Detecting Bird species by analysing the bones using machine learning.

\*Note: Sub-titles are not captured in Xplore and should not be used

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 4th Given Name Surname  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 5th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—the problem presented by the Zoological Society of London requires the identification of the species of the deceased animal. This paper focuses on the part of identifying the bird species by analysing the bone dimensions. As we proceed with the findings we will review the similar researches carried out in this specific domain and identify that the main research gap at the moment is to reduce the cost of the system to be implemented as WWF is planning to implement this system globally. Furthermore, the methodology to create this system would be discussed and at this point of research we will mainly focus on the Exploratory Data Analysis(EDA) and the data- preprocessing. Moving on, the model architecture for training the model would be selected which is random forest. Then the justifications of the model selection are presented in the discussion section and finally the future work would be discussed which is followed by this paper.

Keywords—bones, birds, species, detection, machine learning

# Introduction

The primary objective of this research is to provide a conceptual framework for the creation of a system that can identify the species of a bird based on the measurements of a bone. The creation of this system will lead to the preservation of the environment by identifying the location where a bird died and the real habitats of the concerned bird species. This will be accomplished by determining the spot where a bird died. The authors of [1] indicate that since 1970, a significant number of animal species have been extinct as a direct result of hunting as well as the disruption of their natural environment brought about by the expansion of human constructions. Therefore, technological intervention is essential in order to determine the causes for the mortality of the animals that are happening in order to preserve the ecosystem and the biodiversity. However, there have been a lot of studies that have tried to categorise all of the animal species by using deep learning algorithms. However, these solutions have a very high cost due to the fact that a lot of computational power is needed to categorise the system, and the hardware requirements both for collecting data and training the model are quite pricey[2]. The authors of the study [3] were able to create a system that is capable of classifying birds based on their auditory signals by making use of deep learning techniques. In addition, they were able to assess the performance of the system by using a two-stage CNN-LSTM model. This method is remarkable because, in most cases, it is not essential for the bird to be recorded by the camera; hence, it enables the system to identify the birds that are present in that particular neighbourhood. Another piece of research [4], provides us with the methodology and a different approach to detect birds. This method involves using the data from camera traps to determine the species of animals. They used data from camera traps to determine the species of animals, and the model architecture that they used for their system was faster RCNN, which was finalised after comparing faster RCNN to YOLO.

The remaining portion of the study will concentrate almost entirely on the background, which will discuss all the comparable researchers as well as the reasons why such a system is required. Next comes the approach that was used when developing the system underneath it. After that, there would be a short conversation on the reasoning behind our decision to use this particular training model rather than one of the many other alternative training models. In addition, before we get to a conclusion, we will discuss the next task, which will contain the modelling instruction.

# Background

The Zoological Society of London is the organisation that first recognised the necessity for the system because of their desire to identify species by the sorts of animal bones they possess and record which ecological category each species belongs to. Because of the problem that the wildlife population has decreased by 69% since 1970 and can disturb the ecology of the current living world, which can eventually lead to problems such as global warming, the Zoological society of London has taken this initiative because [1] have published in their recent report that the wildlife population has decreased by 69% since 1970. This information was published in their most recent report.

Using both medical and technical methods, there are a number of different approaches that may be used to determine the species of the bird. The primary goal of this study is to identify animals via the use of technology; hence, we will be focusing on more conventional forms of machine learning and will examine ongoing efforts to categorise animal species. The specialists in the fields of archaeology and palaeontology were the ones who were responsible for doing these classifications in the past [5]. These traditional professions are able to identify the species of birds by examining the bones of the birds, but the processing of the data takes time, and the resources that are necessary impose a recurring expense. The identification of birds is possible via the use of object detection and acoustic signal processing since birds have numerous distinguishing characteristics; nevertheless, our challenge is to identify the species of a bird by analysing its skeletal structure. The research that was carried out by [6] resulted in the creation of a system that could identify the different species of animals. This system made use of object segmentation by mask-RCNN to predict the bounding boxes, which was then followed by the noisy annotation generator and the addition of synthetic data to expand the dataset. Since we are only training the model for birds and employing a deep learning algorithm would result in an exponential rise in the cost of the system, we do not like to follow this strategy, despite the fact that it is an excellent method in general if there are more classes to begin with. As the current researchers have been able to classify the species using deep learning algorithms, which can drive up the cost of systems, we are going to propose a system that makes use of more conventional machine learning algorithms and gives us the highest possible level of performance and accuracy.

