Machine Learning-Based Approach to Identify Bird Ecological Groups using Skeletal Measurements

Aiswarya Kattappilly Ravi  
*Computer Science dept.   
Liverpool John Moores University*Liverpool, United Kingdom  
csmakatt@ljmu.ac.uk

Abstract: This paper addresses the problem of bird species classification using machine learning techniques. The aim is to accurately classify different bird species based on their unique characteristics. Bird bone classification plays a crucial role in various scientific disciplines, including ornithology, paleontology, and forensic anthropology. Traditional manual identification methods are time-consuming and prone to errors.

In this study, I focus on investigating the application of machine learning algorithms to analyze a dataset of bird bones provided by the Zoological Society of London. I develop and compare different ML models, including Random Forest, Support Vector Machine (SVM), and Decision Tree classifiers, using a dataset of bird bone samples.

The models are trained, validated, and evaluated using cross-validation techniques to assess their performance. I analyze and interpret the results, considering factors such as accuracy, precision, recall, and F1-score. Additionally, I consider computational efficiency, interpretability, and the specific requirements of the problem at hand to select the most suitable model.

The findings demonstrate the efficacy of ML algorithms in accurately classifying bird bones, offering potential for improved efficiency and accuracy in various domains involving avian skeletal analysis.

Keywords—Random Forest Classifier, SVM, Decision tree Classifier.

# Introduction

Birds have long been a source of fascination for researchers and nature enthusiasts alike, thanks to their diverse characteristics and ecological adaptations. The classification of birds into different ecological groups based on their living environments and habits provides valuable insights into their behaviours, morphology, and evolutionary history. However, traditional methods of identifying bird species and their ecological groups have relied on labour-intensive and error-prone manual observations, often requiring specialized skills.

Fortunately, recent advancements in machine learning have shown remarkable progress in various domains, including image recognition, natural language processing, and data analysis. As a result, there is a growing interest in exploring the application of machine learning algorithms to automate and improve the classification of bird species and their associated ecological groups. Such advancements can significantly enhance our understanding of avian diversity and contribute to conservation efforts [1][2].

In the field of non-deep learning approaches, traditional machine learning algorithms such as Random Forest, Decision Tree, and Support Vector Machine (SVM) have been widely used for bird species classification. These algorithms leverage handcrafted features extracted from bird images or audio recordings to train models that can classify different species. Previous studies have reported promising results using these algorithms, achieving high accuracy and demonstrating the feasibility of using non-deep learning methods for bird classification[3]-[8].

On the other hand, deep learning approaches have gained significant attention in recent years for various image classification tasks. Convolutional Neural Networks (CNNs) have shown remarkable performance in tasks such as object recognition and image classification. Researchers have also explored the application of CNNs in bird species classification, training models on large-scale datasets and achieving impressive accuracy levels. These deep learning approaches leverage the power of deep neural networks to automatically learn features from raw image data, eliminating the need for manual feature engineering [9][10].

Recognizing the potential of machine learning in addressing the challenges associated with identifying bird species and assigning them to ecological groups based on skeletal data, the Zoological Society of London has launched a research project to investigate the feasibility of using machine learning techniques to analyse and classify bird species using bone measurements. By automating the process, valuable time and resources can be saved while potentially reducing human error. This research has the potential to revolutionize the field of avian classification and contribute to our understanding of the natural world.

Despite the progress made in bird species classification research, there are still limitations that need to be addressed. Firstly, the existing literature lacks comprehensive evaluations and comparisons of different algorithms specifically for bird classification. While individual studies have reported high accuracy, it is essential to have a comparative analysis to determine the most effective approach. Additionally, the interpretability of models and their computational efficiency are important considerations, especially in real-world applications and scenarios where explain ability and efficiency are crucial factors.

In this study, I propose to address these limitations by conducting a systematic evaluation and comparison of Random Forest, Decision Tree, and SVM algorithms for bird species classification. I will train and evaluate these models on a labelled bird dataset, considering metrics such as accuracy, precision, recall, and F1-score. Moreover, I will analyse the interpretability and computational efficiency of each model. By considering these factors, I aim to identify the most suitable approach for accurate and efficient bird species classification.

The rest of this paper is organized as follows:

Section II includes the challenges outlined by the Zoological Society of London, data analysis, machine learning, and its role in classifying species using skeletal data.

Section III describes description of the dataset along with

appropriate Exploratory Data Analysis (EDA).

Section IV contains methodology chosen for solving the problem and performance evaluation.

Section V explains challenges and justification of chosen methodology.

Section VI gives the information about the future work.

# background

The Zoological Society of London has presented a challenge that involves the classification of species using skeletal data. Skeletal data refers to the measurements and characteristics of animal skeletons, which can provide valuable information for species identification and classification. Data analysis is a crucial aspect of this challenge as it involves extracting meaningful insights from the collected skeletal data. By analysing the skeletal features, patterns, and measurements, researchers can identify distinguishing characteristics that differentiate one species from another. This analysis may involve various techniques such as statistical analysis, dimensionality reduction, and feature extraction to uncover the relevant information embedded in the skeletal data [11].

Machine learning techniques have shown great potential in species classification tasks. These algorithms can automatically learn patterns and relationships from large datasets, allowing them to make accurate predictions and classifications. In the context of species classification using skeletal data, machine learning algorithms can be trained on labelled datasets where the skeletal features are associated with known species. The trained models can then be used to classify new skeletal samples and assign them to the appropriate species.

