifies the signal components according to the frequency in which they oscillate within a determined range. While the classification methods focus on the assessment of the static or dynamic differences of the characteristics extracted from the observed sample with respect to an established pattern.

In the scientific literature there are several contributions related to the extraction of characteristics and classification of bird songs. These proposals have accuracy higher than 75%. In the pre-processing of bird song data are used different mathematical tools, such as: Window Function, Filter Bank, Discrete Fourier Transform, (DFT), Power Spectral Density (PSD), Short-time Fourier transform (STFT), Discrete Wavelet Transform (DWT), Mel Frequency Cepstral Coefficients (MFCC), Spectrogram, among others. These tools convert the audio into: vector, curve, matrix or image that are later used to extract the features used in the classification.

For the extraction of characteristics, reduction of dimensions, and classification of the edge, methods have been used: simple or complex, traditional or modern. Some of the tools used are mentioned below: Mean, Deviation, Variance, Covariance, Correlation, Euclidean Distance, Spectrographic Cross-Correlation (SPCC) (Khanna et al., 1997; Chen et al., 2020), Discriminant Function Analysis (DFA) (Chen et al., 2020), Linear Discriminant Analysis (LDA) (Hsu et al., 2018), Principal Component Analysis (PCA) (Hsu et al., 2018; González et al., 2019), Gaussian Mixture Model (GMM) Algorithm (Lee et al., 2008), Vector Quantization (VQ) (Lee et al., 2008), Dynamic Time Warping (DTW) (Kogan and Margoliash, 1998), Hidden Markov Model (HMM) (Kogan and Margoliash, 1998), Support Vector Machines (SVM) (Fagerlund, 2007), Decision Tree (Chen and Li, 2013), Artificial Neural Network (ANN) (Sukri et al., 2020) ANN of Self Organizing Map (SOM) types (Tanttu et al., 2003), Recurrent Neural Networks (RNN) (Wan et al., 2022), Convolutional Neural Networks (CNNs) (Zhang et al., 2019), among others.

The study by Khanna et al. (1997) the SPCC uses the spectrograms of the songs of two birds to simultaneously analyze the frequency, amplitude and time, and establish a correlation between them by means of a single coefficient, a value of the coefficient close to one means that the spectrograms are similar. and it is the same species of bird. In Chen et al. (2020), a window is made to divide the complex song of the bird into first, second and third phrases, then the DFA and SPCC are used to determine that only the first phrases meet the requirements or discriminating variables for individual vowel identification. in the species studied. the song is divided into windows of finite duration, the DWT is applied to each window, then the correlation of the DWT coefficients in different sub-bands is established and with this the In Hsu et al. (2018) vector of descriptors is created, the PCA and LDA are used. to reduce the dimensionality of the features, and finally a simple distance-based classifier is used. In González et al. (2019) the Spectrogram of the song is obtained, to these the Eigenface technique is applied based on the PCA, to obtain the eigenvectors of the covariance matrix that best represents the spectrograms, the vectors that present the highest energy are taken, and the song is classified by comparing the training and test vectors. In Lee et al. (2008) GMM and VQ are used to represent the MFCC of the songs in different bells (cluster) or regions (vectors), to later estimate the mean of the GMM or centroids of VQ and form the

prototype vectors of a certain species, and from these, recognize the vector that best fits the song of the species to be classified using K-means. In Kogan and Margoliash (1998) DTW based long continuous song recognition is applied to the song, (synchronized search) to obtain a grid (i,j,k), i is the time frames of the input song, j is each individual template and k is the template counter, then the song is classified through the sum of local metrics d (i,j,k) and the distances between the two multidimensional vectors of the test signal and the pattern, finally the performance is compared with the HMMs. In Fagerlund (2007) the spectrogram of the song is made, then it is divided into syllables, the characteristics of the syllable are extracted using Mel-cepstrum and the set of signal parameters, these characteristics are used in an SVM (Supervised Learning) for the classification of the song. singing. In Chen and Li (2013), the spectrograms of the song are made, it is divided into syllables, the texture characteristics of the syllables are extracted, and the Random Forest method (Decision Tree) is applied for classification. In Sukri et al. (2020), the PSD of the song and the classical ANNs are used for the classification. In Tanttu et al. (2003), tracking of the first harmonic components of the spectrogram is used to extract the characteristics and SOM (Unsupervised Learning) for the classification. Wang et al. (2022) performs the Melspectrogram and MFCC of the songs, these are the inputs for the deep learning model, Long short-term memory (LSTM) is a variant of recurrent neural networks (RNN), which integrates specific gates to recover the short-term or long-term context of the input, LSTM is used to extract features and classify the song. In Zhang et al. (2019), the 3-D convolution kernels of the CNN were used to extract both positional and temporal characteristics from the Mel-spectrogram and with this improve the classification.

The works mentioned above used traditional mathematics and modern tools to classify birds and validated the method through an evaluation of its precision. In comparison to these published materials, the purpose of this research is to test accuracy and validate the computational requirements. To achieve this, two approaches are experiments

In the first experiment the accuracy of the classification is evaluated, using the fewest number of operations in the process. In this, a distance-based classifier is employed; This type of classifier has been used and tested in other works (Kogan and Margoliash, 1998; Hsu et al., 2018). The differences are in the pre-processing and feature extraction techniques. Kogan and Margoliash (1998) and Hsu et al. (2018) used tools like DWT; Khanna et al. (1997) and González et al. (2019) employed spectrograms, Lee et al. (2008), Zhang et al. (2019) Wang et al. (2022) use MFCC. In our case, we employ lighter pre-processing, such as DFT. For feature extraction, we collected and analyzed the mountains and peaks of the DFT. We deliberately avoided advanced statistical techniques such as PCA, as used by González et al. (2019), LDA, as used by Hsu et al. (2018), or DFA, as used in the study by Chen et al. (2020), to reduce the computational cost.

In the second experiment, the spectrogram is used for processing and the CNNs for classification. This is different from other works that use modern learning-based tools (Tanttu et al., 2003; Fagerlund, 2007; Chen and Li, 2013; Zhang et al., 2019; Sukri et al., 2020; Wan et al., 2022) because it focuses on determining the minimum number of samples required in the training if transfer learning is used in the classification, taking precision into account. The procedure is performed on three CNN structures.

The CNNs are a type of artificial neural network used widely for machine training, where artificial neurons are intended to be given the capability that neurons in the primary visual cortex have. The CNNs contain several specialized layers interconnected with each other, in which: the first layers can detect lines, curves and they are specialized until reaching deeper layers that recognize complex shapes (Alzubaidi et al., 2021). In addition, they are feed-forward, that is, in these the artificial neurons and layers do not form cycles, but move in a single direction, from the input layer, passing through the hidden layers to the output layer.

The CNNs require a large amount of input data for their training, and in these the convolutions are the differential factor with respect to other types of artificial neural networks. A convolution entails calculating a scalar product between a selected group of neighboring pixels from the input image and a small matrix known as a kernel or filter. The kernel goes through all the input neurons and generates a new output matrix (new image), with characteristics of the original image that help distinguish one object from another (Ketkar and Moolayil, 2021).

All these interconnections between multiple layers are configured in different architectures that have been designed, implemented and trained with a large number of images and objects, which is why they are commonly used in deep learning applications. With transfer learning, a previously trained network can be taken and used as a starting point to learn a new task (Mathew et al., 2021). Tuning a network with transfer learning is often quicker and easier than training a network with randomly initialized weights.

