Identification of birds species based on bones using machine learning approach

line 1: Muhammad Rizwan Tahir (1023340)  
line 2: *Liverpool John moores university*

line 3: Liverpool, England  
line 4: rizwantahir730@gmail.com

*Abstract*—In recent years, researchers have been working together to deal with the extinction of animal species. The conservation of wildlife is important as it provides stability to the natural mechanism. There are many machine and deep learning algorithms that have been developed so far to identify and classify animal species. This study investigates a task assigned by the Zoological Society of London on the use of machine learning as a potential method for classifying and identifying bird species based on their ecological group and bone characteristics. In this research, exploratory data analysis and data pre-processing of bird’s dataset has been done using different techniques. This research also proposes two classifiers, Random Forest and SVM (Support Vector Machine), for the identification of bird species based on bone types. In further work, both classifiers will be implemented using the Python scikit-learn package, and the outcomes will be compared to determine which model is the best.

Keywords—Artificial Intelligence, Machine Learning, Deep Learning, wildlife, Zoological society of London, DNA barcode, Random Forest, SVM

# Introduction

The conservation of wildlife is essential for maintaining environmental balance. It also provides stability to several natural mechanisms in nature. Animals play a crucial role in maintaining the ecosystem's balance. Moreover, animals make forests more beautiful, and humans are drawn to them because of their scenic beauty. However, unfortunately many species are in danger of being extinct in the current year [1] According to [2], worldwide, extinction of species is being caused by climate change and ecological degradation. To monitor the loss of biodiversity on Earth, automatic wildlife detection is urgently needed. [3] stated that for the protection, continuous observation and up-to-date knowledge of the presence, location, and behavioural changes in an endangered species' environment are necessary.

Technology can significantly reduce the time and effort required for tasks related to animal and ecosystem conservation [4]. Both machine learning and deep learning approaches have been used for the identification of animal species. There are several methods for monitoring wild animals, such as broadcasting, wireless sensors tracking, camera traps, and wildlife voice pattern recognition [2]. According to [5], few of the characteristics that have been used to identify animal species are bones, feathers or hair, teeth, eggs, shells, claws or paws, and ivory. However, depending on the type of problem we are attempting to address, each strategy has benefits and drawbacks.

This research topic is primarily focused on the identification of bird’s species. In prior studies, identification of bird’s species has been done by different approaches, using DNA barcode sequence [6], audio signal processing [7] , and by using protein radioimmunoassay (pRIA) for bone fragments [8]. Although, we can identify species by using these techniques but there are some limitations of these approaches. Outlining limitation of each approach, for DNA barcode sequence [6]; hybridization of different species is mainly the problem, for audio signal processing [7]; background noise or weather noise is the issue and for pRIA [8]; identifying species having poorly reserved bones is the biggest limitation. Consequently, utilising various ways to identify different bird species has certain drawbacks. However, deep learning techniques have recently advanced, and the results for identifying objects and animals in images have been excellent [3]. To achieve the highest accuracy in comparison to different approaches, faster R-CNN in combination with the ResNet152 feature map and the SSD (Single Shot MultiBox Detector) is the best option for images dataset [9]. Although we have a dataset of bird’s species in numerical form (1D), therefore this research proposes Random Forest and SVM (Support Vector Machine) for the implementation part to classify bird’s species. These classifiers will be used to train the model and accuracy of both will also be compared.

The remaining sections will be organised as follows: In Section 2, a background explanation with the importance of bird’s species, prior studies to identify bird species and challenges outlined by Zoological society of London has been covered. The findings of experiments and discussions based on the dataset of bird bones are presented in Section 3. Finally, future work will be discussed in Section 4.

# background

The Zoological Society of London (ZSL) is essential in determining how well the world's wildlife is doing, from keeping track of threatened animals to mapping wildlife populations on a large scale. According to the ZSL, threats to wildlife include the fatal chytrid fungus, illegal trafficking, habitat loss, and climate change. A significant number of animal species may become extinct during the next century [10]. According to Dr Andrew Terry, some of the options available to wildlife travelling across continental Europe are not present in the UK. However, strong environmental safeguards, financial support, and a kickstart to establish the necessary conditions might help animals recover across our nation [11].