# Methodology

In this part, we will discuss how we will analyse the dataset and perform various operations on the data in order to get it into the best possible condition for training or modelling. We would be carrying both of these sections side by side as it brings more clarity. Therefore, analysing the dataset component is known as Exploratory Data Analysis (EDA), and to normalise our data is known as data preparation. This study focuses mostly on how we want to construct a system that will be able to determine the species of a bird by analysing the length of its bones. In order to get the procedure started, the first thing that we did was access the data set and have a look at the issues that we would need to address. After finishing this step, we came to the conclusion that the provided dataset includes 420 birds. These 420 birds are represented by 10 bone measurements, including the length of Humerus (huml), the diameter of Humerus (humw), the length of Ulna (ulnal), the diameter of Ulna (ulnaw), the length of Femur (feml), the length of Tibiotarsus (tibl), the length of Tarsometatarsus (tarl), and (SO). The efficiency of the system is wholly based on this component, making data set analysis an extremely vital step in the development process.

## Data type and size

EDA starts with an analysis of the many kinds of data that are accessible and a planning phase for the subsequent phases. It is necessary for us to pay attention to the missing values and '0' values in classes where '0' may be treated as defective data since the dataset that is supplied is in the form of a csv file, which is in a table form. Because of this, it is necessary for us to train the concerned model.Graphical user interface, text, application

Description automatically generated

Figure ( a screenshot of the type of data stored in the dataset)

The next step is to evaluate whether or not the dataset has enough information to properly instruct the model in the future. On the other hand, if the dataset is inadequate, there are a number of methods that we may use to enhance its size. These methods include the inclusion of synthetic data as well as the addition of more data from online resources such as Kaggle. to analyse the existing dataset, which has sufficient data and consists of 420 items distributed across 11 categories. the goal is to investigate this dataset.

## Data Cleaning

At this point, we are positive that the quantity of data that is easily accessible is sufficient, so we proceed to the process of data pre-processing. After that, we analyse the data by carrying out the python commands, and we receive a tabular representation of the data in the jupyter notebook. After doing an analysis on the dataset, which consisted of 420 rows and 13 columns as shown in figure 2, and afterwards using further python commands, the total missing values in the dataset were determined and are presented in figure 3. We indicated all of these missing values with the letter "Nan," and we also had to filter out the values of "0," which are acting as the value but they are falsifying the data. For instance, if the heart rate of a patient is zero and that data is used to predict whether or not this person is going to have cancer or not, then the data is not genuine because we are using the data of a deceased person.

Graphical user interface, text, application

Description automatically generated

Figure (a representation of the complete dataset)

A screenshot of a computer

Description automatically generated

Figure (the amount of missing values in each dataset)

## The distribution of Data

checking the distribution of the data on a plot using a histogram, which shows us whether or not the data has a Gaussian form. If you want the model to work perfectly, you need make sure the input data set has a Gaussian distribution, which makes it possible for the feature distribution to be even [7]. The distribution showed in the figure 4 is positively skewed.

A screenshot of a computer

Description automatically generated

Figure ( a histogram representation of the dataset)

## Normalising the distribution of data

As was seen earlier in figure 4, the data in question has a positive skew; hence, we need to normalise the data in order to improve the performance of the model and its capacity to remain stable throughout training [7]. Because we are executing operations and altering the data in order to get the best possible performance from the model, this stage may be thought of as the data pre-processing stage. We utilise the powertransform and quantiletransform functions of the scikit package to normalise the data. These functions provide us the ability to normalise the skewness that exists in our dataset[8].

## Balancing the classes

The subsequent stage in the process of preparing our data is to carry out an analysis of the classes and determine whether or not the classes are evenly balanced. Following the completion of the class balance checks, we will proceed to balance the class by supplying synthetic data to be used for balancing [9].

