**Neural Networks for Bird Species Classification Based on Sounds**

**ABSTRACT**

This report focuses on the use of neural networks to classify bird species based on their vocalizations. The dataset used in this study was obtained from the Birdcall competition, which includes preprocessed spectrograms of bird vocalizations from 12 different species. The report presents the findings of the use of both dense and convolutional neural networks for binary classification between two species ‘Blue Jay’ and ‘Barn Swallow’, as well as classification of all 12 species. The report highlights the challenges and important considerations in designing models for sound classification tasks. The network architectures and hyperparameters are carefully selected and compared to enhance performance. The results demonstrate the effective classification of bird species using the custom neural networks. The report discusses the challenges and considerations in designing models for sound classification tasks.

The binary classification model achieved a perfect score of 100% on the training set and a validation accuracy of 93.3%, indicating good generalization ability. The model also showed steady improvement in its performance during the training process and did not overfit the training data. The model developed for classification of 12 bird species achieved a satisfactory test set accuracy of 69.84% and demonstrated good generalization ability. However, the model showed some signs of overfitting to the training data, which may be addressed through further fine-tuning of the architecture and parameters. Overall, the study shows promising results for the use of neural networks in sound classification tasks.

**INTRODUCTION**

The Birdcall competition data used in this report is obtained from Xeno-Canto, a crowd-sourced bird sounds archive. The competition aims to develop effective ways of monitoring bird populations, assessing habitat quality, pollution levels, and the success of conservation efforts. The Center for Conservation Bioacoustics (CCB) at the Cornell Lab of Ornithology is leading the initiative to collect and interpret sounds in nature and develop innovative conservation technologies.

One of the main challenges of the competition is to analyze complex soundscape recordings that contain a variety of bird vocalizations and environmental sounds. These recordings are challenging to process as they often contain overlapping species calls, anthropogenic sounds like airplane overflights, and non-bird calls like chipmunks. The objective of the competition is to build accurate detectors and classifiers that can identify various bird vocalizations within the intricate soundscapes. ([Cornell](https://www.kaggle.com/c/birdsong-recognition))

The dataset used in this report consists of short recordings of individual birds, as well as soundscape recordings that are longer and include multiple species calls simultaneously. The dataset for this report is carefully selected and includes 10 high-quality sound clips, each under 60 seconds, for 12 different bird species commonly found in the Seattle area. These clips have been chosen to represent the highest quality available for each species.

The data has undergone preprocessing, which includes subsampling the sound clips to a sample rate of 22050 Hz and identifying the "loud" parts of the recordings. From these "loud" parts, 2-second windows with bird calls are selected for further analysis. Spectrograms are then generated for each 2-second window, resulting in a 343 (time) x 256 (frequency) "image" of the bird call. Finally, all bird calls for all clips in each species are saved individually, resulting in an uneven number of samples in each species.

**BACKGROUND**

Deep Learning and Neural Networks: An Overview

Deep learning is a subfield of machine learning that utilizes artificial neural networks with multiple layers to process and learn from data. Neural networks are designed to simulate the structure and function of biological neurons and are composed of interconnected computational units called neurons. These neurons can process complex computations on input data and are capable of learning.

NEURAL NETWORKS

Neural networks are machine learning models that draw inspiration from the structure and function of biological neurons. They are composed of interconnected layers of computational units called neurons, which are capable of learning and processing complex computations on input data.

There are two types of neurons in a neural network - linear and nonlinear, depending on the activation function used in their neurons. The activation function is a mathematical function that transforms the output of a neuron into a nonlinear function of its inputs.

A linear neuron computes the output by performing a linear function on its inputs. On the other hand, a nonlinear neuron applies a nonlinear function to its inputs. The output of a linear neuron is a weighted sum of its inputs, whereas the output of a nonlinear neuron is a nonlinear function of its inputs. The formula for a linear neuron is expressed as f(X) = A1X1 + A2X2 + ... + An\*Xn + Ak, where X1, X2, ..., Xn are input values, A1, A2, ..., An are the weights of the inputs, and Ak is the bias term. The formula for a nonlinear neuron using the sigmoid activation function is g(z) = 1 / (1 + e^-z), where z is the weighted sum of the inputs plus the bias term.