All species inside an environment contributes significantly to its maintenance. Birds, along with other native species, are essential to maintaining an ecosystem's balance. They are regarded as species that support stable population levels among all living creatures in the larger web of life [12]. The entire ecology is put in danger by even a little decline in their population[13]. In the trophic hierarchy, birds can be found at many different levels. They are both top predators and mid-level consumers. They restore plants to ecosystems through pollination or seed dispersion. Numerous birds are scavengers that aid in the swift removal of carcasses and the nutrient recycling in the environment to keep a healthy habitat. However, variations in avian distribution are frequently the initial indicator of an environmental disturbance [14]. BirdLife International found in a survey that many bird populations are in decline globally. About 1500 bird species are on the verge of extinction, and it is difficult to estimate when they will do so[13]. Therefore, for conservation, it's essential to develop a quick and efficient method for identifying the components of a region's biodiversity and tracking changes over time [14].

In recent art-of-the-work [7] stated, classification of bird species is a challenging that scientists are familiar with, and it has long been thought of as a scientific study. Additionally, conventional categorization techniques such as morphology and DNA are well-known ways to distinguish between different bird species. In [6], the author conducted research to identify bird species based on traditional approach by using short DNA barcode sequences from a specified section of the genome. However, invasion of the genome is widespread, as shown by DNA sequence information and other molecular techniques which can results in hybridization of some species [15], which limits the barcode classification. According to the [8], identifying species from bones is crucial to forensic, anthropological, and archaeological investigation. In this research author used protein radioimmunoassay (pRIA) to identify animal species from bone and bloodstains. Moreover, [16] stated that, a bootstrap consensus tree is produced after phylogenetically analysing published proteins and short peptide sequences extracted from bones to identify birds. The author of [17] investigated the ancient collagen or protein chains COL1A1 and COL1A2, and they then provided a system of proteome profiling to categorise extinct bird species. According to the author, mass spectrometry analysis may extract collagen peptide sequences with a high degree of coverage from 46 000 years old chicken, dodo, and auk bones. Researchers were able to accurately categorise historical bones down to the family level using these sequences. However, the identification of a species might be challenging or impossible when the bones are fragmented or poorly protected [8].

According to earlier research, experts have been looking into several technologies for years to accomplish automatic identification of species.

# METHODOLOGY

This section of the paper is divided into six parts. In first part we will describe our dataset of birds, second part will be about the analysis of dataset by using some graphs, third part is data pre-processing, will perform certain actions after the analysis of dataset so it will be ready for modelling, in fourth part there will be a discussion of selection of model architecture, in fifth part model training will be discussed and finally in the last part inference and evaluation of selected model.

## Data Description

Birds come in various varieties, including pigeons, ducks, ostriches, and penguins. While some bird species can fly, others cannot. Land birds can run, some can swim underwater, and yet others can wade in shallow water. Birds are divided into various ecological groupings based on their habitats and dietary preferences. In this paper, we will be dealing with 6 ecological bird groups and each bird has been labelled by its ecological group.

* Swimming Birds (labelled as: SW)
* Wading Birds (labelled as: W)
* Terrestrial Birds (labelled as: T)
* Raptors (labelled as: R)
* Scansorial Birds (labelled as: P)
* Singing Birds (labelled as: SO)

The dataset we have contains 420 different bird species. For the representation of each bird there are 10 measurements which are listed below.

* Length and Diameter of Humerus
* Length and Diameter of Ulna
* Length and Diameter of Femur
* Length and Diameter of Tibiotarsus
* Length and Diameter of Tarsometatarsus

The type of all measurements are continuous float values (mm), and empty strings are used to represent any missing values. This dataset contains bones from the Natural History Museum of Los Angeles County. They are members of 153 genera, 21 orders, and 245 species.

## Exploratory Data Analysis

Many data scientists perform exploratory data analysis (EDA), which involves investigating and summarizing data sets, frequently using data visualization techniques. The purpose of performing EDA is to find obvious errors, better understanding of data patterns, detect outliers and can find the correlations between the variables. Different visualization techniques have been performed on the bird’s dataset to extract insights from the data. Scripting language, python has been used to check the insights of bird’s dataset. Let’s check the insights by importing the dataset in python.

**Overview of dataset**

Graphical user interface, application

Description automatically generated

**Fig. 1 – Overview of dataset**

Fig 1 shows the overview of bird’s dataset. As we can see, there are 12 variables and 420 observations in the dataset. Dataset has two types of variables numeric and categorical, from 12 variables, 11 are numeric and one of them is categorical. Fig 1 also shows some other statistics like the percentage of missing cells which are 0.3% in this dataset.

**Dataset information**

Text

Description automatically generated

**Fig. 2 –** **Dataset variables**

In Fig 2, there is a list of columns of the dataset, with their number of occurrences and data types. All the variables are numeric other than the type which is an object.