Graphical user interface, application

Description automatically generated

## Defining a corelation of features

The creation of a correlation between features is another stage in the EDA that is highly significant because it enables our system to make accurate predictions about the values of characteristics that have a high correlation coefficient [10]. As a result, it is the most essential component of both our EDA and our pre-processing. Second, thanks to this correlation, we are able to eliminate characteristics that are superfluous and so save the processing RAM that would have been needed to carry out operations on these features. Figure 5 is an illustration of the correlation matrix for our dataset. A strong correlation is one that is more than 0.7, an average correlation falls between 0.7 and 0.3, and a correlation that is less than 0.3 is considered to be a weak correlation. As a result, in order to get a better comprehension of the correlations, we plot the correlation heat map that is shown in figure 6. This colour codes the matrix and presents better correlation visuals to analyse; however, the matrix is difficult to read, and as a result, we eliminate the upper reflection of the matrix, which leaves us with the heat map that is shown in figure 7.

Graphical user interface, application

Description automatically generated

Figure (correlation matrix created using python)

Graphical user interface, application

Description automatically generated

Figure (Heat map of corelation)

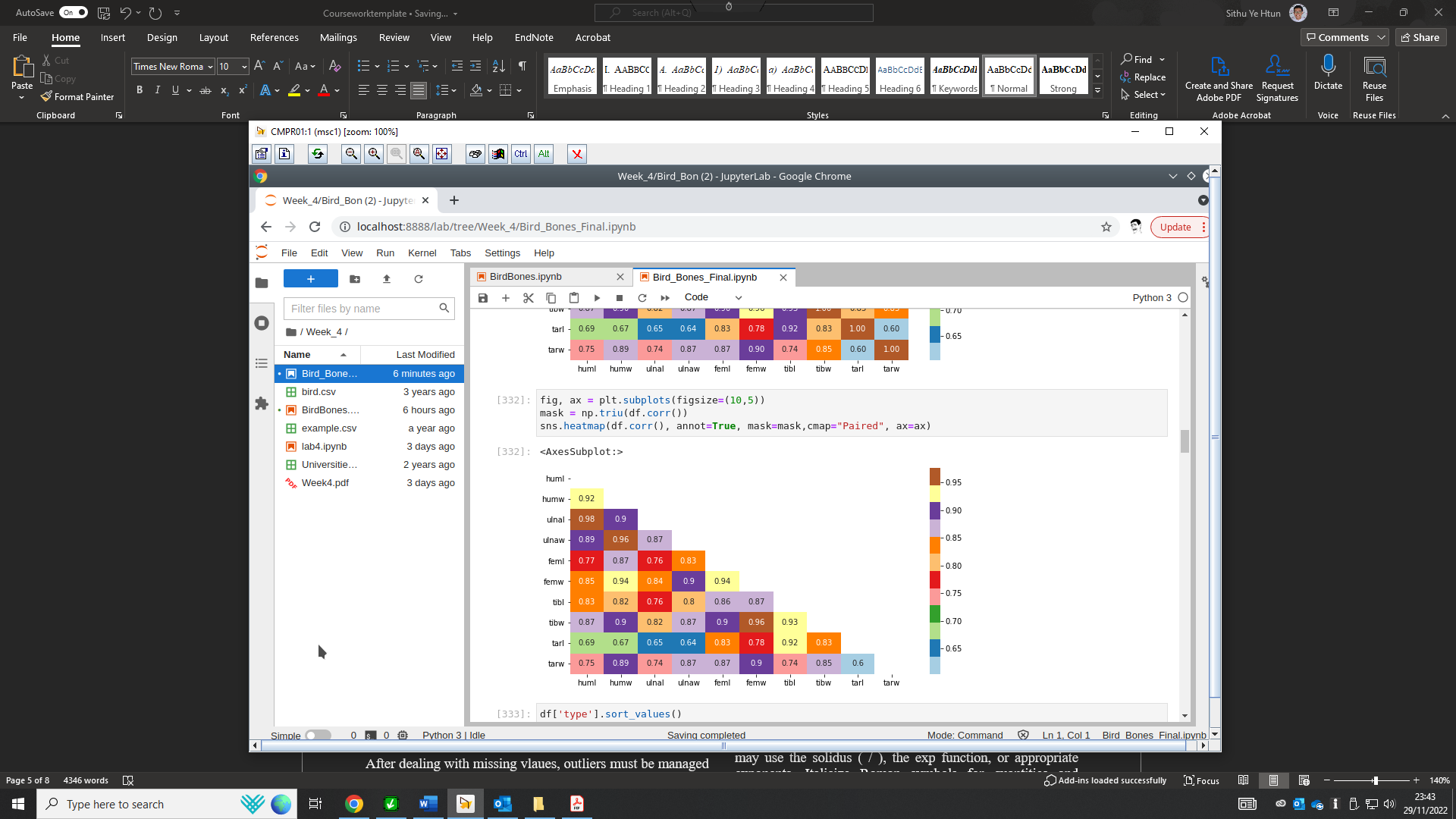


Figure (lower half of the heat map)