ResNet-50 is a residual convolutional neural network, created by Microsoft 50 layers deep. A residual connection implies that the output of a layer is a convolution of its input plus its input, see Fig. 1. This preserves knowledge and increases performance during training. The goal of this ultra-deep network is to be free from the vanishing gradient problem, applying the branch path concept (He et al., 2016).

The VGG19 is a convolutional neural network of the VGG16 family, created by developers at the Visual Geometry Group (VGG) at the University of Oxford. It has a total of 19 layers, 16 are convolutional and 3 are fully connected layers, see Fig. 1. VGG proposed filters with small sizes instead of large filters (Simonyan and Zisserman, 2014).

Inception-V3 is a 48-layer-deep convolutional neural network created by Google. Inception-V3 has symmetric

and asymmetric blocks. Each block consists of several convolutional layer, pooling layer (max pooling and average pooling), contacts, dropout layers, and fully connected layers, (Szegedy et al., 2016) (Fig. 1).

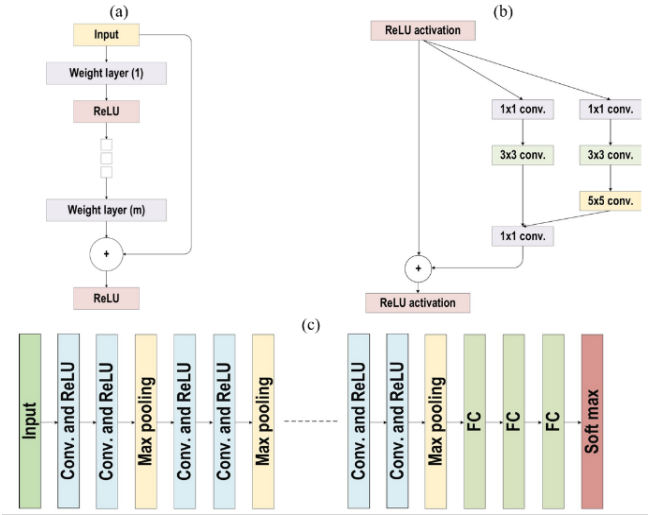

Fig. 1. CNN architectures: (a) ResNet50, (b) InceptionV3, (c) VGG19

These tools and methods are mostly designed to be implemented on computers. It is for this reason that in this work it is required to determine a light and precise method that can be implemented in small single board computers or midrange or low end smartphones. Devices that are inexpensive and can be deployed or used on a large scale over wide areas. This would facilitate and intensify the observation, study and protection of birds.

As a continuation, the structure of the paper contains the following sections: Materials and Methods, Results, Discussion, Conclusion, Acknowledgments, and References. In the Materials and Methods section, the experiment's design is described, this section is divided into data and pre-processing, feature extraction, classification, and evaluation metric. In Results, the data obtained in the evaluation are shown and analyzed applying the mentioned metrics. In the Discussion section, the results found are compared with the results of other cited studies, and future work is proposed. And in the Conclusion section, we present the main contribution of the work with respect to the results obtained is presented.

2. MATERIALS AND METHODS

The following methodology is designed and articulated for the two bird song classification approaches. Light and traditional approach and the CNN approach.

2.1 Data and Pre-processing

The audio recordings of birdsong used for extraction, training and validation are obtained from the Xeno-Canto foundation (2021) and the Alexander von Humboldt Research Institute for Biological Resources (2020). The songs belong to 31 species of birds, the number of audios obtained are 574. The audios belong to bird songs recorded in locations in Colombia. The differentiation between song, claim and request and location is done manually. The audios are homogenized into: samples per audio channel of 1024 (length_buffer) and a sample rate (fs) of 44100 Hz. These recordings are separated into 2 groups and used according to the proposed classification models.

2.1.1 Data and Pre-processing for the Proposed Lightweight Model

Group 1 has 403 audios. These are used for feature extraction and validation of the proposed light classification model, 13 audios for each of the 31 species to be classified. The recording group 1 is applied the DFT. See Equation 1, where x[n] is the input vector to analyze, N is the length of the buffer x[n], X[k] is the DFT of the signal x[n], and fs is the sample rate (Wang et al. 2022). Then the DFT is estimated in decibels (dB) at full scale (Decibels Full Scale, dBFS) Y[k], to define the amplitude levels based on the maximum available level, where 0 dB is the maximum audio power, see Equation 2. Next, the DFT elements below -40 dB (ambient noise) are removed.

$$X[k] = \sum_{n=0}^{n=N-1} x[n] e^{\frac{-j2\pi kn}{N}} \quad (1)$$

$$Y[k] = 20 \log \frac{X[k]}{X_{maximum\ possible}} \quad (2)$$

2.1.2 Data and Pre-processing for the CNN Models

Group 2 has 210 audios. These are used to train and validate the ResNet50, VGG19 and InceptionV3 CNN models. 70 audios for each of the three species to be classified (Mimus gilvus, Crypturellus soui and Cistothorus apolinari).

The experiments were initially planned to be conducted with the same number of species and using songs collected exclusively in Colombia. However, this was not feasible due to the limited number of songs available for each species in the Xeno-Canto foundation at the time of the study. The 31 species in experiment 1 did not have a sufficient number of samples. Furthermore, for experiment 2, it was essential that the classes to be classified were balanced and had a large number of songs for each of the species. Taking these factors into consideration, only three species were chosen for experiment 2.

These species are chosen for having a high number of songs stored in Xeno-Canto foundation (2021). These birds are shown in Fig. 2. The Spectrogram is performed on group 2 of recordings. Spectrogram is a graphic representation of the variations of frequency and intensity (colors) over time of the audio signal. This is done based on the STFT. The signal is divided into time windows w[n] of the same duration (length m) and the DFT is applied to each window.

These windows overlap to avoid spectral ringing (Zhang et al., 2019). See Equation 3.

$$STFT\{x[n]\} = \sum_{n=-\infty}^{n=\infty} x[n]w[n-m]e^{-2\pi fn} \quad (3)$$

The result is a 2D matrix, with the first dimension representing the frequency segments and the second dimension corresponding to the time indices. Our spectrograms have a Resolution Bandwidth (RBW) of 21.53 Hz ((fs/2)/ length_buffer), a time resolution of 0.046 s (1/RBW), and use the color map shown in Fig. 3, to represent intensity.

The 210 spectrogram images are edge cropped, resized based on CNN input (ResNet50 224 x 224, VGG19 224 x 224 and InceptionV3 299 x 299) and separated into 3 classes based on the bird species they correspond. 30% of the images are marginalized to validate the model, and 70% are used for training. Validation images are chosen randomly.

Fig. 4 shows an example of the images that are input to the CNN for input pre-processing. Subsequently, the input preprocessing function (preprocess_input) designed specifically for each neural network is applied. This function regularly normalizes the intensities of the image pixels.

2.2. Feature Extraction

2.2.1 Feature Extraction in the Proposed Lightweight Model

For the lightweight classification model, features are extracted from the Spectrum and stored in two vectors. The lowest frequency (fmin) and the highest (fmax) and the frequencies with the four highest tones (mountain peaks) were identified. To define the values of the vector elements, three (j = 1, 2, 3) bird songs are analyzed for each fully identified species and the average of these values is obtained. See Equations 4 and 5, and Figs. 5 and 6.

$$R_j = [f1_j, f2_j, f3_j, f4_j] \quad (4)$$

$$BW_j = [fmin_j, fmax_j] \quad (5)$$

2.2.2 Feature Extraction in the CNN Models

CNN uses the Convolutional Layer, Activation Layer and Pooling Layer to extract the features. Convolutional Layer extract important features by identifying local relationships among the data points in the input layer. The feature vector is obtained by convolving the kernel over the image pixels and summing the pixels (Ali et al., 2021). See Equation 6.

