**International Journal of Ecosystems and Ecology Science (IJEES)** **Vol. 13 (3):** 39-44 (2023)
https://doi.org/10.31407/ijees https://doi.org/10.31407/ijees13.3

# MACHINE LEARNING METHODS FOR UNDERSTANDING BIODIVERSITY AND ITS CONSERVATION

**Endri Xhina[1], Inva Bilo[2], Ana Ktona[1*], Anila Paparisto[3], Orion Liçi[1], Xhuliana Qirinxhi[4], Dritan Haxhiu[1]**

[1*]*Department of Informatics University of Tirana Tirana, Albania;*
[2]*Department of Mathematics Informatics and Physics University of Gjirokastra Gjirokaster, Albania;*
[3]*Department of Biology University of Tirana Tirana, Albania;*
[4]*Department of Biology -Chemistry Fan S. Noli University;*

[*]Corresponding author Ana Ktona, e-mail: ana.ktona@fshn.edu.al; endri.xhina@fshn.edu.al; ibilo@uogj.edu.al; anila.paparisto@fshn.edu.al; orion.lici@fshn.edu.al; xh.qirinxhi@yahoo.com; dritan.haxhiu@fshstudent.info;

Received June 2023; Accepted July 2023; Published August 2023;

DOI: https://doi.org/10.31407/ijees13.306

## ABSTRACT

The exponential expansion of digital data and advances in machine learning techniques present great opportunities for biodiversity research. This paper aims to explore the application of machine learning algorithms to analyze and interpret some biodiversity data pertaining to Class Insecta, specifically focusing on Order Odonata. This study seeks to demonstrate the power of machine learning in generating valuable insights that facilitate species identification and contribute to the advancement of conservation efforts. This paper encompasses the application of various machine learning algorithms to data from Class Insecta, Order Odonata, for classification, clustering and prediction tasks, and the evaluation of their performance.

**Keywords:** biodiversity, machine learning methods, species identification

## INTRODUCTION

Ecologists face a lot of challenges in conserving biodiversity. [1] Biodiversity conservation is essential for the functioning of ecosystems, the provision of vital ecosystem services, genetic resilience, cultural significance, economic benefits, ethical considerations, and scientific exploration. [2, 3] Protecting and managing biodiversity is crucial for the long-term well-being of both nature and humanity. Every species, regardless of its size or prominence, contributes to the overall functioning, stability, and resilience of ecosystems. Each species plays a role in ecological processes, trophic interactions, and genetic diversity, making them integral to biodiversity conservation efforts. The identification and classification of species are fundamental to biodiversity conservation. They provide the basis for understanding species diversity, informing conservation planning, monitoring, and assessing biodiversity, understanding ecosystem functioning, and developing legal and policy frameworks. Machine learning methods, computational algorithms designed to enable computers to learn from and make predictions or decisions based on data without being explicitly programmed, have been increasingly used in understanding biodiversity and its conservation, including identification and classification of species. [4, 5, 6, 7] This paper applies various machine learning algorithms to biodiversity data of Class Insecta, Order Odonata, and assesses their performance while comparing them to one another.

## MATERIAL AND METHOD

### Methodology and Data

Data on biodiversity is collected from two reliable and accessible online databases. Different machine learning techniques are applied to these data. The first database is The Odonate Phenotypic Database (OPDB), an online data resource for dragonfly and damselfly phenotypes (Insecta: Odonata) from all over the world. This database consists of a variety of morphological, life-history and behavioral traits, and biogeographical information collected from various sources in the literature. The database is accessible at www.odonatephenotypicdatabase.org/. OPDB consists of many attributes, but for this study were taken into consideration 15 attributes and 144 instances. Table 1 describes the attributes for a better understanding of the OPDB dataset.