Single-Layer and Multi-Layer Neural Networks

A single-layer neural network consists of only one layer of neurons that process input data. This type of network is limited in its ability to represent complex patterns in the data and is mainly used for simple classification tasks. However, a multi-layer neural network, also known as a deep neural network, consists of multiple layers of neurons that process input data. These networks can represent more complex patterns in the data and are widely used in various tasks such as image recognition, natural language processing, and more. To illustrate the power of neural networks, the MNIST dataset is often used. It is a dataset of handwritten digits that range from 0 to 9. The task is to identify the correct digit from its image. By using deep neural networks, remarkable accuracy can be achieved, surpassing human performance in some cases.

Neural Network Architecture

The basic neural network architecture consists of three interconnected layers of artificial neurons: input layer, hidden layer, and output layer. The input layer receives information from the outside world and processes it. The hidden layers take their input from the input layer or other hidden layers and analyze the output from the previous layer, processing it further before passing it on to the next layer. Finally, the output layer provides the result of all the data processing by the neural network. The number of nodes in the output layer depends on the nature of the problem, with a binary classification problem having a single output node and a multi-class classification problem having multiple output nodes. ([amazon](https://aws.amazon.com/what-is/neural-network/#:~:text=A%20neural%20network%20is%20a,that%20resembles%20the%20human%20brain.))

