Machine Learning for Conservation: Evaluating Deep Learning and Feature Extraction in Bird Species Classification in New Zealand

1st Mahsa Mohaghegh
Auckland University of Technology (AUT)
Auckland, New Zealand
0000-0003-2228-8300

2nd Khaula Alizai
Auckland University of Technology (AUT)
Auckland, New Zealand
khaulaalizai786@gmail.com

3rd Minh Hoang
Auckland University of Technology (AUT)
Auckland, New Zealand
tumihoang2002@gmail.com

4th Kapil Patel
Auckland University of Technology (AUT)
Auckland, New Zealand
kappatel408@gmail.com

5th June Lee
Auckland University of Technology (AUT)
Auckland, New Zealand
ssr2801@autuni.ac.nz

***Abstract*—Automated classification of bird sounds plays an important role in monitoring and protecting biodiversity. Recently, similar efforts have been carried out for birds from all over the world, but New Zealand is one that has been overlooked. Hence in this study, we will be comparing feature extraction methods and machine learning models using a dataset that primarily contains bird species from New Zealand. Machine learning models such as Gated Recurrent Unit (GRU), Long Short-term Memory (LSTM) Recurrent Neural Network, Artificial Neural Network (ANN), and Convolution Neural Network (CNN) were used for audio classification. The accuracies achieved from the training of these models resulted in GRU-MFCC with 0.78, LSTM-MFCC with 0.91, ANN-MFCC with 0.091, and CNN-MFCC with 0.997 accuracy respectively. In order to design a user interface that can anticipate bird sounds and identify them appropriately, we employed our highest-performing model, CNN, with MFCC acting as the extractor.**

***Index Terms*—audio classification, feature extraction, accuracy, models, machine learning, user interface, predict**

I. Introduction

Birds play an important role in maintaining a balanced ecosystem by removing pests and acting as pollinators. New Zealand, known for its unique avian biodiversity, could particularly benefit from such innovative conservation efforts. With the decreasing number of species of birds [3], it is vital to monitor the avian population in New Zealand to prevent them from becoming extinct such as birds from the flightless family as they can be extinct due to lurking predators around. While manual methods can be used to keep an eye on bird numbers, this may be time-consuming and expensive. However, automated methods combining acoustic sensors and automated bird categorization algorithms may be utilized to quickly and efficiently assess a species' conservation status with minimal human input and expertise.

Birds utilize sound for a number of functions, including creating territories for male birds, attracting a partner for mating, responding to their surroundings, and determining whether or not there is a threat. Previously, ornithologists, who are experts in bird sound, would identify bird cries and record and report on the birds found in different locations. However, it is becoming more and more difficult to find ornithologists for the manual classification of birds. Another issue is that with every passing generation of ornithologists, the amount of knowledge passed to the next generation is waning resulting in loss of valuable expertise [4]. Therefore, it is vital to be able to store and train artificial intelligence to be able to recognize the different bird species and help maintain the species conservation status. As a result, manual classification of birds is becoming increasingly undesirable, and there is a trend towards automation rather than manual classification. In order to maintain our ecology and prevent a drop in the number of birds and bird species, Machine Learning and Deep Learning Models are trained with bird sound data to apply the classification of bird calls.

In this study, data is gathered, segmented, and cleaned utilizing Audacity software in order to train and evaluate using four distinct deep learning models. Before the features are retrieved for training and testing, augmentation and oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) are used to balance and supplement our dataset to obtain additional sample points in order to prevent overfitting and under-fitting obstacles.

II. Literature Survey

Audio classification has been a popular topic for studies ever since machine learning became popular and flourished around the 1940s [21]. The majority of these attempts found the most success by applying a type of machine learning model called CNN.

To increase the accuracy of audio categorization, researchers in the fields of machine learning and artificial intelligence have applied a variety of deep learning-based studies. When multiple machine learning models are used to do this task, it has been shown to be successful. This section presents an overview of the most important findings from earlier research, highlighting crucial insights on the subject.