Table 1. Description of OPDB database

| No. | Atribute | Description |
|---|---|---|
| 1 | GenusSpecies | Taxonomy of record. Names are taken from the World Odonata List |
| 2 | Genus | Taxonomy of record. Names are taken from the World Odonata List |
| 3 | Species | Taxonomy of record. Names are taken from the World Odonata List |
| 4 | Family | Taxonomy of record. Names are taken from the World Odonata List |
| 5 | SubOrder | Taxonomy of record. Names are taken from the World Odonata List |
| 6 | body_colors | Visual perception of a body color (black, blue, brown, green, orange, red, yellow) |
| 7 | body_colortypes | How the color is produced physically (pigment, structural, pruinescence) |
| 8 | body_patterns | Patterns found on any part of the body (plain, striped, spotted, or a combination of above) |
| 9 | continents | What continents are populations known to exist (Afrika, Asia, Australia, Europe, north America, south America) |
| 10 | aquatic_habitats | The type of aquatic habitat that the adults primarily exist in and around (ephemeral, lake, pond, river, stream, wetland) |
| 11 | body_lengths | Total length (mm) |
| 12 | hindwing_lengths | Length of the hind wing (mm) |
| 13 | wing_pigment_symmetry | Do front and hind wings have color to an equal extent (symmetric pigment, hindwing more pigment, forewing more pigment) |
| 14 | wing_pigment_color | The colors present in the wing (brown, red, yellow, black, orange, iridescent, amber, prunescence, blue or multiple) |
| 15 | wing_pigment_placement | Where the pigment or color is located (basal, middle, tip, full, or multiple) |

The other database is a database dedicated to Dragonfly in some European countries (at the moment only Central European countries). This can be found at www.dragonfly-database.eu/adults. To study the Odonate Europe dataset, there are 13 attributes and 76 instances that were included in the implementation of machine learning algorithms. Table 2 describes the attributes for a better understanding of this dataset.

Table 2. Description of Odonate Europe Database

| No. | Atribute | Description |
|---|---|---|
| 1 | Suborder | Taxonomy of record |
| 2 | Family | Taxonomy of record |
| 3 | Species | Taxonomy of record |
| 4 | Abdomen length –Max | Maximal adult body length in milimeters |
| 5 | Abdomen length – Min | Minimal adult body length in milimeters |
| 6 | Body size – Max | Maximal adult body size in milimeters |
| 7 | Body size – Min | Minimal adult body size in milimeters |
| 8 | Coloration | Coloration of body (metallic green/blue, azure blue, red, dark, yellow/dark) |
| 9 | Wingspan – Max | Maximal wingspan in milimeters |
| 10 | Wingspan – Min | Minimal wingspan in milimeters |
| 11 | Flight period – Start | The start of the flight period in months |
| 12 | Flight period – End | The end of the flight period in months |
| 13 | Flight period duration | The flight period duration in months (numeric) |

Before applying machine learning algorithms, first we performed removing or correcting incomplete/inconsistent data, as well as formatting the data into a usable format for applying machine learning algorithms. After cleaning and preprocessing is complete, the data is ready for exploration and visualization. The Machine Learning process can be broken down into the following phases: Problem Definition, Gathering Data, Data Preparation & Preprocessing, Feature Extraction, Dataset Testing and Performance Evaluation as shown in Fig.1.

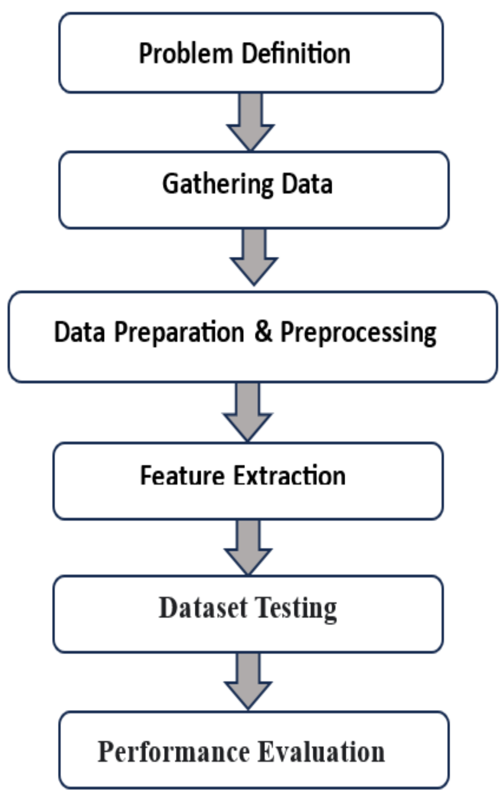


Figure 1. Steps of Machine Learning Process.