**Missing Values**

Chart, bar chart

Description automatically generated

**Fig. 3 –** **Missing Values**

In the above Fig 3, matrix shows the columns having missing values in our dataset. In this matrix, the vertical axis shows the total number of records and horizontal axis shows the number of columns in the dataset. The horizontal white line is the representation of missing values. The percentage of missing value is 0.3% in the dataset.

**Outliers**

**Graphical user interface, application, table, Excel

Description automatically generated**

**Fig. 4 – Outliers in the dataset**

The above table in Fig 4 shows the total number of occurrences, mean, standard deviation, minimum and maximum values in each column. From Fig 4, the yellow highlighted values in the table are outliers of the dataset. As we can see, this table just shows the 25%, 50%, 75%, minimum and maximum number of occurrences of the values but other hidden outliers can be seen by plotting the graphs of each column. To understand how to see outliers from graph for each column or variable just look at Fig 5.

Chart, histogram

Description automatically generated

Fig. 5 – To check outliers in each column

There are many ways to identify outliers in dataset, through mathematical functions and from visualization. Visualization can be done by plotting box plot, scatter plots and histogram. In the above Fig 5, histogram shows the outliers of “tibl” with fixed 50 size bins. Here, we can see that one value in the start and other values above 200 are outliers.

In the next part “Data pre-processing”, will discuss the different techniques to remove outliers and will also discuss the proposed one for this dataset.

**Class Imbalance**

**Table

Description automatically generated with low confidence**

**Fig. 6 – Imbalance classes**

Fig 6 represents the classes of “type” categorical variable in the dataset. As we can see, these classes are not balanced because the number of observations of each class is different from each other. Class “SO” contains 128 occurrences while class “T” just contains 23.

Class balance is significant because it facilitates model training by preventing the model from becoming biased towards one class.

**Data Skewness**

Skewness is a statistic for the asymmetry or distortion of symmetrical patterns in a dataset. Skewed data can reduce the model’s capacity to describe typical cases because it deals with rare occurrences on extreme values.

Text

Description automatically generated

**Fig. 7 – Data Skewness**

Fig 7 shows the asymmetry of the probability distribution of variables about their mean values in the dataset. Consequently, data normalization is needed so we can deal with data skewness.

**Correlation**

**Chart

Description automatically generated**

**Fig. 8 – Correlation between the variables**

Fig 8 shows the correlation between the variables. The correlation of ‘id’ variable with other variables is weak because we can see the negative association from Fig 8. Furthermore, the “id” column is an auto incremented value against each record which acts as identifier, but we can track each record through index. Moreover, other variables are independent and identically distributed so “id” can be eliminated or dropped out.

The correlation between other variables ranges from 0.61 – 0.98. As from the matrix we can see that the strongest correlation is between “ulnal” and “huml”, weakest is between “tarl” and “tarw”.

## Data pre-processing

Now we can clean our data from the raw form after comprehending and analysing the dataset throughout the EDA process.

**Column Dropping**

As we discussed from Fig 8 in the correlation part of EDA that “id” has negative association with other variables so we can drop it.

Application

Description automatically generated with medium confidence

**Fig. 9 – Dropping of id variable**

Fig 9 shows that “id” variable has been removed from the dataset.

**Handling Missing values**

Missing values introduce bias into the dataset, they can affect how well the model performs. For the construction of a model, the loss in values may hold important insights or data. Therefore, to address the missing values there are some possible ways such as list-wise deletion, simple imputation and regression imputation. For this dataset we use simple imputation to handle missing values.

Chart, box and whisker chart

Description automatically generated

**Fig. 10 – Mean of each class**

Fig 10 shows the mean value of each class belongs to categorical variable “type”. Simple imputation technique takes the mean value of each class.

Graphical user interface, text, application

Description automatically generated

**Fig. 11 – Imputation function and result after handing missing values**

Fig 11 represents the code of imputation function in the left and on the right-side results are shown after performing this function to each column. This piece of code returns the mean value of each class and replaces it with the missing cell wherever found in the column.

**Removing Outliers**

There are different techniques to handle outliers from the dataset, deleting observations, transforming values, imputation and separately treating. As we can see from fig 4 that we have outliers in the dataset. Since there aren't many outliers in this dataset, the strategy of deleting observations has been selected. By removing the outliers, correlation between variables also gets stronger.

Graphical user interface, text, application, email

Description automatically generated

**Fig. 12 -Removing Outliers**

In the above Fig 12, a piece of code is the function to remove outliers from the dataset. In this function, q1 is the first quartile of the dataset and q3 is the third quartile of the dataset. The quartile range has been set to remove the values which are not in the quartile range.