Machine learning models for species classification using skeletal data can employ various approaches, including both traditional non-deep learning methods and deep learning methods. Non-deep learning methods, such as decision trees, random forests, and support vector machines, can provide interpretable and efficient solutions. On the other hand, deep learning approaches, such as convolutional neural networks (CNNs), can automatically learn hierarchical representations of the skeletal data, potentially capturing intricate patterns and improving classification accuracy [12].

Data analysis is the process of inspecting, cleaning, transforming, and modelling data in order to discover meaningful patterns, draw conclusions, and make informed decisions. It involves applying various techniques and methods to extract insights and information from data, ultimately leading to a deeper understanding of the underlying phenomena or trends. By applying statistical techniques, visualization methods, and modelling approaches, data analysis uncovers patterns, trends, and relationships in the data that may not be apparent initially. It helps in identifying opportunities, solving problems, mitigating risks, optimizing processes, and improving overall decision-making.

Despite the significant advancements in data analysis and machine learning techniques, there are still challenges and limitations in species classification using skeletal data. These challenges include dealing with incomplete or noisy data, handling class imbalance, selecting relevant features, and addressing potential biases in the data collection process. Moreover, the interpretability of deep learning models may be a concern in certain applications where explain ability is crucial.

To overcome these challenges, this research aims to propose an effective methodology for species classification using skeletal data. The proposed approach will involve comprehensive data analysis techniques to pre-process and explore the skeletal data. Machine learning models will be trained and evaluated to achieve accurate and interpretable species classification. The research will also focus on addressing the limitations and challenges mentioned earlier, such as handling incomplete data and ensuring the robustness and generalizability of the models [13].

The proposed methodology will be a significant contribution to the field of species classification using skeletal data. It will provide a comprehensive and effective approach to overcome the challenges and limitations in this area.

The research will also have practical implications, as accurate species classification is crucial in various fields, such as forensic science, archaeology, and biology. By addressing the challenges and limitations in species classification using skeletal data, this research will pave the way for more accurate and reliable species identification in the future.

# Methodology

Conducting Exploratory Data Analysis (EDA) on the bird bone dataset is crucial in understanding the relationships between bone measurements and the ecological groups of birds. This dataset comprises 420 bird samples, each represented by 10 measurements related to the length and diameter of specific bones. The dataset is structured as a CSV file with 420 rows and 11 columns. The columns include an ID column, bone measurement columns (humerus length and diameter, ulna length and diameter, femur length and diameter, tibiotarsus length and diameter, tarsometatarsus length and diameter), and an ecological group (type) column. This dataset is a collection from the Natural History Museum of Los Angeles County, representing 21 orders, 153 genera, and 245 species. It provides valuable insights into the physical characteristics of different bird species and their ecological groups.

By analysing this dataset, I can gain valuable insights into the physical characteristics of birds and how they relate to their ecological groups.

## EDA

The dataset comprises measurements of various bird species categorized into six ecological groups: Swimming Birds (SW), Wading Birds (W), Terrestrial Birds (T), Raptors (R), Scansorial Birds (P), and Singing Birds (SO). The bone measurements are continuous float numbers (mm), and any missing values are indicated by empty strings (Fig. 1).

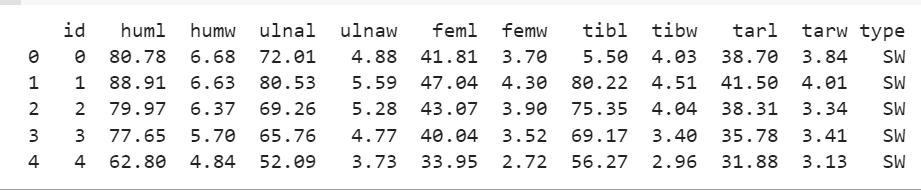


Fig. 1 dataset of bird

#### Data Cleaning

I checked the null values in dataset, and I saw that there exist some null values and which is shown in Fig. 2. To handle the null values in the Bird Bone dataset, I used the fillna() method from the pandas library to replace the missing values with mean value of the corresponding column.

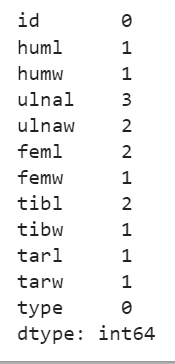
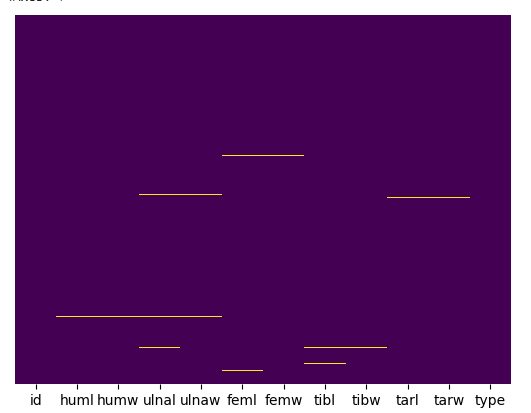
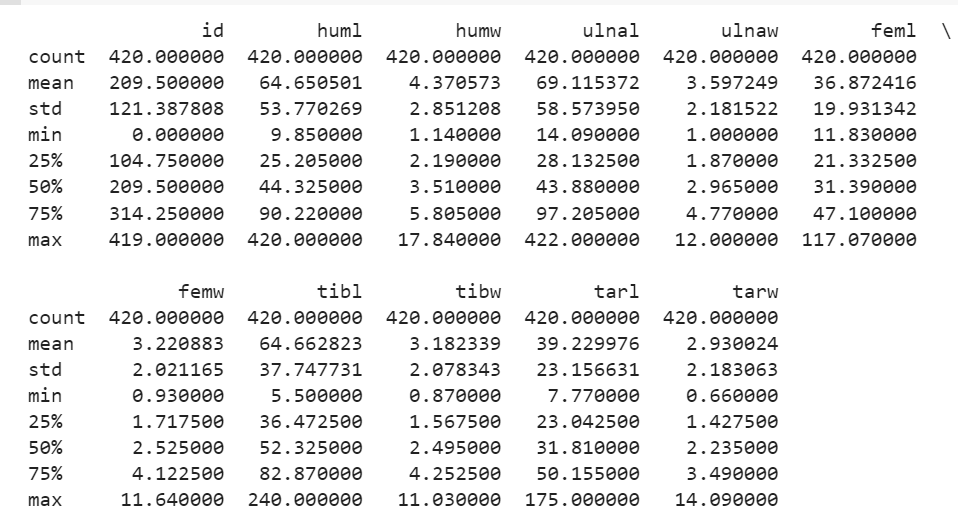
 