To automatically identify bird species from video recordings, researchers at [7] constructed two distinct pre-trained neural network models, Res-Net50V2 and EfficientNetB0, using image and audio processing and classification techniques. They utilized a dataset with 13,700 recordings from 137 different bird species that were compiled from various sources. EfficientNetB0 exhibited an accuracy of 92.4%, which is less than model ResNet50V2's accuracy of 97.1% on the test set, although both models demonstrated excellent accuracy. ResNet50V2 also shows faster and more efficient training. In the end, the two models were integrated and had a 90% accuracy rate. This research automated the categorization of bird species, ending the arduous task for ornithologists and subject matter specialists.

Another work by Fan, Ying, and Yue [10] focuses on strengthening bird protection and generalizing the majority of current bird sound identification algorithms. The study used a sizable dataset with 264 species that was built from several Kaggle tournaments and tagged appropriately. With MobileNetV3, a form of CNN model, as the foundation, a lightweight feature extraction model was built in this study to classify birds. Other models, including MobileNetV1, MobileNetV2, and ResNet50, have also been used for comparison. 95.12% accuracy in the test set and 100% accuracy in the training set were attained by the selected model, which is outstanding. All other lightweight models fared better than this one, with the exception of ResNet50, whose accuracy was 2.25 percent greater. The proposed solution demonstrated in this study improved model classification accuracy with fewer parameters and calculations.

Arunodhayan and Danny [11] participated in the BirdCLEF 2022 challenge in 2022, which aimed to identify rare and endangered bird species using soundscape recordings of Hawaii. The contest ran from February through May. 14,853 brief audio recordings of 152 different bird species were uploaded by Xeno-Canto users and included to the training dataset, together with metadata including pertinent details about the recording sites, the type of bird chirp, and other information. 5,500 recordings totalling 1 minute each made up the test dataset, which was used to evaluate 21 endangered bird species. The preprocessing and augmentation of data was the study's suggested strategy before training the models. Mel-spectrograms were created from raw audio recordings and any audio files that simply included background noise were deleted. Eight additional augmentation approaches, including Gaussian noise, Pink noise, Tanh distortion, and Denoise transform, were also used to increase the model's resilience. Gain, Normalisation of Loudness, Mixing, and Vertical Roll. The SED-based models DenseNet121, ResNet50, and EfficientNetB0 were employed. With an accuracy of 79%, the studies showed that EfficientNetB0 beat the other models. This research shows that it is possible to monitor climate change, endangered bird species, and overall quality of life.
In the next section, we will describe the methodology used in detail.

III. Methodology

In this study, audio classification of bird sounds is done with two feature engineering techniques. The first feature extraction technique used was the Mel spectrograms extraction and the second feature extraction technique used was the Mel Frequency Cepstral Coefficients (MFCC) feature engineering. Before the use of feature engineering techniques data collection, data segmenting, and data augmentation techniques were used. Mel spectrograms and MFCC features were extracted as input for the CNN, LSTM, ANN, and GRU machine learning models in order to get different accuracies and make a comparison between them.

A. Data Pre-processing

Audio or sound is perceived through the traverse wave of the vibration pressure of air. This analog signal information is sampled at intervals to allow digital workarounds for computer-based inputs. This digital information as an audio signal is distributed numerically over a set time. The quality or resolution of an audio signal is captured by the sample rate, which is the capture of the wave in a sample in time. To achieve accurate representation and to limit aliasing, the sampling rate must be double the frequency of an audio; known as the Nyquist Shannon sampling theorem [5]. However, this digital representation in waveform is not the best for training a model, as the capture of patterns is flawed and not ideal. To address this, spectral representation of audio to an image-like data is a popular choice for deep learning using spectrogram, Mel-Spectrogram, and Mel Frequency Cepstral Coefficient.