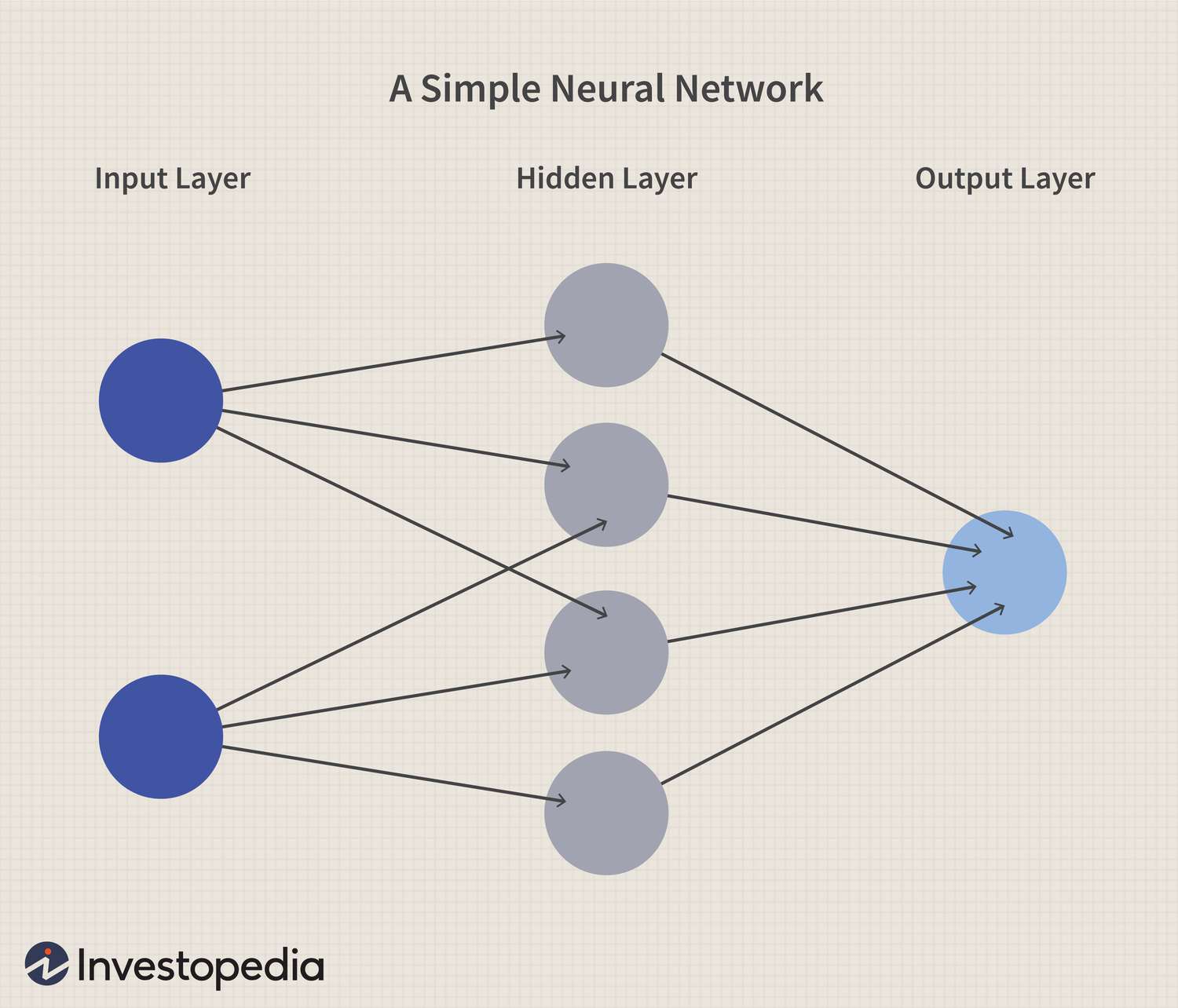


Fig 1: Neural Network Architecture - [investopedia](https://www.investopedia.com/terms/n/neuralnetwork.asp)

Neural Network Applications

Neural networks are widely used in various applications, including image classification, document/text classification, convolutional neural networks, and recurrent neural networks.

Image classification

Image classification is a common application of deep learning, where neural networks are used to classify images based on their content. Convolutional Neural Networks (CNNs) are widely used for this task due to their ability to recognize patterns in images. These networks are trained on large image datasets to accurately classify images. Data augmentation techniques such as rotation, scaling, and flipping can be used to generate more training data and improve the accuracy of the model.

Document/Text classification

Neural networks can be utilized for document or text classification, where the input text is transformed into a sparse matrix and passed through a neural network for classification. Recurrent Neural Networks (RNNs) are especially suitable for analyzing sequences of text data, where the order of the input matters. To enhance the performance of RNNs in text classification tasks, lagged variables and auto-regression techniques can be employed.

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of neural network that are ideal for processing sequential data, such as text, speech, or time-series data. The network is designed to maintain a "memory" of previous inputs, which enables it to consider the order and context of the input data. RNNs can predict future values in time-series data, and their performance can be improved by utilizing techniques such as lagged variables and auto-regression. The architecture of RNNs allows them to understand the order and context of the input data, making them useful for a wide range of applications, including natural language processing, speech recognition, and time-series data analysis.

Sequential models, a type of RNN, are ideal for processing data that has a temporal dimension. Sequential models process input data in a sequential manner, where the output of each step becomes the input of the next. This creates a feedback loop that enables the model to remember previous inputs, crucial for processing sequential data.

Long Short-Term Memory (LSTM) is an advanced type of RNN that overcomes issues associated with traditional RNNs, such as the vanishing gradient problem. LSTMs use a more complex architecture that incorporates a memory cell, which allows them to retain information for a more extended period. This makes LSTMs suitable for tasks such as speech recognition and natural language processing.

Activation functions play a crucial role in RNNs, including sequential models and LSTMs. Activation functions add non-linearity to the model, which helps to learn complex patterns. Common activation functions include the hyperbolic tangent and the rectified linear unit.