Fig. 2 Null values in dataset

#### Descriptive Statistics

By analysing these statistics, I gain insights into the distribution and variation of bone measurements. I observed the average, range, and spread of each measurement, identify any potential outliers, and understand the overall characteristics of the bone sizes in the dataset.

Fig. 3 Statistic of dataset

The Fig. 3 provides the descriptive statistics for each bone measurement column in the dataset.

* Count: The number of non-missing values in each column. All columns have a count of 420, indicating that there are no missing values.
* Mean: The average value of each column. It represents the central tendency of the data.
* Std: The standard deviation, which measures the spread or dispersion of the data. Higher values indicate greater variability.
* Min: The minimum value in each column, representing the smallest observed value.
* 25%: The first quartile, which is the value below which 25% of the data falls.
* 50%: The second quartile or median, which is the value below which 50% of the data falls.
* 75%: The third quartile, which is the value below which 75% of the data falls.
* Max: The maximum value in each column, representing the largest observed value.

#### Detect outliers and resolve

Outliers are data points that deviate significantly from the majority of the data. I used Z-score to detect outliers in dataset, it is a statistical measure used to detect outliers by calculating how many standard deviations a data point is away from the mean. A Z-score above or below a certain threshold (typically 3) indicates an outlier. To detect outliers using Z-score:

* Calculate the Z-score for each data point in a column,
* Set a threshold (e.g., 3) for outlier detection
* Identify data points with Z-scores exceeding the threshold as outliers.

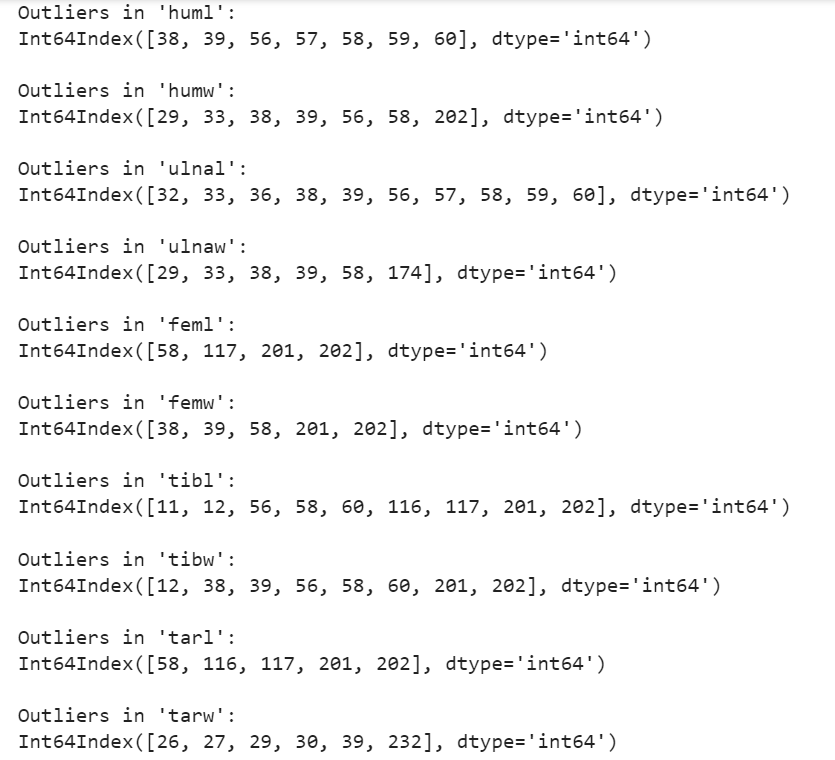


Fig. 4 Outliers in dataset

Instead of removing the outliers (Fig. 4) entirely, I minorize the data by replacing the extreme values with less extreme values within a certain range. This approach helps minimize the impact of outliers while retaining some information. I visualized the new dataset by using box plots to ensure that the outlier was removed (Fig. 5).