The spectrogram is the product when each of the time frames or windows undergoes Short Time Fourier Transformation (STFT). This mathematical computation converts the audio signal into a spectrum representation against time. This allows the visualization of the amplitude of an audio signal over time by the color intensity [5]. Since spectrograms are linearly represented by frequencies at a specific time, small variations in amplitude across time can be challenging for accurate analysis.

Mel-Spectrogram is a transformation applied to the spectrogram tailored to better align with human auditory perception. This transformation maps the original frequencies to a perceptually relevant scale, that essentially reduces frequency resolution of less significant areas of the spectrogram, whilst allowing faster computational on emphasising the important audio information in the

spectrogram. Additionally, this transformation is logarithmic, which compresses amplitude values at a lower frequency range while retaining high-intensity components. These additions for feature extraction are valuable for bird species identification for audio analysis. Mel Frequency Cepstral Coefficient is another further transformation on the Mel-Spectrogram by discrete cosine transformation (DST). The audio was sampled at 22050 Hz using a Python library Librosa, a popular audio tool for audio processing in Python. For the data Pre-Processing 3 steps were used in order to get a large dataset, balance the dataset, and also to clean and segment the audio dataset and they are below:

1. Dataset Segmenting and cleaning
   Our dataset is constructed from a trusted source and from three sites that provided audio from avian species native to New Zealand. It consists of more than 4700 audio files across 44 native New Zealand species. The audio had varying audio lengths from 10 seconds to 5 minutes long. A lot of the audio has background noise with such vehicles and other bird noises, which is accompanied by varying calls on the different bird species. Furthermore, these audios were formatted as mp3, where the convergence to Wav was done by the audio software Audacity to expand the data quality without loss. Additionally, manual audio cleaning was done to remove sections of blank and unwanted audio overlapping the raw audio, manual silencing was done on sections that were interfering with the bird call as shown in Figure 3.

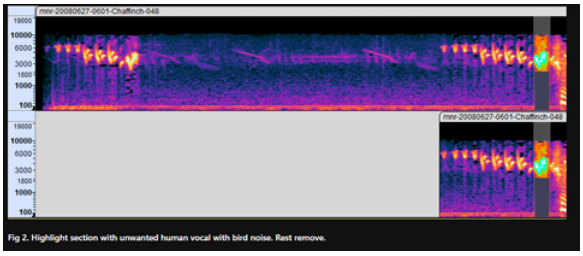

Fig. 1: Cleaning the audio files.

   To further clean this, a Python library to reduce noise was used to apply non-stationary noise gating on the audios to change threshold requirement overtime on unwanted audios with statistical calculations [8]. However, the audio clips still contained background and unwanted noises alongside the bird calls. The further tune was done by applying a high-order Butterworth filter pass to threshold unwanted noises. The Wav files were then segmented using Audacity into three-second intervals that contained the bird calls to field higher audio counts, capture relevant information for training, as well as lower computational load for the models.

2. Synthetic Minority Oversampling Technique (SMOTE)
   SMOTE is a synthetic sampling technique that uses oversampling to balance minority classes. Because our dataset comprises imbalance classes, this strategy addresses bias and overfitting. To generate synthetic data, the strategies employ the K nearest neighbor algorithm. The SMOTE does not consider the data locations of the surrounding majority class when providing synthetic data for the minority class. As a result, the classes may overlap or generate noise, rendering this method ineffective for high-dimensional data classification. Since Mel-Spectrogram and MFCC are the features we will use, the high-dimensional aspect will not be a problem.

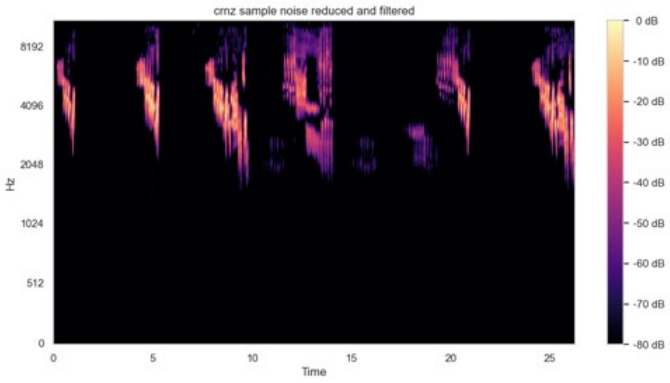

Fig. 2: filtered and noise reduced Mel-spectrograms.