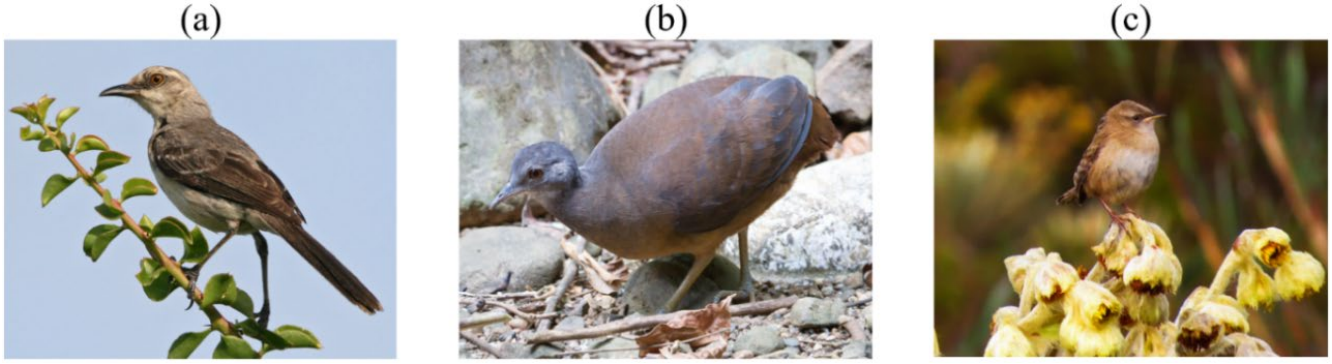

Fig. 2. Birds to classify with CNN; (a) Mimus gilvus, (b) Crypturellus soui, (c) Cistothorus apolinari

Fig. 3. Color map of Spectrogram

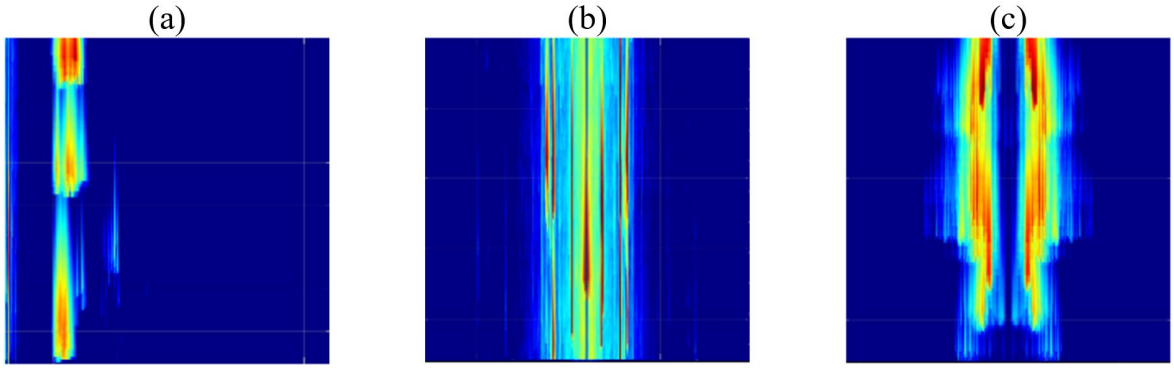

Fig. 4. Spectrogram of Birds to trained: (a) Mimus gilvus, (b) Crypturellus soui, (c) Cistothorus apolinari