***Machine learning techniques***

***Data Visualization techniques***

To gain insight into the distribution and patterns of biodiversity, it is important to visualize data with the help of graphs or plots for a qualitative understanding of the different characteristics and hidden relationships among attributes, which is not possible by simply looking at the datasets. This is part of data pre-processing step which is the most time-consuming phase. By using visual elements, we quickly identify the trends, and this helps in data cleaning, selecting variables and data reduction. These techniques help to get an overview of the distribution of each separate attribute in the dataset and relationships between them. Figure 2 shows an example of scatter plot visualization to review the pairwise relationships between two attributes (body_lengths and hindwing_lengths).

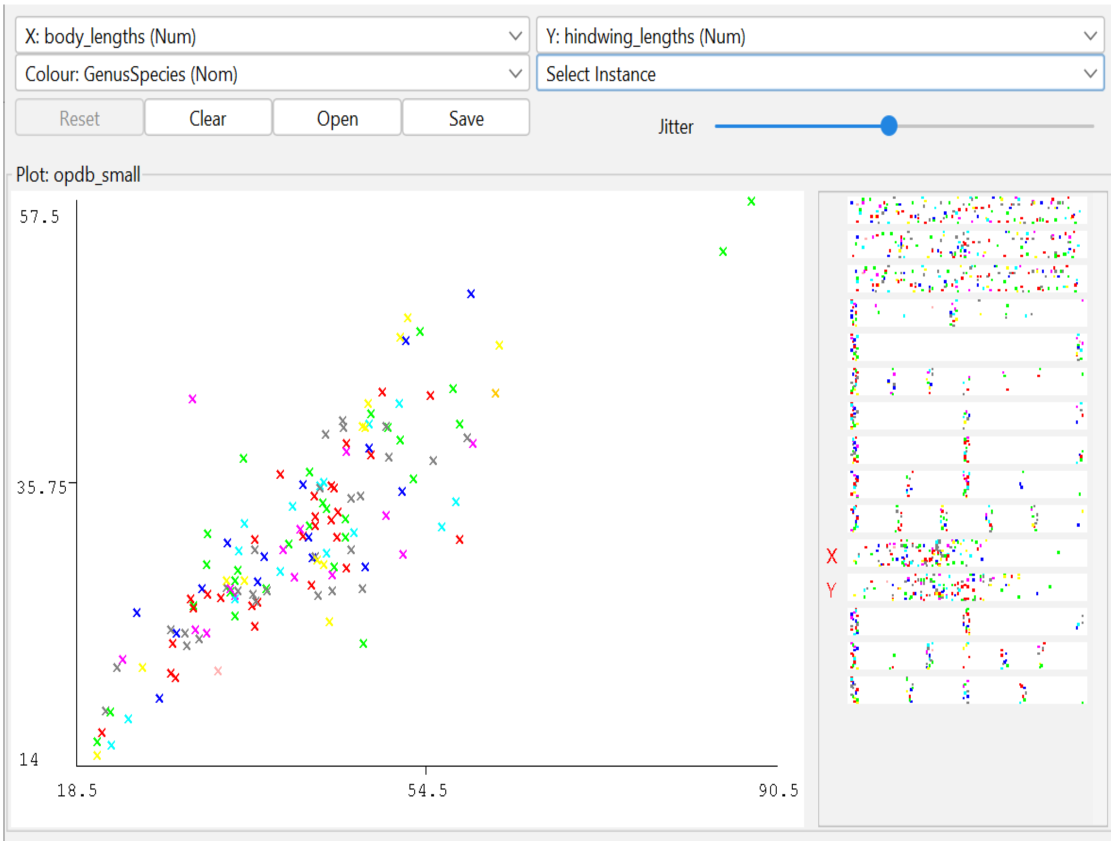


Figure 2. Relationship between body_lengths and hindwing_lengths attributes.

***Identification and classification of species***

Techniques applied to datasets mentioned above are supervised (classification) and unsupervised techniques (clustering). Classification algorithms take a known set of input data and known responses to the data and train the model to generate reasonable predictions. Algorithms we applied for performing classification in two chosen datasets, include: decision tree, random forest, ***k***-nearest neighbors, Naive Bayes, neural networks.

Clustering algorithms are based on separating data categories by similar features. Two representatives of the clustering algorithms that we performed are the K-means and the expectation maximization (EM) algorithms.