3. Data Augmentation
   The risk of overfitting must be minimized because there aren't many sound samples available. The process of data augmentation involves adding more data. There are several common methods for enhancing audio data, including time shifting, noise addition, time stretching, and pitch scaling. Three data augmentation approaches, Time Stretching and Pitch Scaling, and Speed changing have been used in this research project and are discussed below.

   a) Time Stretching
      Time stretching is a method for changing an audio stream's pace or duration without changing its pitch or other characteristics. The Python audio editing package Librosa makes it simple to stretch out the passage of time. The pace and length of the audio can be changed by using various rate values. The Comparison of audio before and after time stretching can be seen in Figure 5.
   b) Pitch Scaling
      in terms of effect, pitch scaling is the opposite of time stretching. Here, the duration of the signal is kept constant while variable rate values are applied to alter the pitch of an audio transmission, producing a new sound as shown in Figure 6.

   c) Speed changing
      By increasing the speed of the audio clips, we can get more audio samples without changing their duration. This offers us more data points to analyze and determine whether the model can still function well when the speed is increased.

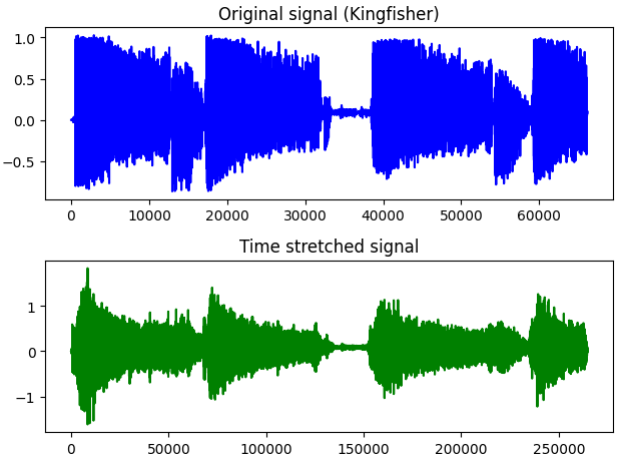

Fig. 3: Comparison of Time stretched audio for Kingfisher bird.

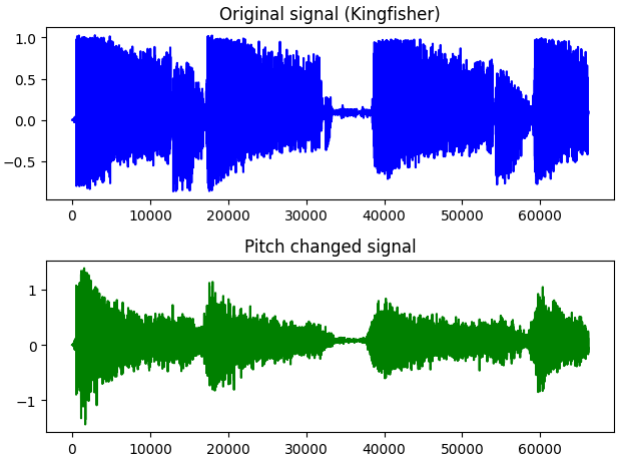

Fig. 4: Comparison of pitch scaling audio for Kingfisher bird.

B. Models Selected

Computer programs called machine learning models are used to find patterns in data or make predictions via learning. They are made using algorithms used in machine learning that are trained on either labeled, unlabelled, or mixed data. Applications for machine learning models include audio categorization, predictive analytics, picture identification, and many more. Although there are many different kinds of machine learning models, they may be divided into three major groups: reinforced learning, unsupervised learning, and multi-agent learning. With the use of labeled training data, supervised learning models are taught to generate predictions or categorical determinations. Models trained using unlabelled data, known as unsupervised learning, find patterns, clusters, or structures within the data without explicit direction. Models that use reinforcement learning acquire knowledge by experimenting and interacting with their surroundings.