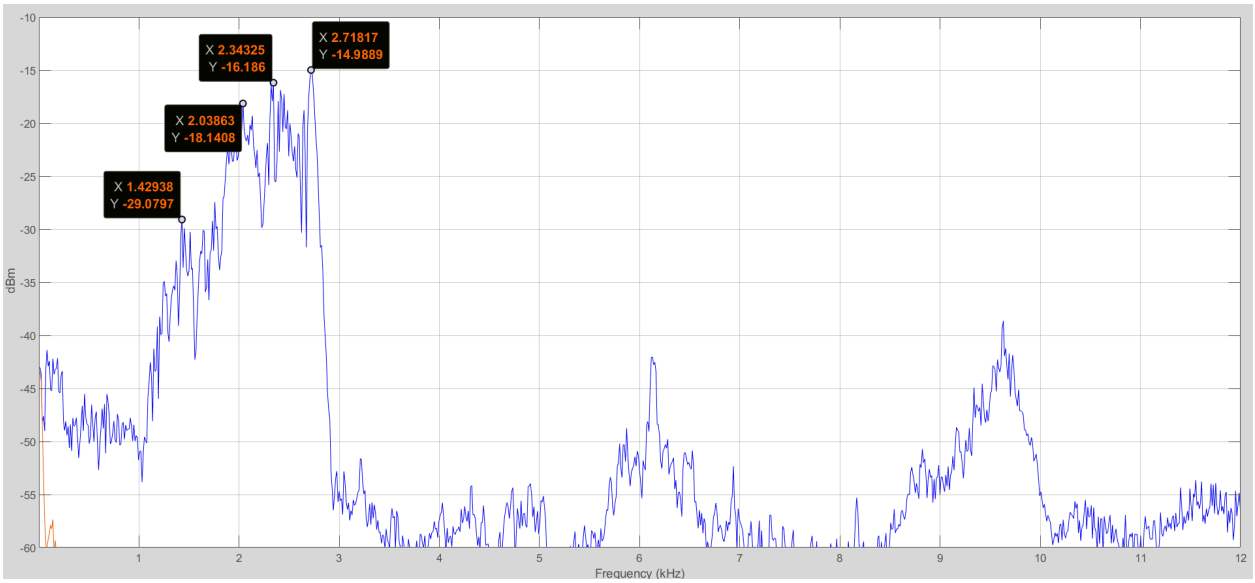

Fig. 5. Feature extraction of spectrum

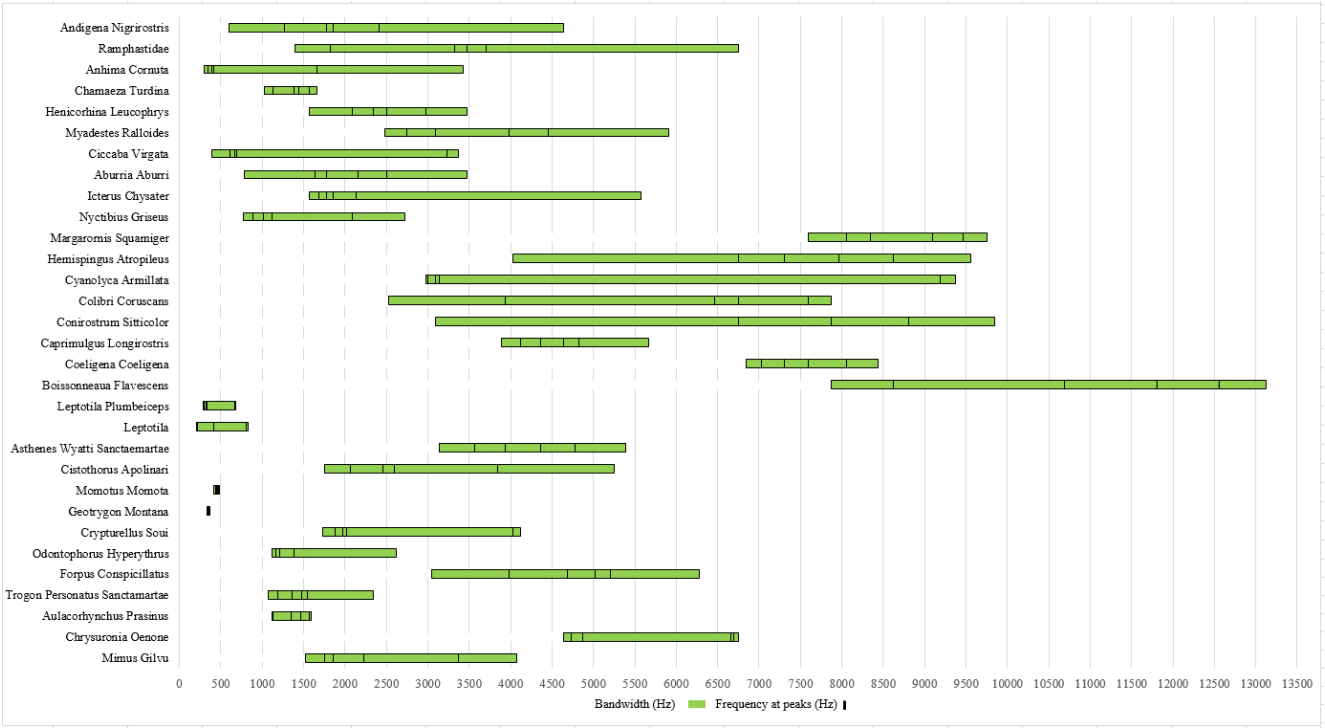

Fig. 6. Characteristics vectors

$$Feature\ Vector = \sum(I_{k\times k} + W_{k\times k}) + B \quad (6)$$

Here, $I_{k\times k}$ is the input local receptive field on which the convolution operation is performed. $W_{k\times k}$ and B are filter weights, kernel size, and filter bias. The obtained feature maps are entered in the activation layer (Ali et al., 2021).

The Activation Layer applies a Rectified Linear Unit (ReLU) transformation to the feature map, introducing nonlinearity into the model (Nair and Hinton, 2010). In Equation 7, F(I) is the output of ReLU and I is the input, F(I) is I if I is positive, F(I) is 0 if I is negative (Ali et al., 2021).

$$F(I) = \max(0, I) \quad (7)$$

The Pooling Layer divides the feature map into small, non-overlapping pooling kernels. This layer reduces the dimensionality of the data and the number of model parameters. There are two main types of pooling: max pooling and average pooling.

The training phase of the models is configured to

maintain the weights of the layers of the previously trained base models, and the training is added with the images of our spectrograms. For training, 15 epochs are used. The training is carried out with a different number of input images: 10, 20, 30, 40, 50, 60 and 70 spectrograms for each species. The training is done without the data augmentation technique and with data augmentation. The data augmentation technique consists of generating a sequence of images that are used in training from the horizontal rotation (10 degrees) and zoom (10%) of the original images.

2.3. Classification

2.3.1 Classification in the Proposed Lightweight Model

In the simple method of classifying the bird from the song, the process shown in the flowchart of Fig. 7, is implemented. The song of the bird (x [n]) to be identified is captured and then the spectral analysis (Y [k]) is performed. From Y[k], the frequency tones are identified and the recorded song is characterized, obtaining the record with the lowest and highest frequencies (CWB [fmin, fmax]) and the record with the frequencies with the peaks (CR [f1, f2, f3, f4]).

Using the information of the characterization vectors of the 31 songs (j = 31) previously stored in the records BW_j [$fmin_j$, $fmax_j$] and R_j [$f1_j$, $f2_j$ $f3_j$, $f4_j$] and the records CWB [fmin, fmax] and CR [f1, f2, f3, f4] with the information of the recorded song we proceed to identify the bird. First, a pre-classification is carried out from the comparison of the fmin and fmax of the CWB vector with all the BW_j records, all birds whose song is not within the registered bandwidths are discarded. If there are no preselected birds, the value of "unidentified bird" is returned. Subsequently, the Euclidean distances are calculated by comparing the CR [f1, f2, f3, f4] with all the R_j [$f1_j$, $f2_j$ $f3_j$, $f4_j$], establishing the decision variable U for each case. The bird with the lowest U value is recognized as the bird that sings.

2.3.2 Classification in the CNN Models

CNNs use the Fully Connected Layer and Softmax Layer for classification. The Fully Connected Layer converts the three-dimensional array obtained from the previous layers into a one-dimensional vector using a convolution operation. See Equation 8.

$$Z_{V_0\times 1} = W_{V_0\times V_i} I_{V_i\times 1} \,.\, B_{V_0\times 1} \tag{8}$$

V_i and V_o are the input and output vector size, Z is the output of layer, W is weight. I, is input and B is Bias (Ketkar and Moolayil, 2021).

In Softmax Layer, a standard exponential function is applied to each element and then these values are normalized by dividing by the sum of all exponentials, to ensure that the sum of all output values is 1, as shown in Equation 9. Softmax Layer with this calculates the normalized class probabilities for each class in n classes. This information ("confidence score") is used for classification.

$$P(\vec{Z}) = \frac{e^{Z_i}}{\sum_{j=1}^{n} e^{Z_j}} \tag{9}$$

Where, $P(\vec{Z})$ is the output vector and represents the confidence score vector. $(\vec{Z})$ is input vector to the softmax function. Zi are the elements of the input vector. e^{Z_i} is the standard exponential function is applied to each element of the input. The summation is the normalization term, and n is the number of classes in the multi-class classifier (Ketkar and Moolayil, 2021).

In this work, the base models for each CNN (ResNet50, VGG19 and InceptionV3) previously trained with the ImageNet database are used, without the final classification layers. The final layers are customized to: reduce the dimensions (Pooling Layer) of the feature maps (GlobalAveragePooling2D), model generalization by reducing overfitting (Dropout Layer) by removing 20% of the input values (keras.layers.Dropout (0.2)), create a Fully Connected Layer with 512 neurons with rectified linear unit, and create a fully connected output layer (Softmax Layer) with 3 neurons to classify the 3 birds. Finally, the model is built with the base model and the final custom layers.

To reduce the error made by the CNN, the learning rate is optimized in the model compilation process. In our case we always employ the Adaptive Moment Estimation (Adam) optimizer. This optimizer adjusts the weight matrix (W) using the gradient descent method using adaptive estimation of first and second order moments. In addition, it is computationally efficient and has low memory requirements (Kingma and Ba, 2014).

2.4. Evaluation Metrics

To evaluate the performance of the models, the metric of Equation 10 is used.

$$\text{Accuracy} = \frac{correct\ classifications}{all\ classifications} \tag{10}$$

In addition, the loss function that measures how well the model is performing in the specific problem is used, for this, sparse categorical crossentropy is used. See Equation 11. Where Y_{true} is the truth data, Y_{pred} is model's predictions (Ketkar and Moolayil, 2021; Wang et al. 2022).

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} Y_{true} \times ln\,(Y_{pre}) \tag{11}$$

Fig. 8 shows an example of applying these metrics in training and validation in each epoch. Additionally, the use of RAM in the training and validation of each of these architectures is analyzed, as a metric for the consumption of computational resources.