## Model architecture

We had multiple other model architecture such as SVM, KNN, and we'll explore the reasons why we eliminated these options and selected random forest as our model instead of these architectures. The model architecture that we selected for our model which is able to handle multiclass problems is called random forest.

# Discussion

In the following paragraphs, we are going to talk about the model architecture that we have decided to use for our system. The random forest model architecture is the one that we have decided to go with. This model is essentially just a collection of decision trees. Nevertheless, we will investigate the functioning of various systems that were under consideration, and then we will come to a conclusion about why we have chosen random forest as the architecture for our model.

## SVM( Support Vector Machine)

The author of [10] explains that Support Vector Machines (SVMs) attempt to differentiate between two or more classes by using the most appropriate hyperplane. They use a one against all strategy regardless of whether they are working on a binary or multiclass issue. Changing the method kernel allows for the hyperplane to take on linear, quadratic, or polynomial forms, among other possible configurations. Although modifying the kernel may help improve class separation, using complicated hyperplanes can cause overfitting. In the event that the data is linear, the graphic on the left of the slide before this one will adequately depict the best hyperplane that should be used. In addition, if the data are nonlinear, the hyperplane that is specified in the centre would be sufficient. It is recommended that you do not use the picture on the far right since it would result in overfitting. In an ideal world, we would want our hyperplane to be able to bend and twist like a flexible cane. In their most fundamental form, hyperplanes are decision boundaries that serve the purpose of assisting in the separation of data points that belong to distinct classes. After that, distinct data classes are assigned to the sites of interest that are located on each side of the hyperplane. SVMs provide us the ability to fit many hyperplanes to our data at the same time.

Chart, scatter chart

Description automatically generated

Figure ( the figure of SVM type of decision making boundaries)

When choosing the optimal hyperplane, one should aim to maximise the value that exists between the hyperplane and the closest datapoint for each class. This choice should also take into account how the data is normalised. Therefore, on the basis of these characteristics, it is a good solution for the multiclass issue; but, since we have normalised the distribution of our data, which the SVM is sensitive to, we are unable to use this model for our system.

## KNN (K Nearest Neighbour)

According to the [10], the k-nearest neighbour (KNN) method is both one of the simplest algorithms and one of the easiest algorithms to put into practise. KNN is a method for supervised machine learning that may be used to the solution of issues involving classification as well as regression. There is no need to construct a model, fine-tune any parameters, or make any assumptions while using KNN. When determining the distance between two points, the complete dataset is put into memory, and then additional observations are displayed against the Euclidean distances calculation. A prediction is obtained by taking the average of the labels produced by the KNN and applying it to the observation. In the process of regression, the k observations that are most similar to the most recent observation are averaged. When doing classification, the mode is what is utilised to decide which class to employ. This model suffers from the same issue as SVM does: while it performs well for classification, it is very sensitive to any attempt to normalise the distribution of the features. Figure 9 provides an illustration of how the KNN may classify data.

A picture containing scatter chart

Description automatically generated

Figure (the graph shows how the K detect nearest neighbours)

## Random forest

According to [10] decision trees are a type of supervised machine learning algorithm that can be applied to a variety of different types of problems, including regression, binary, and multiclass issues. The construction of a number of predefined decision trees is the foundation upon which random forests are built. Bagging and a random selection of features are the foundations upon which the decision tree algorithm was built in order to construct a collection of decision trees. There are two stages involved in the construction of a decision tree.

First, the predictor space should be broken up into non-overlapping regions, and then, second, each region should have its most frequent class label predicted. A prediction of the class is given by each individual tree in the forest. As shown in figure 9, the model will make its prediction based on the category that received the most votes.