An artificial neural network called a convolutional neural network (CNN) is made to do classification tasks for pictures and sounds. They take their cues from the way the human visual system operates, in which information is processed hierarchically by many layers of neurons. CNNs typically consist of 4 main types of layers: Conv2d with, BatchNormalization, Activation, and MaxPooling2D. Convolutional layers detect patterns from the input by applying filters. The pooling layers reduce spatial dimensions of the data by downsampling, which helps retain important information while reducing computational complexity. Fully Connected layers make the final decisions, such as classifying an image, by processing features extracted from previous features.

A Long Short-Term Memory (LSTM) is a species of recurrent neural network (RNN) designed to handle sequential data, thus being particularly beneficial for applications like audio classification. This model has the ability to recall long-term dependencies in data which can benefit us by analyzing the time-dependent features in the audio clips. The LSTM model has the following layers: LSTM, lSTM, Dense, Dropout, and Dense. The LSTM layer is the layer with the most significance as it memorizes the neurons and updates the model through the iterations. This ensures it's taking both short-term and long-term dependencies. It is very beneficial for audio as it encapsulates the sequential features which can help model temporal dependencies for audio classification analysis.

A Gated Recurrent Unit (GRU) is another variant of recurrent neural network (RNN) that is designed to handle sequential data, allowing it to be beneficial for audio classification tasks. GRUs implement gating mechanisms to manage the quality of information flow. This enables the model to gather temporal dependencies in the data along with being more efficient in computation. This has a slightly simplified architecture of conventional RNNs. The model has layers GRU, GRU, and Dense. The GRU layer commands how the data is adapted and proceeds through the model. This facilitates the model to gather the dependencies between distinct time frames. This model is favorable as it provides an equilibrium for efficiency and sophistication. Another advantage is they are less inclined to overfitting than standard RNNs which helps keep purposeful knowledge while disposing of unwanted details. Being swift in computations makes it a great candidate for running instantaneous classifications on audio.

An Artificial Neural Network (ANN) contains inner linked layers of artificial neurons. This is a common feed-forward network that is utilized for machine learning models. ANNs possess hidden layers that are accountable for learning unique features from mel spectrograms and Mel-frequency cepstral coefficients. ANNs abandon the temporal dependencies and treat the audio as 1D arrays. They can predict classes with less data than other models and generalize relatively well. This is favorable as it has the potential to adapt to new conditions and can work exceptionally well with inadequate data.

C. Software and Programming language used

Python was chosen as the project's preferred language, and packages like pandas, TensorFlow, and sci-kit-learn were

utilized to develop the system.

Google Colab was used extensively for the core development and training process since it offered 32 GB of RAM, free access to GPU and TPU, and made it simple for users to communicate and cooperate on their work. The model was trained on the GeForce RTX 2040 Super GPU from NVIDIA.

IV. Results and Evaluation

Before we could train the model we needed to extract features to allow our model to learn the various patterns and attributes for each individual audio file. One feature we applied was the Mel spectrogram, we used nmels = 128. For each file, it is separated into 128 evenly spaced frequencies which are distanced as it is heard by the human ear. In previous related work, Mel-frequency cepstral coefficients were also applied to audio classification to extract features. This computes the Discrete Fourier Transform for the audio file. The features were processed to accompany the model's input with a 3-dimensional input shape for the CNN model and RNNs architecture models. The models were trained and results are provided in Fig.6 below. By observing the results, we indicate that CNN with MFCC is performing the best by attaining the topmost value for validation accuracy and acquiring low loss from our models.