The confusion matrix is also used as an evaluation metric. It is employed to assess how models classify the different

classes. This tool helps identify the number of correctly and incorrectly classified samples for each class. The matrix records the location and count of cases of True Positives (TP), False Positives (FP), and False Negatives (FN) for each class. This indirect evaluation allows for an assessment of accuracy, recall, specificity, and F1-score, providing an overview of the classification model's performance.

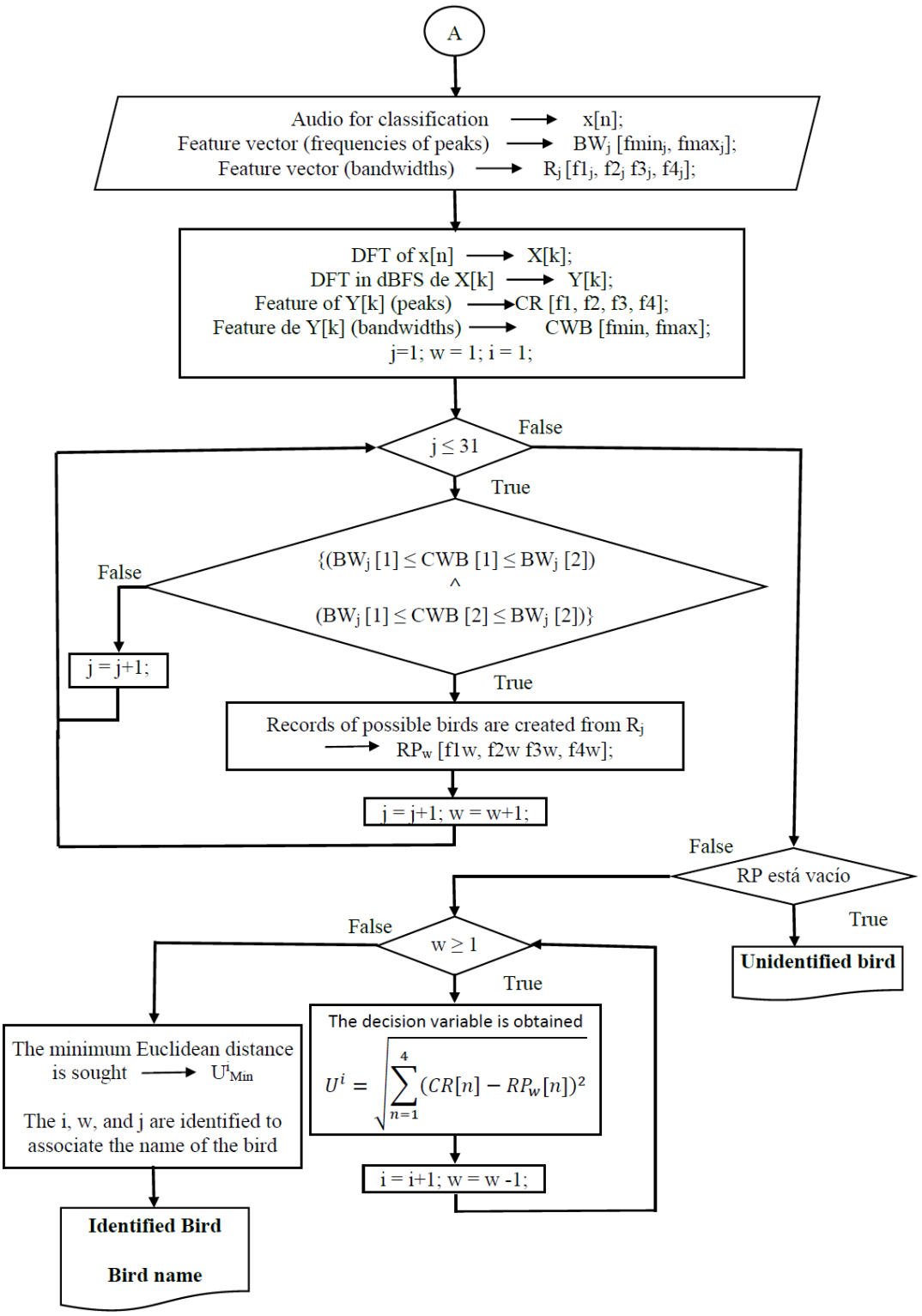

Fig. 7. Classification algorithm

3. RESULTS

3.1 Evaluation Metrics in the Proposed Lightweight Model

Fig. 9, shows the accuracy of proposed lightweight model. The model has an accuracy of 0.8. Additionally, it is identified that 72% of the incorrect ones are due to an error in the classification, and the remaining 28% of failures are because the spectrum cannot be associated with some species in the database.