Time series forecasting is a prevalent use case for RNNs, where the objective is to predict future values in a time-series dataset. RNNs are well-suited for this task because they can recognize patterns in the data over time and understand its sequential nature. Techniques such as auto-regression and lagged variables can further improve the performance of RNNs in time-series forecasting tasks. [IBM](https://www.ibm.com/topics/recurrent-neural-networks)

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of deep learning neural network that has greatly enhanced the accuracy of image recognition and classification tasks. The architecture of a CNN is composed of several layers, each of which is engineered to extract specific features from an input image. Some of the most crucial layers in a CNN include convolutional layers, ReLU layers, pooling layers, and fully connected layers. ([towardsdatascience](https://towardsdatascience.com/wtf-is-image-classification-8e78a8235acb))

A picture containing diagram, text, screenshot, design

Description automatically generated

Fig 2: CNN Architecture - [dshahid](https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529)

The initial layer in a CNN is the convolutional layer. It applies a convolution operation to the input image to extract meaningful features, like edges and shapes. Convolutional layers use filters that can be adjusted during training to create a feature map that highlights relevant features of the input image. By continually learning and adjusting the filters through exposure to a dataset of labeled images, the CNN can extract essential features from new images.

Following the convolutional layers are the ReLU layers, which introduce non-linearity into the CNN to enable it to learn more complex patterns. ReLU layers employ an element-wise activation function that discards negative pixel values and retains positive ones. This rectification helps prevent the vanishing gradient problem, which occurs when gradients become too small to adjust the network's parameters in deep neural networks.

Pooling layers come next in the CNN architecture, and they are used to reduce the dimensionality of the feature maps generated by the convolutional layers. These layers cluster the neurons' outputs in feature maps and combine them into a single neuron in the next layer. This process reduces the number of parameters in the network and guards against overfitting.

Finally, the fully connected layers link every neuron in one layer to every neuron in the next layer, resulting in the network's output. These layers perform image classification by utilizing the learned features from the previous layers to categorize images into specific classes or labels.

NEURAL NETWORK TECHNIQUES

Fourier Transforms

Fourier transforms are mathematical operations that enable the conversion of signals from their original domain, such as time or space, to their frequency domain. These transformations are

widely used in neural networks for processing signals, like images and speech, because they facilitate the identification of significant patterns and features that are essential for the task at hand. ([Kartik Chaudhary](https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520))

Spectrograms

Spectrograms are visual representations that display the frequency content of a signal over time. They are generated by performing Fourier transforms on small time segments of the signal, resulting in a set of frequency spectra that are then arranged over time to form a spectrogram. Spectrograms are frequently used in signal processing tasks, such as speech recognition and music analysis. They offer an effective way to visualize the frequency components of a signal, making it easier to recognize crucial features.

One-Hot Encoding

One-hot encoding is a technique used to represent categorical data as binary vectors. Each category is assigned a unique index, and a vector of binary values is generated with a 1 in the position corresponding to the index of the category and 0s in the other positions. One-hot encoding is extensively used in neural networks, especially in natural language processing tasks, where words are represented as one-hot vectors. This technique enables the network to distinguish between different categories and learn the relationships between them more easily.

**METHODOLOGY**

DATA SELECTION AND PREPROCESSING

This report utilized data from the Birdcall competition, which is sourced from Xeno-Canto, a platform that crowdsources bird sounds. The original dataset was massive, containing many sound clips for various bird species, with a total size of around 36 GB. To make the dataset more manageable, we selected 12 species and handpicked the highest quality sub-60-second sound clips available for each species. This resulted in a more manageable dataset of 10 sound clips per species, with a total size of around 56 MB.

For preprocessing, we were provided with both the original sound clips and preprocessed spectrograms. However, it is suggested to work with preprocessed spectrograms as they are easier to start with. The preprocessing steps involve subsampling the original sound clips to half their sample rate, resulting in a sample rate of 22050 Hz. We then identified the "loud" parts of the sound clip, which are parts greater than 0.5 seconds in duration, and selected two-second windows of sound where bird calls were detected. For each of these selected windows, a spectrogram was generated, resulting in a 343 (time) x 256 (frequency) "image" of the bird call. Finally, we saved all bird calls for each species individually, which resulted in an uneven number of samples for each species.