Models	Extractor	Loss	Accuracy	Val Loss	Val Accuracy
CNN	Mel Spec	0.124	0.962	7.667	0.351
	MFCC	3.797	1.000	0.010	0.997
LSTM	Mel spec	0.369	0.885	0.507	0.869
	MFCC	0.306	0.909	0.592	0.911
GRU	Mel Spec	0.207	0.950	0.706	0.778
	MFCC	6.283	1.000	0.003	1.000
ANN	Mel Spec	1.931	0.444	1.472	0.579
	MFCC	3.556	0.078	3.607	0.091

From the initial modest bird sound dataset, applying audio preprocessing techniques has allowed our model to learn the features and give valuable predictions on bird sounds. The Confusion Matrix of CNN can be seen in Fig.7 below.

The MFFC extractor has exceeded the Mel spectrogram extractor as the model generates higher accuracy with MFCC. Evaluating the accuracies of MFCC, we capture the following accuracies 100%, 91%, 100%, 09% effectively suggesting that CNN-MFCC is the model to operate for model inference. The trained CNN model will predict on unseen data to manufacture a result.

Furthermore, Gradio was utilized to incorporate our model into the Machine learning web application. This allows a simple user interface to engage with the classifier. The user is able to input an audio file which is ultimately processed by the model to compute which bird is in the audio. This displays what the model perceives of the audio from our classes. Fig.8 below shows our interface in action.

V. Conclusion and Future Work

The goal of this study was to find a way to correctly identify New Zealand bird species according to the sound

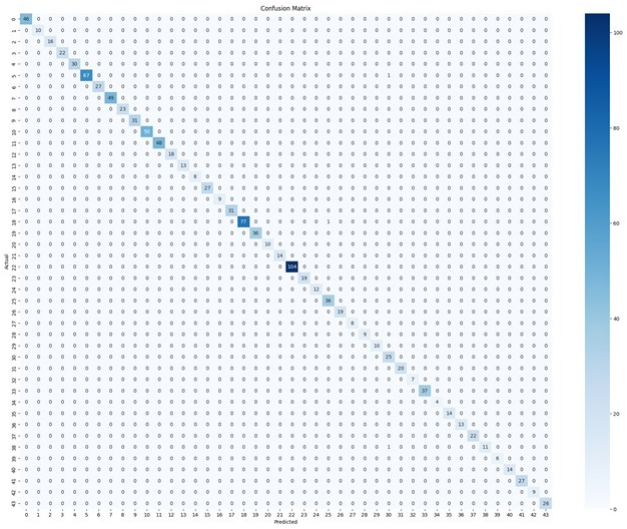

Fig. 5: CNN Confusion Matrix.

Fig. 6: Interface in action.

they make through Machine Learning methods. Before training models and making predictions, audio preprocessing is an important step to make the data suitable for machine learning. We handled this step by re-sampling our dataset, ensuring all audio samples have the same sample rate, and feature extraction, in which we extract relevant features like MFCCs and spectrograms. The four models we selected for this study are CNN, LSTM, GRU, and ANN. CNN outperformed other models with an accuracy of 99.7% . We assume our dataset was large enough with 101,552 audio files across 44 species after data augmentation and applying SMOTE technique. Machine learning algorithms learn better when provided with large datasets. Additional data on relevant New Zealand bird species will be obtained in the future edition of this work to strengthen the model's robustness during training. Other data augmentation and preprocessing techniques may also be investigated. Finally, more models may be applied in the future for better comparison.

Acknowledgment

We would like to express our gratitude to the Palmerston North City Council (PNCC) staff for their assistance during our study.

References

[1] Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev. 2019 Feb;11(1):111-118. doi: 10.1007/s12551-018-0449-9. Epub 2018 Sep 4. PMID: 30182201; PMCID: PMC6381354.

[2] E. J. Henri and Z. Mungloo–Dilmohamud, "A Deep Transfer Learning Model for the Identification of Bird Songs: A Case Study for Mauritius," 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Mauritius, Mauritius, 2021, pp. 01-06, doi: 10.1109/ICECCME52200.2021.9590917.