**Graphical user interface, application, table

Description automatically generated**

**Fig. 13 – Dataset description after removing outliers**

Fig 13 represents the description of the dataset after removing the outliers. As the number of observations has been removed due to outliers, the total number of observations has reduced from 420 to 381.

**Class Balancing**

As discussed before in Fig 6, there is imbalance of all the classes contained in dataset. Class imbalance results in models that have poor predictive performance.

Class balancing can be handled using resample the training set. There are two methods of re-sampling, under-sampling and over-sampling.

The dataset is balanced through under-sampling by minimizing the size of the abundant class. In contrast, oversampling is employed when there is insufficient data. As we have an insufficient number of observations, which is 420, oversampling is the best option.

**Data Skewing**

Data that produces an unequal distribution of curves on a graph is skewed data. When the statistical distribution's curve appears warped to the left or right, we may tell that the data are skewed.

Chart, histogram

Description automatically generated

**Fig. 14 – Normal Distribution of variable “feml”**

Fig 14 shows the normal distribution of one column of the dataset. To handle skewness or for normal distribution of the dataset logarithm function of “numpy” library has been applied to each column.

**Converting Categorical to Numerical**

In the dataset, there are 11 numeric datatypes and 1 categorical. Machine learning and deep learning algorithm requires input and output variables as a number. So, categorical values need to be converted to numeric ones.

One method is to get dummy values of the categorical column and then concatenate with the dataset and remove the existing categorical column.

Graphical user interface, application

Description automatically generated

**Fig. 15 – Converting categorical to numerical**

Fig 15 shows that the categorical variable “type” has been converted to numerical values.

## Model Architecture Selection and Justification

There are several model architectures of machine learning and deep learning that have been developed so far. Each model has a specific usage, and it depends upon the type of problem and dataset.

The dataset of bird’s bones lies on the category of supervised machine learning because the dataset is in labelled form. The predictive outcome of the dataset is in categorical form, so classification techniques are needed to develop the model. Random forest and SVM (Support vector machine) classifiers have been selected as both can deal with the classification problems. These models have the best predictive results among classification problems. Both models will be implemented in the future work and accuracy will be compared between them.

## Model Training

The dataset will be divided into training and test sets for model training. 80% of the data will be in the training set, and 20% will be in the test set. We will set hyperparameters for the method we'll be utilizing after dividing the data into different sets. Hyperparameters are used to regulate the model's learning process. As we selected Random Forest and SVM both for the modelling. Let’s discuss the model training of each model one by one.

**Random Forest Training**

The bagging approach is used to train Random Forests. Bagging, also known as bootstrapping, is the process of fitting a model to smaller subsets of the training data, aggregating the predictions, and randomly selecting sections of the training data.

The hypermeters that will be used for the training of this model are class\_weight, criterion, max\_depth and n\_estimators.

**SVM (Support Vector Machine) Training**

SVM classifies data by mapping the input to a high-dimensional feature space, which allows it to do so even when the data points cannot be separated linearly. The data are changed to allow the hyperplane representation of the separator once a separator between the categories has been found.

The hypermeters that will be used for the training of this model are C, kernel, degree and gamma.

## Inference and Evaluation

Once the training of the model has been completed, the next step is to inference and evaluate the trained model. The model then will be deployed to process external data, as a result, it will be capable of looking for and identifying whatever it has been trained to identify. The most important thing to evaluate your model is not to train it on the whole dataset. The data should be split into training and test sets. It’s better to divide dataset such as 70-80 % to training and 20-30% to test sets. Moreover, to avoid the possibility of models becoming overfit to the training set, it is crucial to incorporate fresh data while evaluating it. Additionally, the model shouldn't be underfitted, as this indicates a high level of bias and pays little attention to the external data.

There are several metrics that will be used for the evaluation of the model such as bias-variance trade off, confusion matrix, accuracy, precision, recall and mean squared error.

# Discussion

The preservation of animals is crucial for preserving the ecological equilibrium. Unfortunately, though, several species are in risk of being extinction in the current year [1]. In Artificial Intelligence, scientists are working assiduously together to find different ways to identify animal species and to protect them. There are many pre-build machine algorithms to solve different types of problems such as regression, classification, clustering etc. In this paper, the dataset of bird’s bones has been used for the classification of bird’s species. After conducting an exploratory data analysis, we came up with the decision to use Random Forest and SVM classifiers in future work for identification purposes because earlier research has demonstrated that these models provide the best predictive outcomes for classification problems. Random forest can handle large datasets efficiently and SVM can give high performance. We will compare the results of both classifiers to check the best one among them.