BINARY CLASSIFICATION USING NEURAL NETWORK

This report demonstrates binary classification between two bird species, Blue Jay and Barn Swallow, using spectrogram images as input data. The preprocessed spectrograms for both species were loaded and combined into a single dataset X, with a binary label variable indicating the species for each spectrogram. The data was split into training and testing sets, with a 70/30 split.

Two different neural network models were defined for this task. The first model was a dense neural network with two hidden layers of 128 units each, a dropout layer to reduce overfitting, and an output layer with two units and a SoftMax activation function. The second model was a convolutional neural network with two convolutional layers, each followed by a max pooling layer, and two dense layers, with the second layer having two units and a sigmoid activation function. Both models were compiled using binary cross entropy as the loss function, and accuracy as the evaluation metric, with different optimizers being used for each model (Adam for the first and RMSprop for the second). Before training the models, the input data were reshaped to have four dimensions, with the fourth dimension representing the number of channels (in this case, 1).

CLASSIFICATION BETWEEN DIFFERENT BIRD SPECIES

This also demonstrates a neural network model capable of classifying 12 different bird species based on their spectrograms. The 12 species of birds included in this study are American Crow, Barn Swallow, Black-capped Chickadee, Blue Jay, Dark-eyed Junco, House Finch, Mallard, Northern Flicker, Red-winged Blackbird, Stellar's Jay, Western Meadowlark, and White-crowned Sparrow. Here, the spectrogram data for each bird species is loaded and combined into a single array. The data is then divided into training and testing sets using a 70:30 ratio. To prepare the data for the model, the label variable is converted to a matrix using one-hot encoding.

Next, the architecture of the model is defined using the Keras API in R. The model comprises of two convolutional layers, each followed by a max-pooling layer. The output of the second max-pooling layer is flattened and connected to two dense layers with a ReLU activation function and a SoftMax activation function, respectively. The model is compiled with the "categorical\_crossentropy" loss function, the "rmsprop" optimizer, and accuracy as the evaluation metric. Prior to training the model, the input data is reshaped to a 4-dimensional array with dimensions (number of samples, height of spectrogram, width of spectrogram, number of channels). Finally, the model is trained on the training set for a set number of epochs and evaluated on the test set.

**RESULTS**

Binary Classification Between Blue Jay and Barn Swallow Species

The results presented in Fig 3, showcase the success of a neural network model trained to classify spectrogram sounds of two bird species, Blue Jay and Barn Swallow. The model underwent training for 20 epochs with a batch size of 16, and the performance was evaluated on both the training and validation sets.

The analysis of the results reveals that the model has achieved remarkable accuracy on the training set, with a perfect score of 100%. This signifies that the model has effectively memorized the training data and is capable of accurately classifying Blue Jay and Barn Swallow based on their spectrogram sounds. Moreover, the validation accuracy of 93.3% indicates that the model is generalizing well to new data and is not overfitting.

A graph of loss and loss

Description automatically generated with low confidence

Fig 3: Binary Classification Between Blue Jay and Barn Swallow Species

The loss function analysis shows that the model has steadily decreased its training loss with each epoch, indicating that it is progressively learning the underlying patterns in the training data and improving its performance. However, the validation loss has fluctuated after an initial decrease, indicating that the model may be overfitting the training data. However, overall, the model effectively distinguished between Blue Jay and Barn Swallow and did not overfit the training data. Testing on diverse datasets can confirm its ability to classify the two bird species.