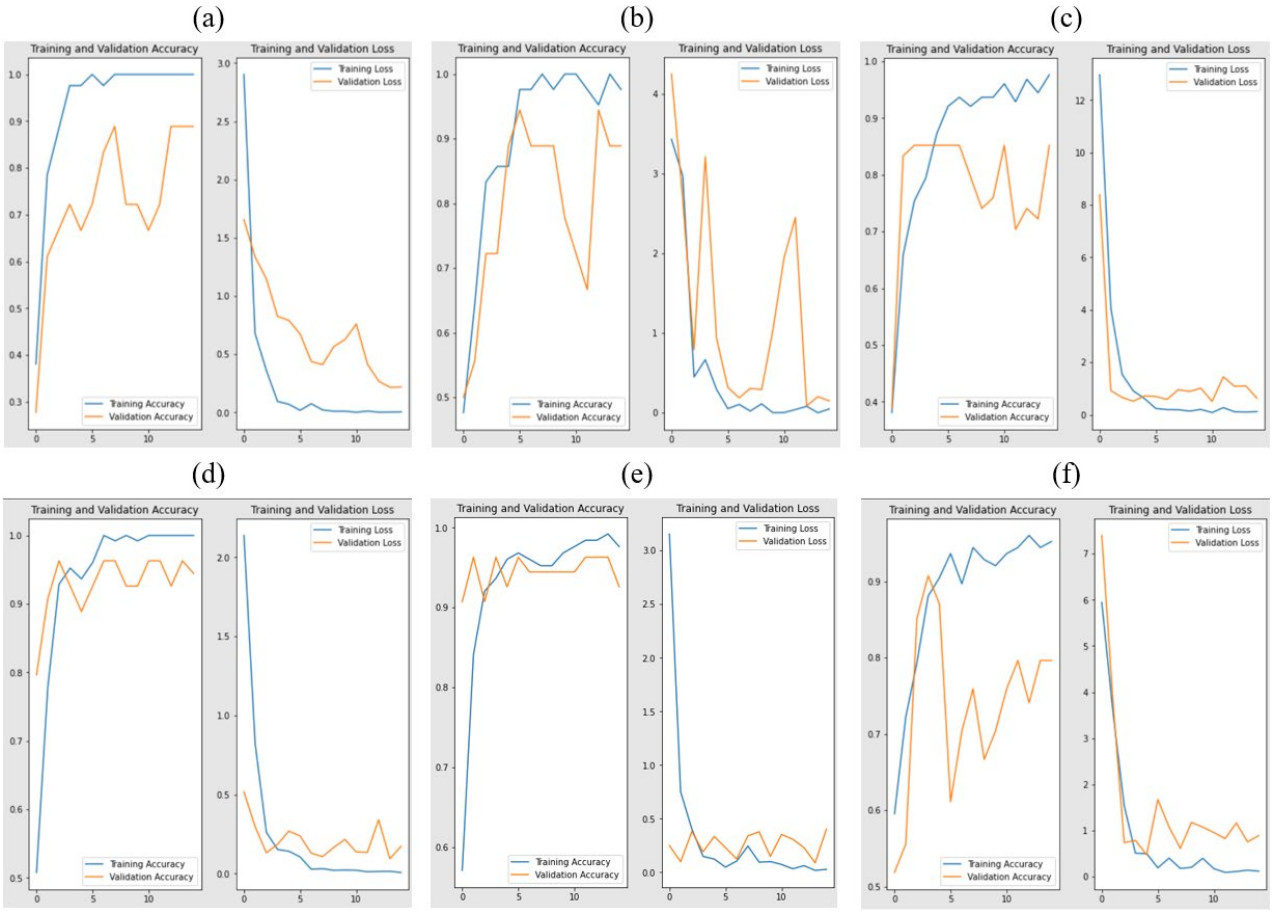

Fig. 8. Precision and Loss in the training and validation, with input of 60 spectrograms per species to the model: (a) ResNet, (b) VGG19, (c) InceptionV3, (d) ResNet with data augmentation, (e) VGG19 with data augmentation, (f) InceptionV3 with data augmentation

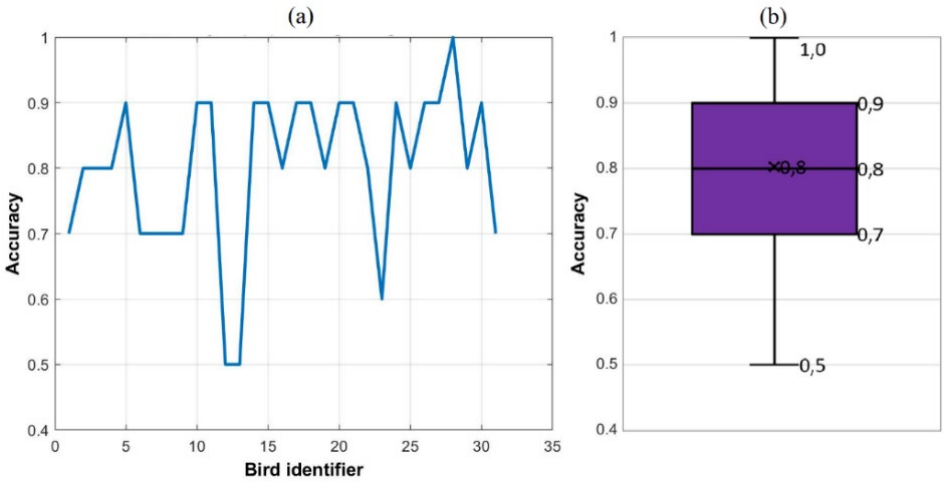

Fig. 9. Validation Accuracy: (a) by species, (b) boxplots of accuracy

Proposed lightweight model has some losses (sparse categorical cross-entropy) of 2.32. The birds that most influence losses are Leptotila plumbeiceps and Leptotila verreauxi. These are birds of the same family and order (Columbidae, Columbiformes) and have very similar songs. RAM usage is very low compared to CNN models. This is below 1 GB.

3.2. Evaluation Metrics in CNN Models

Accuracy and Loss of the CNN models analyzed as a function of the number of spectrograms per species used for training and validation are shown in Figs. 10 and 11. The model that has the best accuracy and lowest loss with the least number of samples for training is the ResNet 50. All the CNN models analyzed have an acceptable accuracy and an acceptable loss, with an input of more than 30 spectrograms per species.

The data augmentation technique does not represent a significant improvement in the accuracy of the models. This is due to the natural verticality of the spectrogram. If 30 spectrograms per class are used, and 30% of these samples are used for validation, the confusion matrix shown in Fig. 12 is obtained. The matrix shows that there is difficulty in classifying the species Crypturellus Soui. In the validation, false positives are observed in the VGG19 and InceptionV3 models.

Fig. 13 shows that there is no significant variation in RAM consumption depending on the number of spectrograms per species used for training and validation. It can also be seen that the model that requires slightly less RAM is ResNet50, this is due to its architecture and way of working.

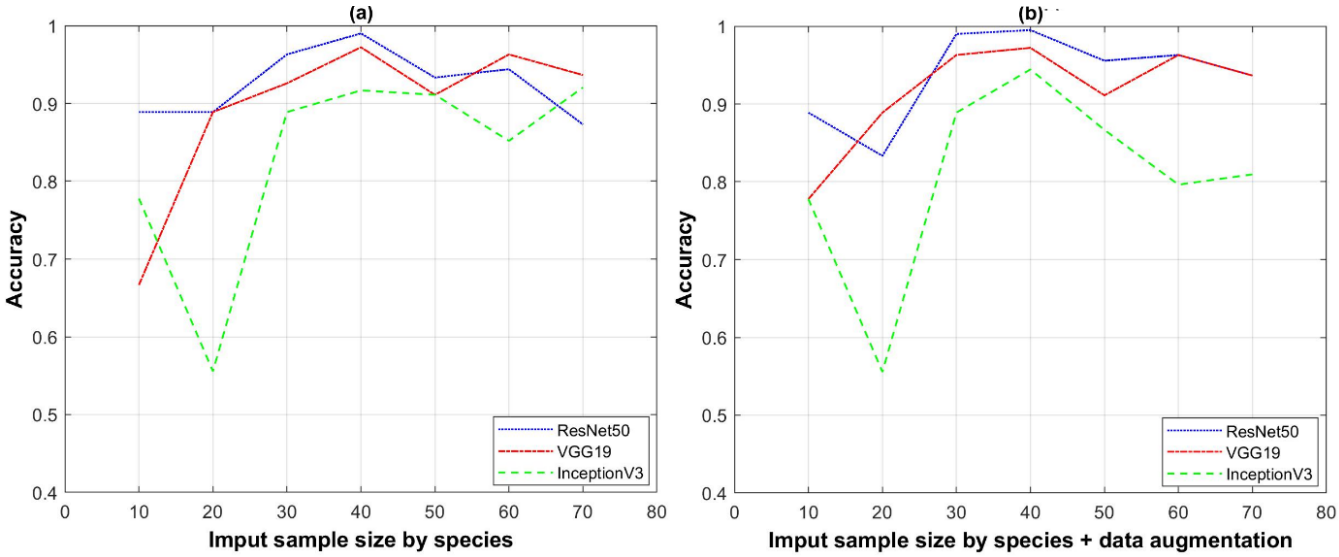

Fig. 10. Validation Accuracy: (a) without data augmentation, (b) with data augmentation