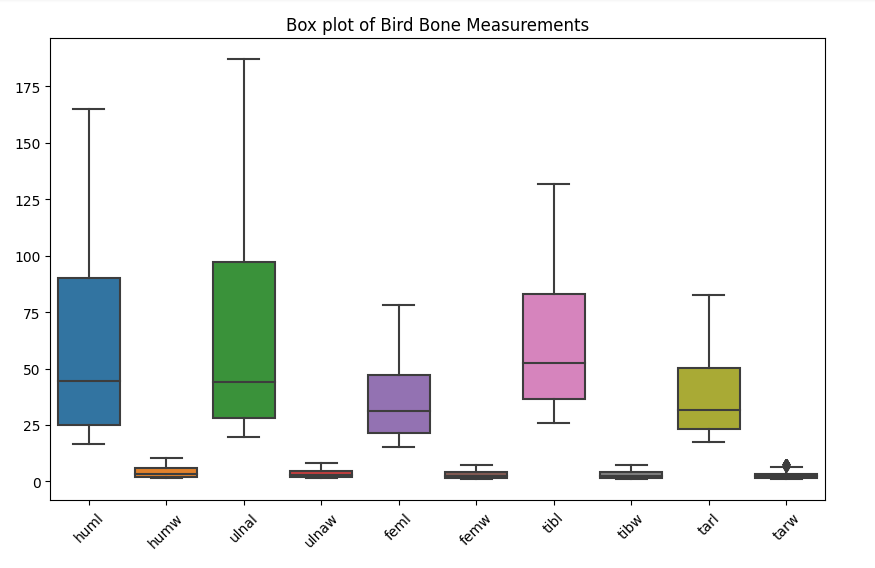


Fig. 5 Boxplot

#### Balance the class

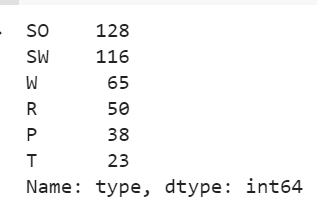
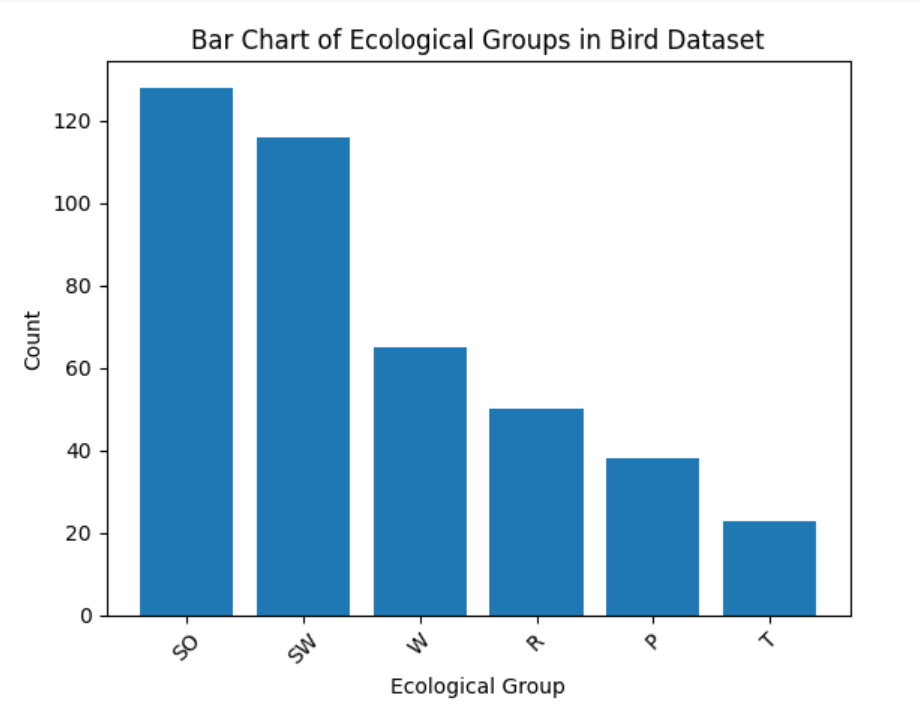


Fig. 6 Frequency of ecological group

Fig. 6 provides insights into the distribution and representation of different bird ecological groups in the dataset. It shows that Singing Birds and Swimming Birds are the most abundant, while Terrestrial Birds have the least representation.

The frequency analysis of the ecological groups in the bird dataset highlights the presence of class imbalance. Class imbalance refers to an unequal distribution of samples across different classes in a dataset. The analysis reveals that the 'SO' (Singing Birds) class has the highest frequency with 128 occurrences, indicating a relatively dominant presence. On the other hand, the 'T' (Terrestrial Birds) class has the lowest frequency with only 23 occurrences, suggesting a significant under-representation. This class imbalance can pose challenges in training machine learning models as it may result in biased predictions and reduced accuracy for minority classes. It is important to address class imbalance to ensure fair and reliable model performance. To overcome this class imbalance, I use oversampling with the Synthetic Minority Over-sampling Technique (SMOTE) [14][15]. SMOTE generates synthetic samples for the minority classes by interpolating new instances between the existing minority class instances. This helps to increase the representation of the minority classes without duplicating the existing samples. By applying SMOTE, the dataset can be rebalanced, and the classifier will have a more balanced distribution of samples to learn from, potentially improving its performance in predicting the minority classes.

#### Feature skewness

Feature skew refers to the asymmetry or lack of symmetry in the distribution of a feature (variable) in a dataset. Skewness measures the degree and direction of this asymmetry. It is an important statistical property as it provides insights into the shape and characteristics of the data distribution. Skewness can have three possible values:

* Positive Skewness: If the tail of the distribution is longer on the right side and the majority of the data points are concentrated on the left side, the distribution is said to be positively skewed or right skewed.
* Negative Skewness: If the tail of the distribution is longer on the left side and the majority of the data points are concentrated on the right side, the distribution is said to be negatively skewed or left-skewed.
* Zero Skewness: If the distribution is perfectly symmetric, the skewness value will be zero.