CLASSIFICATION BETWEEN TWELVE DIFFERENT BIRD SPECIES

The neural network model developed to classify 12 bird species based on their spectrograms underwent a 20-epoch training process with a batch size of 16, with each epoch taking around 5 to 6 seconds to complete. The resulting test set accuracy of 69.84% is satisfactory, given the complexity of the classification task. The model demonstrated good generalization ability, with a validation loss of 2.19 indicating that it can classify unseen data accurately.

The model's training history showed a steady increase in accuracy for the first 10 epochs, followed by a fluctuating accuracy plateau. The model achieved an impressive training set accuracy of 98.64%, suggesting that it may have overfitted to the training data. Future improvements to the model's performance and avoidance of overfitting may be possible through more fine-tuning of the model's architecture and parameters.

A picture containing text, line, diagram, plot

Description automatically generated

Fig 4: Classification between all twelve species

The bird species classification neural network model based on their spectrograms showed a promising performance that could benefit from future optimizations to further enhance its accuracy.

**DISCUSSION**

The results demonstrate the effectiveness of utilizing neural network models to classify bird species based on their spectrogram sounds. The first model, which focused on distinguishing between Blue Jay and Barn Swallow, achieved high accuracy on the training set, indicating its ability to effectively memorize the training data. Additionally, the model performed well on new data, with a validation accuracy of 93.3%, demonstrating its capacity to accurately classify unseen data.

The second model, which classified 12 different bird species, achieved a satisfactory test set accuracy of 69.84%, given the complexity of the task. The model demonstrated good generalization ability, with a validation loss of 2.19, indicating its accuracy in classifying new data. However, the high training set accuracy of 98.64% suggests that the model may have overfitted to the training data. Improving the model's architecture and parameters through fine-tuning could lead to better performance and avoid overfitting in future iterations During the training process, it was observed that certain bird species posed more challenges in terms of accurate prediction. For instance, species such as Mallard had distinctive and clear calls, which made them easier to identify, resulting in no misclassifications. On the other hand, species like Black-Capped Chickadee, Blue Jay, House Finch, and Northern Flicker had calls that were somewhat similar in sound. This similarity in calls made it more difficult to differentiate between these species, increasing the difficulty of accurate classification.

Other models, such as decision trees, support vector machines, and logistic regression, could have been used for this task. However, neural networks were suitable because they can handle complex relationships between features and have shown great success in sound classification tasks. Neural networks can also learn patterns in data and improve performance with more training data.

**CONCLUSION**

In summary, the study demonstrates the successful application of neural network models in classifying bird species based on their spectrogram sounds. The first model, designed for the binary classification between Blue Jay and Barn Swallow, achieved perfect accuracy on the training set and demonstrated good generalization ability with a validation accuracy of 93.3%. This indicates that the model effectively learned the distinctive characteristics of these species and can accurately classify them.

The second model, developed for the classification of twelve different bird species, achieved a satisfactory test set accuracy of 69.84%. Although the model showed good generalization ability and low validation loss, there are indications of overfitting due to its high training set accuracy of 98.64%. To address this issue, future improvements can be made by fine-tuning the model's architecture and parameters to enhance its performance and prevent overfitting.

Overall, the findings highlight the potential of neural network models in bird species classification, offering valuable insights for ecological research and conservation efforts. Further advancements can be pursued by optimizing model architecture, refining parameters, and exploring larger datasets and advanced techniques. These efforts can lead to improved classification accuracy and expand the practical applications of bird species identification in various domains.

**CITATIONS**

1. <https://aws.amazon.com/what-is/neural-network/#:~:text=A%20neural%20network%20is%20a,that%20resembles%20the%20human%20brain>.
2. <https://www.investopedia.com/terms/n/neuralnetwork.asp>
3. <https://towardsdatascience.com/wtf-is-image-classification-8e78a8235acb>
4. <https://www.ibm.com/topics/recurrent-neural-networks>
5. <https://www.kaggle.com/datasets/rohanrao/xeno-canto-bird-recordings-extended-a-m>
6. <https://xeno-canto.org/>