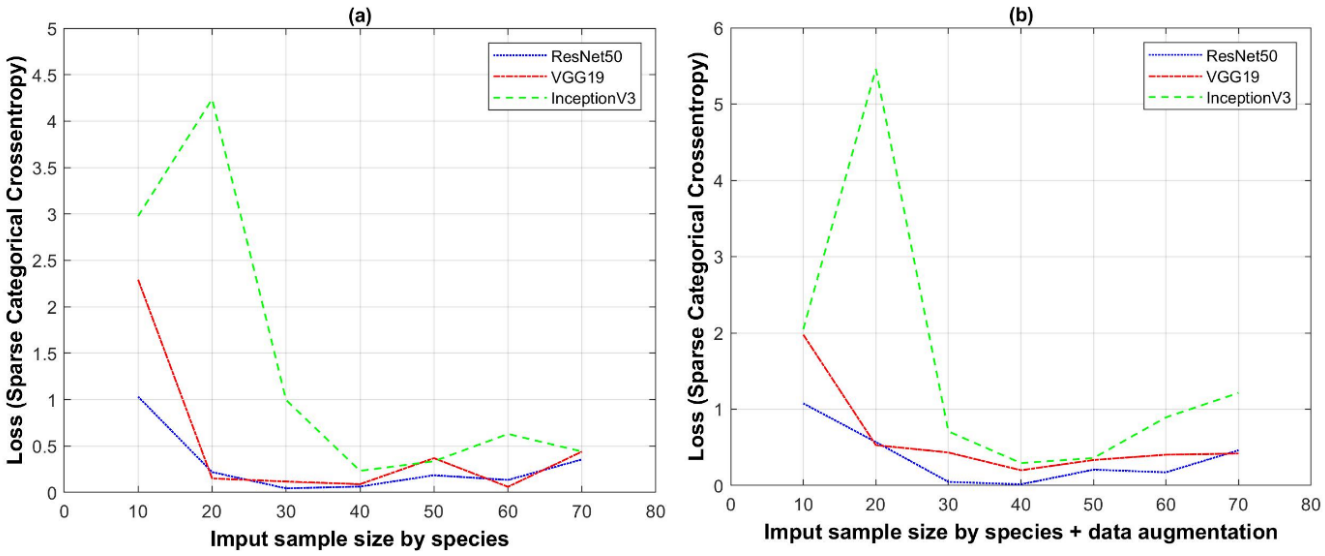

Fig. 11. Validation Loss: (a) without data augmentation, (b) with data augmentation

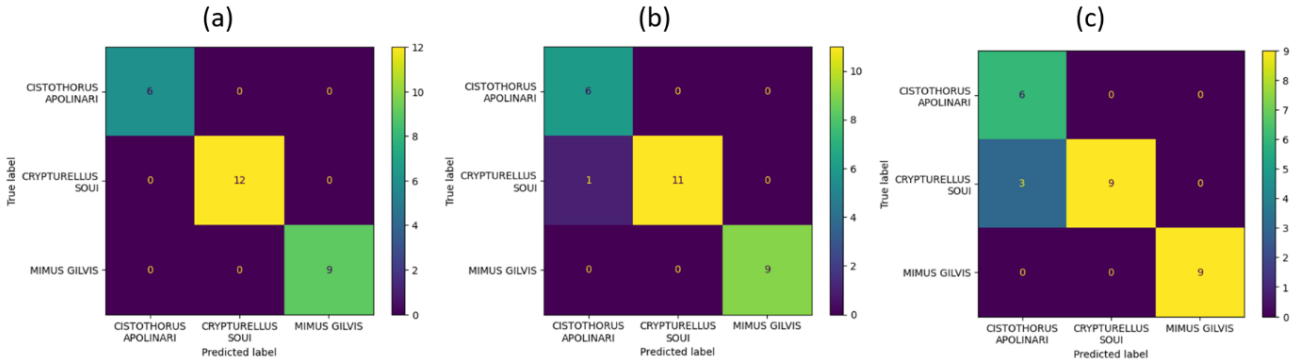

Fig. 12. Confusion matrix: (a) ResNet50, (b) VGG19, (c) InceptionV3

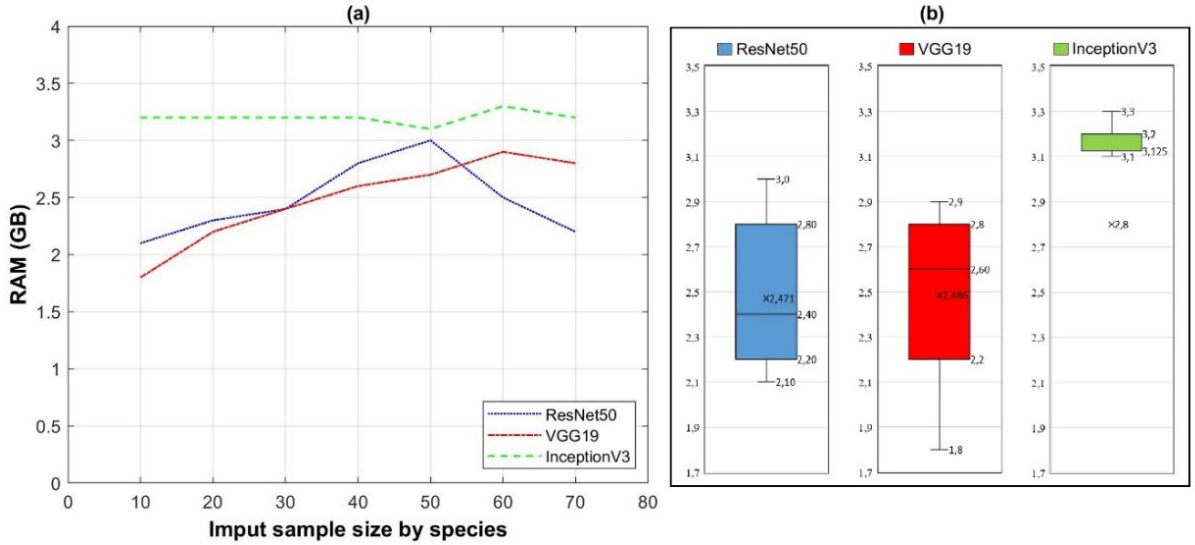

Fig. 13. RAM usage, (a) depending of number of input sample per species, (b) boxplots of RAM usage in models

4. DISCUSSION

The results of accuracy, loss and computational requirements obtained in the proposed and adapted models are acceptable when compared with those published in previous studies.

For example: Wang et al. (2022) use MelSpectrogram and MFCC fusion as input for the deep learning model, Zhang et al. (2019) use the continuous frame sequence of the MelSpectrogram as input, and we use as input an image with the spectrogram in a fixed window of frequency and time and a specific color map. Wang et al. (2022) use an LSTM that is a variant of RNN to classify, Zhang et al. (2019), we use an SFLN that is a variant of a CNN, and we use and evaluate 3 CNN architectures. Wang et al. (2022) utilize the songs from the Xeno-Canto Foundation. Zhang et al. (2019) take the songs from the Xeno-Canto Foundation, manually labeling the species and discarding audios with complex or erroneous sounds y we do too, but we restrict the songs to a geographic area of observation. Wang et al. (2022) classified 264 species using varying numbers of samples for training and achieved an accuracy of approximately 75%. Zhang et al. (2019) classified 4 species, utilizing different numbers of samples per species for training (with more than 337 samples per species), resulting in a precision of 97%, in our case, we classified only 3 species, obtaining varying precision values depending on the number of samples used for training, with a maximum precision of 98%. Wang et al. (2022) sought an efficient model for identifying a large number of species. Zhang et al. (2019) proposed a linear network using spectrogram frames and continuous frame sequences for classification, in our case, we are searching for an efficient and computationally lightweight model for classification. Due to the aforementioned reasons, the results of the studies cannot be directly compared, however, it can be shown that the results obtained are in an acceptable range.

The proposed lightweight model has a low computational requirement, and an accuracy lower than that of the CNN models, but this model supports the fact that the characteristics extracted (mountain peak frequencies, low and high frequency) from the DFT of the song serve to classify them.