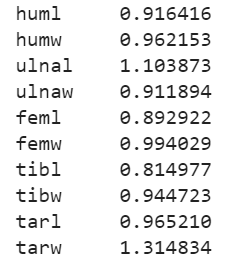


Fig. 7 Skewness of features

From Fig. 7, it’s clear that all features exhibit positive skewness, indicating that their distributions are skewed towards higher values. The higher the skewness value, the more pronounced the skewness in the distribution. Skewness can impact data analysis and modelling, and in some cases, it may be necessary to apply appropriate transformations to mitigate the skewness and achieve a more symmetric distribution if required for the analysis or modelling task. To address the skewness in the dataset, I apply a transformation to the skewed features. One commonly used transformation is the logarithmic transformation, which can help reduce the skewness of positively skewed features. The logarithmic transformation is a mathematical operation that applies the natural logarithm function to the data.

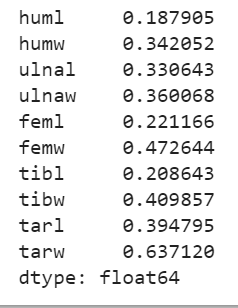


Fig. 8 Skewness of features after applying logarithmic transformation

#### Label Encoding

Many machine learning algorithms require numerical input. By encoding categorical variables into numerical labels, it ensures that the data is in a format that can be processed by these algorithms. By converting categorical variables into numerical representations, we can incorporate them into the analysis and model training process. This increase in dimensionality can impact the computational resources required for training the model. Hence, I Label Encoder to convert type column values to numerical values.

#### Feature selection using Chi square

Feature selection is an essential step in machine learning to identify the most informative features from a dataset. One popular technique for feature selection is chi-square analysis, which measures the dependence between categorical features and a categorical target variable. By calculating the chi-square, it determines the significance of the relationship between each feature and the target. Features with higher chi-square scores are considered more relevant and likely to contribute valuable information for prediction. By selecting features based on chi-square analysis, reduce the dimensionality of the dataset and improve the efficiency and accuracy of machine learning models [16].

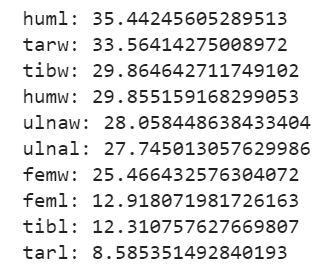


Fig. 9 Chi square value of features

#### Fig. 9 represents the chi-square scores for each feature in descending order. The higher the chi-square score, the more significant the relationship between the feature and the target variable. In this case, the "huml" feature has the highest chi-square score of 35.44, indicating a strong relationship with the target variable. Similarly, "tarw," "tibw," "humw," and "ulnaw" also have relatively high chi-square scores, suggesting their importance in predicting the target variable. On the other hand, "tarl" has the lowest chi-square score of 8.59, indicating a weaker relationship with the target. These chi-square scores can guide feature selection, allowing us to prioritize the most informative features and discard less relevant ones, leading to improved model performance.

#### Correlation Matrix

A correlation matrix is used to measure the linear relationship between variables in a dataset. It provides a comprehensive overview of how each variable is related to every other variable in the dataset. The correlation matrix is typically represented as a square matrix, where each element represents the correlation coefficient between two variables. Correlation coefficients can take values between -1 and 1, where:

* A value of 1 indicates a perfect positive correlation, meaning that the variables move in perfect harmony in the same direction.
* A value of -1 indicates a perfect negative correlation, meaning that the variables move in perfect opposition to each other.
* A value of 0 indicates no linear correlation between the variables.

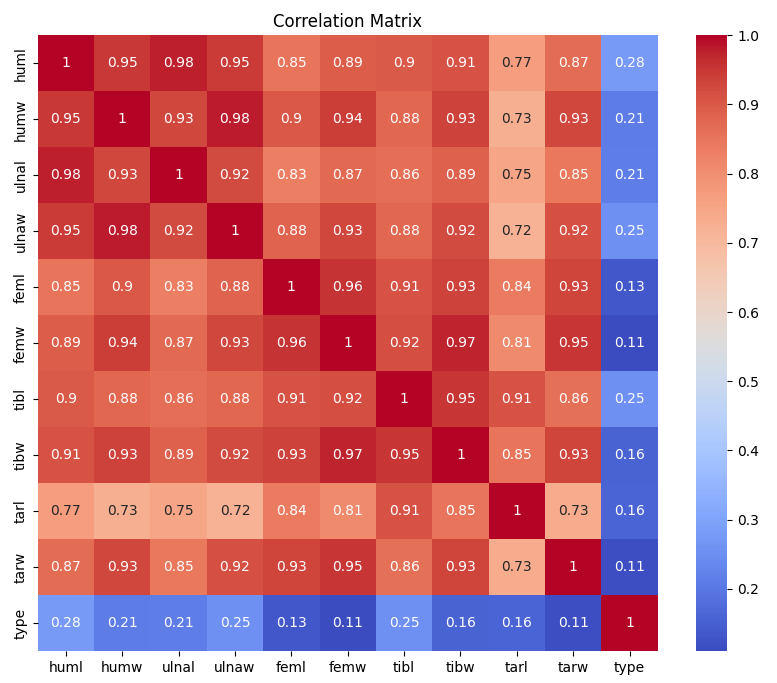


Fig. 10 Correlation Matrix

Fig. 10 shows that huml is highly correlated with humw, ulnal, ulnaw, feml, femw, tibl, tibw, tarl, and tarw. humw is highly correlated with ulnal, ulnaw, feml, femw, tibl, tibw, tarl, and tarw. ulnal is highly correlated with ulnaw, feml, femw, tibl, tibw, tarl, and tarw. ulnaw is highly correlated with feml, femw, tibl, tibw, tarl, and tarw. feml is highly correlated with femw, tibl, tibw, tarl, and tarw. femw is highly correlated with tibl, tibw, tarl, and tarw. tibl is highly correlated with tibw, tarl, and tarw. tibw is highly correlated with tarl and tarw. tarl is highly correlated with tarw.