Diagram

Description automatically generated with medium confidence

Figure ( the random forest model's functionality)

It's possible that the separation will grow more difficult as additional data points are added. Configuring a decisions forest may involve making use of a variety of different parameters, which may include the following list of options: the Maximum depth of the decision trees is the initial factor that determines the maximum depth that every decision tree in the forest may have. An increase in the depth of the tree may result in an improvement in accuracy; however, this comes with the potential for some overfitting as well as an increase in the amount of time spent training. Second, the amount of random splits that should be used while constructing each node of the tree is referred to as the Number of random splits per node. The term "split" refers to the process of randomly dividing the characteristics present at each node of the tree. Finally, Minimal number of samples necessary to establish a terminal node (leaf) in a tree: the minimum number of instances that must be gathered in order to produce any leaf node in a tree. After looking at all of these model architectures and analysing our issue, we came to the conclusion that, firstly, we have a problem with multi-class identification; consequently, we filtered the supervised learning algorithms that are used for multi-classification; consequently, these three models were further investigated to determine whether or not any of these algorithms are sensitive to the EDA and data pre-processing techniques that we have applied to our data; we discovered that two of these algorithms are sensitive to the manipulation of the data distribution and eliminating the skew in distribution..

# Future work

In this paper we have dealt with the first step of creating an AI algorithm for detecting the species of the birds using the EDA and data pre-processing techniques to prepare the data to be input for training the model. In Future, we aim to train the model using the random for model architecture as explained before. Whereas, we cannot solely depend on the performance of the one model as we will not be able to compare the results to check if we have the best performance available. Hence the second model architecture we are going to use is KNN and to conclude our future work we’ll implement both of these models separately and display the results to justify our model architecture selection.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

[1] WWF, ‘Living Planet Report 2022 | WWF’, 2022. https://livingplanet.panda.org/en-GB/ (accessed Dec. 04, 2022).

[2] M.-F. Tsai and H.-J. Tseng, ‘Enhancing the identification accuracy of deep learning object detection using natural language processing’, *J Supercomput*, vol. 77, no. 7, pp. 6676–6691, Jul. 2021, doi: 10.1007/s11227-020-03525-2.

[3] N. I. Hasan, ‘Bird Species Classification And Acoustic Features Selection Based on Distributed Neural Network with Two Stage Windowing of Short-Term Features’. arXiv, Jan. 01, 2022. doi: 10.48550/arXiv.2201.00124.

[4] S. Schneider, G. W. Taylor, and S. Kremer, ‘Deep Learning Object Detection Methods for Ecological Camera Trap Data’, in *2018 15th Conference on Computer and Robot Vision (CRV)*, May 2018, pp. 321–328. doi: 10.1109/CRV.2018.00052.

[5] K. A. Prassack, M. C. Pante, J. K. Njau, and I. de la Torre, ‘The paleoecology of Pleistocene birds from Middle Bed II, at Olduvai Gorge, Tanzania, and the environmental context of the Oldowan-Acheulean transition’, *Journal of Human Evolution*, vol. 120, pp. 32–47, Jul. 2018, doi: 10.1016/j.jhevol.2017.11.003.

[6] Z. Zhou, G. Hassena, B. C. Weeks, and D. F. Fouhey, ‘Quantifying Bird Skeletons’, p. 5.

[7] Google, ‘Normalization | Machine Learning’, *Google Developers*. https://developers.google.com/machine-learning/data-prep/transform/normalization (accessed Dec. 04, 2022).

[8] Y. Shinde, ‘What is skewness in data? How to fix skewed data in python?’, *Medium*, Aug. 27, 2021. https://yashowardhanshinde.medium.com/what-is-skewness-in-data-how-to-fix-skewed-data-in-python-a792e98c0fa6 (accessed Dec. 04, 2022).

[9] D. S. Goswami, ‘Applying SMOTE for Class Imbalance with just a few lines of code Python’, *Medium*, Feb. 18, 2021. https://towardsdatascience.com/applying-smote-for-class-imbalance-with-just-a-few-lines-of-code-python-cdf603e58688 (accessed Dec. 04, 2022).

[10] P. Fergus and C. Chalmers, ‘Supervised Learning’, in *Applied Deep Learning: Tools, Techniques, and Implementation*, Cham: Springer International Publishing, 2022, pp. 63–93. doi: 10.1007/978-3-031-04420-5\_3.