Comparing the results obtained in the evaluations of the

three CNN architectures commonly used in classification, ResNet50 from Microsoft, VGG19 from the VGG and InceptionV3 from Google, the model based on ResNet50 is chosen as the best option to classify the song of birds. This is taking into account that ResNet50 is the model that best fits the specific conditions of the problem: less number of spectrograms of bird songs per species for training, less computational resources for training, an accuracy greater than 0.9 and a loss less than 1.

In order to improve the results, found, it is proposed in a future work to carry out a hierarchical and hybrid model that combines traditional techniques or tools with modern ones. In this future model, a ResNet50 model would be used at the beginning, the reduction and classification layers would be eliminated at the end, the feature vectors found by the previous layers would be added the feature vectors found manually in this work and vectors with other features obtained with traditional mathematical tools used in audio and image processing (MFCC, DWT, among others), this new set of vectors will be the inputs to RNN, CNN, SVM or other techniques and models used for reduction and classification

5. CONCLUSION

It is possible to classify the song of the 31 traditional birds of Colombia with a model: computationally light, with an accuracy of 0,8. Model using traditional mathematical tools. This model starts with the pre-processing of the ridge using the DFT, then extracts the features by traversing the DFT vector and identifying the frequencies where the main elements of the ridge or mountain peaks are recorded, these features are the main input of an algorithm of classification using the concept of Euclidean distance.

To assess the CNN models studied in this work, it is necessary to adapt and unify the final layers of the three architectures. In addition to measuring the evaluation metrics based on the number of images used to train the models. Additionally, the evaluation showed that these architectures have similar results that support the decision to use any of these networks, if the number of spectrograms of bird song, per species, is above 30.

The database with the vector of song characteristics of the 31 bird species cataloged as part of Colombian biodiversity by the Alexander von Humboldt Biological Resources Research Institute, and the schematization of 210 songs of three bird species (Mimus Gilvus, Crypturellus Soui, and Cistothorus Apolinari) commonly found in Colombia and the Americas in the form of spectrograms, meets the requirements of the experiments and is optimal for use in future research.

ACKNOWLEDGMENT

This research and APC was funded by Universidad del Magdalena

REFERENCES

Afanasiev, O. 2022. Birdwatching, ornitological tourism and avitourism: Terminological dispute and review of world and Russian practices. Anais Brasileiros de Estudos Turísticos, 12, 1–12.

Ali, L., Alnajjar, F., Jassmi, H., Gocho, M., Khan, W., Serhani, M. 2021. Performance evaluation of deep CNN-based crack detection and localization techniques for concrete structures. Sensors, 21, 1688.

Alzubaidi, L., Zhang, J., Humaidi, A., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M., Al-Amidie, M., Farhan, L. 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8, 1–74.

Catchpole, C., Slater, P. 2003. Bird song: Biological themes and variations. Cambridge University Press. Cambridge. UK.

Chápuli, R. 2018. ¿Hay música en el canto de las aves? Encuentros en la Biología, 11, 21–25.

Chen, G., Xia, C., Zhang, Y. 2020. Individual identification of birds with complex songs: The case of green-backed flycatchers ficedula elisae. Behavioural processes, 174, 104063.

Chen, S.-S., Li, Y. 2013. Automatic recognition of bird songs using time-frequency texture. Proceeding of 5th International Conference and Computational Intelligence and Communication Networks. Mathura, India. 262–266.

Donegan, T., Verhelst, J., Ellery, T., Cortés-Herrera, O., Salaman, P. 2016. Revision of the status of bird species occurring or reported in Colombia 2016 and assessment of BirdLife International's new parrot taxonomy. Conservación Colombiana, 27, 12–36.

Fagerlund, S. 2007. Bird species recognition using support vector machines. EURASIP Journal on Advances in Signal Processing, 038637, 1–8.

González, J., Padrón, J., Barbero, I., Custodio, L., Merchán, F. 2019. Reconocimiento de canto de aves basado en el análisis de componentes principales del espectrograma. Revista de Iniciación Científica, 5, 124–129.

He, K., Zhang, X., Ren, S., Sun, J. 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA. 770–778.

Hsu, S., Lee, C., Chang, P., Han, C., Fan, K. 2018. Local wavelet acoustic pattern: A novel time–frequency descriptor for birdsong recognition. IEEE Transactions on Multimedia, 20, 3187–3199.

Hvenegaard, G., Dearden, P. 1998. Ecotourism versus tourism in a Thai national park. Annals of Tourism Research, 25, 700–720.

Institute Humboldt. 2020. Descarga en tu celular los sonidos de la Biodiversidad colombiana. Retrieved 2022-04-12 from http://www.humboldt.org.co/es/noticias/actualidad/item/107-descarga-en-tu-celular-los-sonidos-de-la-biodiversidad-colombiana

Ketkar, N., Moolayil, J. 2021. Convolutional neural networks. In Deep Learning with Python. Berkeley: Apress. California, USA.

Khanna, H., Gaunt, S., McCallum, D. 1997. Digital spectrographic cross-correlation: Tests of sensitivity. Bioacoustics, 7, 209–234.

Kingma, D. P., Ba, J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kogan, J., Margoliash, D. 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. The Journal of the Acoustical Society of America, 103, 2185–2196.

Lee, C., Han, C., Chuang, C. 2008. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. IEEE Transactions on Audio, Speech, and Language Processing, 16, 1541–1550.

Mathew, A., Amudha, P., Sivakumari, S. 2021. Deep learning techniques: An overview. Proceeding of International Conference on Advanced Machine Learning Technologies and Applications. Singapore. 1141.

Nair, V., Hinton, G.E. 2010. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10). Haifa, Israel.

Ocampo-Peñuela, N., Winton, S. 2017. Economic and conservation potential of bird-watching tourism in postconflict Colombia. Tropical Conservation Science, 10, 1–6.

Quintana, V. 2017. El turismo de naturaleza: Un producto turístico sostenible. Arbor Ciencia, Pensamiento y Cultura, 193, 1–14.

Simonyan, K., Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv, 1409, 1–14.

Steven, R., Morrison, C., Castley, J. 2015. Birdwatching and avitourism: A global review of research into its participant markets, distribution and impacts, highlighting future research priorities to inform sustainable avitourism management. Journal of Sustainable Tourism, 23, 1257–1276.

Steven, R., Rakotopare, N., Newsome, D. 2021. Avitourism tribes: As diverse as the birds they watch. Singapore: Springer.

Sukri, M., Fadlilah, U., Saon, S., Som, M., Sidek, A. 2020. Bird sound identification based on artificial neural network. Proceeding of 2020 IEEE Student Conference on Research and Development (SCOReD). 342–345.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. 2016. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2818–2826.

Tanttu, J., Jari, T., Ojanen, M. 2003. Automatic classification of flight calls of Crossbill species (Loxia spp.). Proceedings of the 1st International Conference on Acoustic Communication by Animals. 239–240.

Wang, H., Xu, Y., Yu, Y., Lin, Y., Ran, J. 2022. An efficient model for a vast number of bird species identification based on acoustic features. Animals, 12, 2434.

Xeno-canto Foundation. 2021. Compartiendo cantos de aves de todo el mundo. Retrieved 2022-06-06 from https://www.xeno-canto.org/about/xeno-canto

Zhang, X., Chen, A., Zhou, G., Zhang, Z., Huang, X., Qiang, X. 2019. Spectrogram-frame linear network and continuous frame sequence for bird sound classification. Ecological Informatics, 54, 101009.