Instead of removing features, I apply dimensionality reduction techniques like Principal Component Analysis (PCA) to transform the correlated features into a smaller set of uncorrelated components. This can help retain most of the information while reducing the dimensionality of the dataset. Hence, I standardize the features using StandardScaler to ensure that all features have zero mean and unit variance. Next, create an instance of PCA and specify the number of components to keep. Then fit the PCA model to the standardized features and transform the features into the reduced dimensional space using fit\_transform. The resulting transformed features are stored in a new DataFrame along with the target variable. Finally, I print the explained variance ratio, which indicates the amount of variance explained by each principal component and display the transformed dataset with reduced dimensionality (Fig. 11).

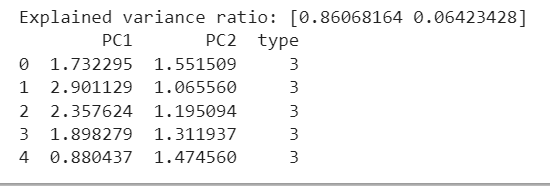


Fig. 11 Dataset with reduced dimensionality

The output shows the explained variance ratio of the principal components obtained through PCA. In this case, the explained variance ratio is [0.86068164, 0.06423428]. The explained variance ratio represents the proportion of variance explained by each principal component. Here the first principal component (PC1) explains approximately 86.07% of the total variance in the dataset, while the second principal component (PC2) explains around 6.42% of the variance. The transformed dataset is also displayed, where each row represents a data point with the corresponding values for PC1, PC2, and the target variable ('type'). The PC1 and PC2 values represent the reduced dimensionality features obtained through PCA.

## Data preprocessing

The data is split into training and testing sets using an 80:20 ratio, where 80% of the data is used for training and 20% is used for testing. The features and target variable are split into X\_train, X\_test, y\_train, and y\_test variables, respectively, to store the respective splits. This splitting allows for independent evaluation of the model's performance on unseen data during testing. The training set is used to train the model, while the testing set is used to assess its performance and generalization capabilities.

## Methodology

#### I have used three models to train the bird dataset.

#### Decision Tree Classifier

The decision tree classifier is a straightforward and interpretable model that can be effective for classification tasks. It partitions the feature space based on decision rules, making it easy to understand and interpret the decision-making process. However, decision trees can be prone to overfitting, particularly when the tree depth is not properly controlled. Regularization techniques can help mitigate overfitting, such as setting maximum tree depth or using pruning methods. Decision trees also provide feature importance measures, allowing for identification of relevant features. While decision trees excel in interpretability and computational efficiency, they may struggle to capture complex nonlinear relationships. In such cases, ensemble methods like random forests or gradient boosting can be more effective.

The hyperparameters of the decision tree classifier were tuned using a param\_grid. This grid consisted of several hyperparameters that were essential for fine-tuning the decision tree model. The criterion parameter determined the splitting criterion, with options such as 'gini' and 'entropy'. The max\_depth parameter set the maximum depth of the tree, providing flexibility with values like 3, 5, 7, or None for an unlimited depth. Additionally, min\_samples\_split defined the minimum number of samples required to split an internal node, while min\_samples\_leaf set the minimum number of samples required to form a leaf node. Finally, the max\_features parameter determined the maximum number of features to consider for the best split, with options like 'auto', 'sqrt', 'log2', or None to include all features. By utilizing grid search with this param\_grid, the decision tree model's hyperparameters were efficiently tuned to enhance its performance and achieve optimal results for the specific task at hand.

#### Random Forest Classifier

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. The classifier was trained on a labelled dataset containing various features relevant to bird characteristics. The Random Forest algorithm operates by creating a multitude of decision trees and aggregating their predictions to obtain the final classification. This approach helps to mitigate the problem of overfitting and improves the overall accuracy and robustness of the model.

To optimize the performance of the random forest model, a grid search was performed to explore different combinations of hyperparameters. The hyperparameters considered in the grid search included the number of trees in the forest (n\_estimators), the maximum depth of the trees, the minimum number of samples required to split an internal node (min\_samples\_split), and the minimum number of samples required to be at a leaf node (min\_samples\_leaf). Multiple values were evaluated for each hyperparameter, such as 100, 200, and 300 for n\_estimators, None, 5, and 10 for max\_depth, and 2, 5, and 10 for min\_samples\_split. The grid search aimed to identify the best combination of hyperparameters that would yield optimal accuracy and generalization performance. By exhaustively searching through the provided parameter grid, the random forest classifier was effectively fine-tuned to achieve the best possible results for bird classification.

#### SVM

Support vector machine (SVM) classifier was employed for the purpose of bird classification. SVM is a powerful supervised learning algorithm that separates data points into different classes by finding an optimal hyperplane in the feature space. The SVM classifier works by mapping the data to a higher-dimensional space, where it tries to find a hyperplane that maximally separates the different classes. It is a discriminative model that focuses on the boundaries between classes rather than the entire distribution of data. SVM has been widely used in various applications due to its ability to handle both linearly separable and non-linearly separable data through the use of different kernel functions. By utilizing a kernel function, SVM can effectively transform the input data to a higher-dimensional space, allowing it to capture complex relationships and achieve accurate classification results. The SVM classifier has proven to be effective in the field of bird classification, demonstrating its capability to accurately distinguish between different bird species based on their unique features and characteristics.