# Future work

In future work, there will be data pre-processing to balance the classes of the dataset using oversampling technique. The above was remaining in data pre-processing, haven’t done so far so it is needed to be done before modelling. Both Random Forest and SVM classifier will be implemented using scikit learn library of python with the best hyperparameters. Both the models than will be compared for the accuracy, keeping in mind, several metrics that can be used for the evaluation of these models.

##### References

[1] K. M. Iraqi, IEEE Computer Society, University of Karachi. Department of Computer Science, and Institute of Electrical and Electronics Engineers, *ICISCT’20 : 2nd International Conference on Information Science and Communication Technology : 8th-9th February 2020*.

[2] C. Duhart, G. Dublon, B. Mayton, G. Davenport, and J. A. Paradiso, ‘Deep Learning for Wildlife Conservation and Restoration Efforts’.

[3] S. B. Islam and D. Valles, ‘Identification of Wild Species in Texas from Camera-trap Images using Deep Neural Network for Conservation Monitoring’, in *2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020*, Jan. 2020, pp. 537–542. doi: 10.1109/CCWC47524.2020.9031190.

[4] P. C. Ravoor and S. T.s.b., ‘Deep Learning Methods for Multi-Species Animal Re-identification and Tracking a Survey’, *Computer Science Review*, vol. 38. Elsevier Ireland Ltd, Nov. 01, 2020. doi: 10.1016/j.cosrev.2020.100289.

[5] S. Gouda, R. G. Kerry, A. Das, and N. S. Chauhan, ‘Wildlife forensics: A boon for species identification and conservation implications’, *Forensic Science International*, vol. 317. Elsevier Ireland Ltd, Dec. 01, 2020. doi: 10.1016/j.forsciint.2020.110530.

[6] P. D. N. Hebert, M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis, ‘Identification of birds through DNA barcodes’, *PLoS Biol*, vol. 2, no. 10, Oct. 2004, doi: 10.1371/journal.pbio.0020312.

[7] M. T. Lopes, L. L. Gioppo, T. T. Higushi, C. A. A. Kaestner, C. N. Silla, and A. L. Koerich, ‘Automatic bird species identification for large number of species’, in *Proceedings - 2011 IEEE InternationalSymposium on Multimedia, ISM 2011*, 2011, pp. 117–122. doi: 10.1109/ISM.2011.27.

[8] J. M. Lowenstein, J. D. Reuther, D. G. Hood, G. Scheuenstuhl, S. C. Gerlach, and D. H. Ubelaker, ‘Identification of animal species by protein radioimmunoassay of bone fragments and bloodstained stone tools’, *Forensic Sci Int*, vol. 159, no. 2–3, pp. 182–188, Jun. 2006, doi: 10.1016/j.forsciint.2005.08.007.

[9] A. Mirugwe, J. Nyirenda, and E. Dufourq, ‘EPiC Series in Computing Automating Bird Detection Based on Webcam Captured Images using Deep Learning’, 2022. [Online]. Available: https://news.mongabay.com/2017

[10] Zoological Society of London, ‘Species Conservation \_ Zoological Society of London (ZSL)’, Accessed: Nov. 23, 2022. [Online]. Available: https://www.zsl.org/conservation/species

[11] Zoological Society of London, ‘The European Wildlife Comeback Report 2022\_ Wildlife can return if given chance to recover \_ Zoological Society of London (ZSL)’, 2022. https://www.zsl.org/science/news/the-european-wildlife-comeback-report-2022-wildlife-can-return-if-given-chance-to (accessed Nov. 23, 2022).

[12] T. Telenský, P. Klvaňa, M. Jelínek, J. Cepák, and J. Reif, ‘The influence of climate variability on demographic rates of avian Afro-palearctic migrants’, *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-74658-w.

[13] BirdLife International, ‘Birds - BirdLife International’, Accessed: Nov. 23, 2022. [Online]. Available: https://www.birdlife.org/birds/

[14] Y. P. Huang and H. Basanta, ‘Recognition of Endemic Bird Species Using Deep Learning Models’, *IEEE Access*, vol. 9, pp. 102975–102984, 2021, doi: 10.1109/ACCESS.2021.3098532.

[15] J. Mallet, ‘Hybridization as an invasion of the genome’, *Trends in Ecology and Evolution*, vol. 20, no. 5. Elsevier Ltd, pp. 229–237, 2005. doi: 10.1016/j.tree.2005.02.010.

[16] M. Y. Stoeckle and D. S. Thaler, ‘DNA barcoding works in practice but not in (neutral) theory’, *PLoS One*, vol. 9, no. 7, Jul. 2014, doi: 10.1371/journal.pone.0100755.