[3] Xie, J., and Zhu, M. (2023). Acoustic Classification of Bird Species Using an Early Fusion of Deep Features. Birds, 4(1), 138–147. MDPI AG. Retrieved from http://dx.doi.org/10.3390/birds4010011.

[4] Randler C, Heil F. Determinants of Bird Species Literacy-Activity/Interest and Specialization Are More Important Than Socio-Demographic Variables. Animals (Basel). 2021 May 28;11(6):1595. doi: 10.3390/ani11061595. PMID: 34071521; PMCID: PMC8229662.

[5] Wright, G. (31 May 2022). What is the Nyquist theorem?.WhatIs.com.https://www.techtarget.com/whatis/definition/Nyquist-Theorem.

[6] Coursera. (2022, May 19). Machine learning models: What they are and how to build them. https://www.coursera.org/articles/machine-learning-models

[7] N. Sharma, A. Vijayeendra, V. Gopakumar, P. Patni and A. Bhat, "Automatic Identification of Bird Species using Audio/Video Processing," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-6, doi: 10.1109/ICONAT53423.2022.9725906

[8] Sainburg, T. (2019). Timsainb/noisereduce: Noise reduction in Python using spectral gating (speech, bioacoustics, audio, time-domain signals). GitHub. https://github.com/timsainb/noisereduce

[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., amp; Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[10] FAN YANG, YING JIANG, AND YUE XU. Design of Bird Sound Recognition Model Based on Lightweight (2022).

[11] Arunodhayan Sampathkumar , Danny Kowerko. TUC Media Computing at BirdCLEF 2022: Strategies in identifying bird sounds in a complex acoustic environments (2022)

[12] Stefan Kahl, Mary Clapp, W Alexander Hopping, Hervé Goëau, Hervé Glotin, et al.. Overview of BirdCLEF 2020: Bird Sound Recognition in Complex Acoustic Environments. CLEF 2020 - Conference and Labs of the Evaluation Forum, Sep 2020, Thessaloniki, Greece

[13] Effendy, Nazrul Ruhyadi, Didi Pratama, Rizky Rabba, Dana Aulia, Ananda Atmadja, Anugrah Yuwan. (2022). Forest quality assessment based on bird sound recognition using convolutional neural networks. International Journal of Electrical and Computer Engineering. 12. 4235-4242. 10.11591/ijece.v12i4.pp4235-4242

[14] Dai, Y., Yang, J., Dong, Y., Zou, H., Hu, M. and Wang, B. (2021), Blind source separation-based IVA-Xception model for bird sound recognition in complex acoustic environments. Electron. Lett., 57: 454-456. https://doi.org/10.1049/ell2.12160

[15] Quan Tang, Liming Xu, Bochuan Zheng, Chunlin He, Transound: Hyper-head attention transformer for birds sound recognition, Ecological Informatics, Volume 75,2023,102001,ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2023.102001

[16] 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(18), 2434. MDPI AG. Retrieved from http://dx.doi.org/10.3390/ani12182434

[17] YO-PING HUANG, (Fellow, IEEE), AND HAOBIJAM BASANTA. Recognition of Endemic Bird Species Using Deep Learning Models (2021)

[18] N. A. and R. Rajan, "Deep Learning-based Automatic Bird Species Identification from Isolated Recordings," 2021 8th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2021, pp. 252-256, doi: 10.1109/ICSCC51209.2021.9528234

[19] Y. Jadhav, V. Patil and D. Parasar, "Machine Learning Approach to Classify Birds on the Basis of Their Sound," 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 69-73, doi: 10.1109/ICICT48043.2020.9112506

[20] Mehyadin, Aska Mohsin Abdulazeez, Adnan Hasan, Dathar Saeed, Jwan. (2021). Birds Sound Classification Based on Machine Learning Algorithms. Asian Journal of Research in Computer Science. 1-11. 10.9734/AJRCOS/2021/v9i430227

[21] Keith D. Foote. A Brief History of Machine Learning (2021). https://www.dataversity.net/a-brief-history-of-machine-learning/