SVM classifier was configured with a parameter grid to explore different combinations of hyperparameters. The parameter grid consisted of the regularization parameter 'C', which controls the trade-off between maximizing the margin and minimizing the classification errors. The 'kernel' parameter specified the type of kernel function to be used, with options including 'linear', 'rbf' (Radial Basis Function), and 'poly' (Polynomial). The 'gamma' parameter determined the kernel coefficient and had two possible values: 'scale' and 'auto'. By systematically varying these hyperparameters, the study aimed to find the optimal combination that maximizes the SVM classifier's performance in accurately classifying bird species based on their distinct features and characteristics.

## Evaluation

In machine learning, evaluation metrics are used to assess the performance and effectiveness of a model in making predictions or classifications. These metrics provide quantitative measures that help understand how well the model is performing and how reliable its predictions are. Here are the evaluation matrices that I used to identify my model’s performance.

#### Accuracy

Accuracy measures the overall correctness of the predictions made by the model. It is calculated as the ratio of correct predictions to the total number of predictions. However, accuracy alone may not be sufficient if the dataset is imbalanced or if certain classes are more important than others.

#### Precision

#### Precision is the ratio of true positives to the sum of true positives and false positives. It measures the ability of the model to correctly identify positive samples. A high precision indicates a low number of false positives.

#### Recall (Sensitivity or True Positive Rate)

Recall is the ratio of true positives to the sum of true positives and false negatives. It measures the ability of the model to correctly identify all positive samples. A high recall indicates a low number of false negatives.

#### F1-Score

The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of the model's accuracy by considering both precision and recall. It is useful when the dataset is imbalanced or when false positives and false negatives have different consequences.

#### Specificity (True Negative Rate)

Specificity is the ratio of true negatives to the sum of true negatives and false positives. It measures the ability of the model to correctly identify negative samples. It is particularly important in binary classification problems.

#### These evaluation metrics helped me to compare different models, selecting the best model for task, and understanding the strengths and weaknesses of the models.

# Discussion

The challenge presented by the Zoological Society of London for 2020 required developing a machine learning methodology for bird classification. This task is of significant importance for biodiversity monitoring and conservation efforts.

In my proposed methodology, I employed a combination of feature engineering, oversampling, and ensemble learning techniques to address the challenge effectively. Feature engineering played a crucial role in capturing relevant information from the given bird dataset. By selecting and transforming the available features, I aimed to enhance the discriminative power of the input data and improve the accuracy of the classification model. To address the issue of class imbalance within the dataset, I incorporated oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique). This approach helped in generating synthetic samples for the minority class, thereby balancing the class distribution and mitigating the bias towards the majority class. By oversampling the minority class, I aimed to prevent the model from being biased towards predicting the majority class, leading to more accurate predictions for rare bird species.

The accuracy, precision, recall, and F1-score for each model were consistently high, indicating that they successfully classified the bird species with great accuracy. The confusion matrices for all models showed minimal misclassifications, with most of the instances being correctly classified across all classes. This suggested that the models effectively captured the underlying patterns and features in the bird dataset. Looking at the individual evaluation metrics, the Random Forest, Decision Tree, and SVM models exhibited similar performance. They all achieved an accuracy of 98.81%, indicating that they correctly classified approximately 98.81% of the instances in the test set. The precision, recall, and F1-score for all models were also consistently high at around 0.99, demonstrating a high level of precision in predicting the positive class (bird species) and a high recall in capturing the actual positive instances.

Furthermore, considering the cross-validation scores, the Random Forest and Decision Tree models exhibited a mean accuracy of 0.8429 and 0.8595, respectively, while the SVM model had a slightly higher mean accuracy of 0.8595. These scores suggested that the models performed consistently well across different validation sets, further indicating their robustness and generalizability. Based on these findings, it can be concluded that all three models, namely Random Forest, Decision Tree, and SVM, were effective and suitable for the bird classification task.

The choice of the best model among these three would have depended on additional factors such as model complexity, interpretability, computational efficiency, and specific requirements of the task. Based on these factors, the Decision Tree model was a suitable choice for the bird classification task.

* Model Complexity: Decision Trees were known for their simplicity and interpretability. They consisted of a series of if-else statements that could be easily understood and visualized. This made it easier to interpret and explain the model's predictions, which was valuable in applications where interpretability was important.
* Interpretability: Decision Trees provided clear and intuitive decision rules based on the features of the data. This allowed domain experts to understand and validate the decision-making process of the model. The interpretability of Decision Trees was particularly beneficial in scenarios where explain ability was necessary, such as in wildlife conservation efforts.
* Computational Efficiency: Decision Trees were computationally efficient, especially when compared to more complex models like Random Forest or SVM. The training and prediction times for Decision Trees were typically faster, making them a practical choice for large datasets or real-time applications.
* Specific Requirements: Decision Trees were known for their versatility and ability to handle both categorical and numerical data. They could capture nonlinear relationships and interactions between features, which were crucial for accurately classifying different bird species based on their unique characteristics.

Considering all these factors, the Decision Tree model struck a good balance between model complexity, interpretability, and computational efficiency. It provided accurate predictions while being relatively easy to understand and compute.

# Future worK

a) Step 1: Load the libraries and dataset

To start the implementation of the classification solution, the necessary libraries and the bird dataset will be loaded. This includes importing the required Python libraries such as scikit-learn, pandas, and numpy, as well as loading the bird dataset into a data frame.

b) Step 2: Data Pre-processing

This includes handling missing values, encoding categorical variables, and scaling numerical features. The goal is to ensure the data is in a suitable format for training the model.

c) Step 3: Feature Selection

It is essential to identify the most relevant features that contribute to the classification task. Feature selection techniques such as correlation analysis, information gain, or recursive feature elimination can be applied to identify the most informative features model.

d) Step 4: Making Test and Train Sets

To evaluate the performance of the classification model, the dataset will be split into training and test sets. The commonly used 80:20 ratio will be applied, where 80% of the data will be used for training the model, and the remaining 20% will be reserved for evaluating the model's performance.

e) Step 5: Building and Evaluating the Model

The next step involves building the model and evaluating its performance. The model will be trained using the training dataset. After training, the model will be tested using the test dataset to assess its accuracy and generalization ability. Evaluation metrics such as accuracy, precision, recall, and F1-score will be calculated to measure the model's performance. Additionally, the confusion matrix will provide insights into the model's ability to correctly classify bird species across different classes.

# References

1. Z. Yutong, "Multi-lingual Transfer Learning for Intent Classification," 2021 3rd International Conference on Natural Language Processing (ICNLP), Beijing, China, 2021, pp. 168-173, doi: 10.1109/ICNLP52887.2021.00034.
2. C. Shi, G. Ji and Y. Wang, "Study of Texture Images Classification Method Based on Fractal Dimension Calculation," 2009 International Joint Conference on Artificial Intelligence, Hainan, China, 2009, pp. 488-491, doi: 10.1109/JCAI.2009.105.
3. J. Liu, Q. Y. Xu and W. S. Chen, "Classification of Bird and Drone Targets Based on Motion Characteristics and Random Forest Model Using Surveillance Radar Data," in IEEE Access, vol. 9, pp. 160135-160144, 2021, doi: 10.1109/ACCESS.2021.3130231.
4. B. Qiao, Z. Zhou, H. Yang and J. Cao, "Bird species recognition based on SVM classifier and decision tree," 2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS), Harbin, China, 2017, pp. 1-4, doi: 10.1109/EIIS.2017.8298548.
5. S. Fagerlund, "Bird Species Recognition Using Support Vector Machines", Eurasip Journal on Advances in Signal Processing, vol. 7, no. 1, pp. 1-8, 2007.
6. C. Kwan et al., Bird classification algorithms: theory and experimental results, vol. 5, pp. 289-292, 2004.
7. S. Fagerlund, "Bird Species Recognition Using Support Vector Machines", Eurasip Journal on Advances in Signal Processing, vol. 7, no. 1, pp. 1-8, 2007.
8. Ricky Mohanty, Bandi Kumar Mallik, Sandeep Singh Solanki, "Automatic bird species recognition system using neural network based on spike", Applied Acoustics, vol.161, pp.107177, 2020.
9. J. Xie, K. Hu, M. Zhu, J. Yu and Q. Zhu, "Investigation of Different CNN-Based Models for Improved Bird Sound Classification," in IEEE Access, vol. 7, pp. 175353-175361, 2019, doi: 10.1109/ACCESS.2019.2957572.
10. . I. Himawan, M. Towsey and P. Roe, "3D convolutional recurrent neural networks for bird sound detection", Proc. DCASE Challenge, pp. 108-112, Sep. 2018.
11. R. Hafen and T. Critchlow, "EDA and ML -- A Perfect Pair for Large-Scale Data Analysis," 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum, Cambridge, MA, USA, 2013, pp. 1894-1898, doi: 10.1109/IPDPSW.2013.118.
12. H. Dale, C. Baker, M. Antoniou and M. Jahangir, "An Initial Investigation into Using Convolutional Neural Networks for Classification of Drones," 2020 IEEE International Radar Conference (RADAR), Washington, DC, USA, 2020, pp. 618-623, doi: 10.1109/RADAR42522.2020.9114745.
13. P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu and C. Zhang, "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks," 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 2021, pp. 13-24, doi: 10.1109/ICDE51399.2021.00009.
14. E. Mbunge et al., "Implementation of ensemble machine learning classifiers to predict diarrhoea with SMOTEENN, SMOTE, and SMOTETomek class imbalance approaches," 2023 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 2023, pp. 1-6, doi: 10.1109/ICTAS56421.2023.10082744.
15. X Guo, Y Yin, C Dong, G Yang and G. Zhou, "On the class imbalance problem", Proc - 4th Int Conf Nat Comput ICNC, vol. 4, pp. 192-201, 2008.
16. . S. Ray, K. Alshouiliy, A. Roy, A. AlGhamdi and D. P. Agrawal, "Chi-Squared Based Feature Selection for Stroke Prediction using AzureML," 2020 Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249117.
17. D. Barik and M. Mondal, "Object identification for computer vision using image segmentation," 2010 2nd International Conference on Education Technology and Computer, Shanghai, China, 2010, pp. V2-170-V2-172, doi: 10.1109/ICETC.2010.5529412.
18. Xin Zhang; Shuo Xu, “Research on Image processing Technology and computer vision”, 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL).
19. M. Shanmuganathan and C. Nalini, "Deep Learning Approaches to Image Classification," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 263-268, doi: 10.1109/ICIRCA48905.2020.9182927.
20. C. Hary and S. Mandala, "Object Detection Analysis Study in Images based on Deep Learning Algorithm," 2022 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2022, pp. 226-231, doi: 10.1109/ICoDSA55874.2022.9862922.
21. K. Li, "Applications of Deep Learning in Object Detection," 2022 International Conference on Computers, Information Processing and Advanced Education (CIPAE), Ottawa, ON, Canada, 2022, pp. 436-442, doi: 10.1109/CIPAE55637.2022.00097.