From the observations made after the implementation of all techniques, it was found that the classification algorithms have greater accuracy to be applied to these datasets. More specifically, table 3 and table 4 show the performance for the algorithms (Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error) with the highest precision that are applied at respective datasets:

Table 3. Performance of algorithms for OPDB dataset.

| | Naïve Bayes | Multilayer Perceptron | Decision Tree | Random Forest | k-nearest Neighbor |
|---|---|---|---|---|---|
| MAE | 0.0043 | 0.0009 | 0.0057 | 0.005 | 0.8841 |
| RMSE | 0.0235 | 0.0021 | 0.0533 | 0.0313 | 0.0349 |
| RAE | 5.1645% | 0.3298% | 41.2587% | 36.5584% | 13.2992% |
| RRSE | 11.9974% | 0.5875% | 64.233% | 37.6539% | 45.9897% |

Table 4. Performance of algorithms for Odoanta Europe dataset.

| | Naïve Bayes | Multilayer Perceptron | Decision Tree | Random Forest | k-nearest Neighbor |
|---|---|---|---|---|---|
| MAE | 0 | 0.0031 | 0.0745 | 0.0464 | 0.0069 |
| RMSE | 0 | 0.0039 | 0.193 | 0.0591 | 0.0415 |
| RAE | 0% | 0.6975% | 16.4989% | 10.2763% | 50% |
| RRSE | 0% | 0.8225% | 40.6749% | 12.4649% | 50% |

## CONCLUSION

This paper aims to demonstrate the potential of machine learning techniques in biodiversity understanding and its conservation. By leveraging these approaches, researchers and conservationists can benefit from improved species identification, enhanced understanding of ecological patterns, and more effective conservation strategies. Machine Learning Algorithms can contribute to the broader field of biodiversity research and pave the way for further advancements in data-driven conservation efforts.

## REFERENCES

1. Cardinale, Bradley & Duffy, J. & Gonzalez, Andrew & Hooper, David & Perrings, Charles & Venail, Patrick & Narwani, Anita & Tilman, David & Wardle, David & Kinzig, Ann & Daily, Gretchen & Loreau, Michel & Grace, James & Larigauderie, Anne & Srivastava, Diane & Naeem, Shahid. (2012). Biodiversity loss and its impact on humanity. Nature. 486. 59-67. 10.1038/nature11148;
2. Forest Isbell, David Tilman, Stephen Polasky, Michel Loreau The biodiversity-dependent ecosystem service debt. Ecol Lett. 2015 Feb;18(2):119-34. doi: 10.1111/ele.12393. Epub 2014 Nov 28. PMID: 25430966;
3. Perring, Michael & Bernhardt-Römermann, Markus & Baeten, Lander & Midolo, Gabriele & Blondeel, Haben & Depauw, Leen & Landuyt, Dries & Maes, Sybryn & De Lombaerde, Emiel & Carón, Maria & Vellend, Mark & Brunet, Jörg & Chudomelová, Markéta & Decocq, Guillaume & Diekmann, Martin & Dirnböck, Thomas & Doerfler, Inken & Durak, Tomasz & Frenne, Pieter & Verheyen, Kris. (2018). Global environmental change effects on plant community composition trajectories depend upon management legacies. Global change biology. 24. 1722-1740. 10.1111/gcb.14030;
4. Lopes, Marcelo & Gioppo, Lucas & Higushi, Thiago & Kaestner, Celso & Silla, Carlos & Koerich, Alessandro. (2011). Automatic Bird Species Identification for Large Number of Species. 117-122. 10.1109/ISM.2011.27;
5. Yuren Sun, Tatiana Midori Maeda, Claudia Solís-Lemus, Daniel Pimentel-Alarcón, Zuzana Buřivalová, (2022) Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation, Ecological Indicators, Volume 145, 2022, 109621, ISSN 1470-160X, https://doi.org/10.1016/j.ecolind.2022.109621;

6. Bhatt, Parth, Ann Maclean, Yvette Dickinson, and Chandan Kumar. 2022. "Fine-Scale Mapping of Natural Ecological Communities Using Machine Learning Approaches" *Remote Sensing* 14, no. 3: 563. https://doi.org/10.3390/rs14030563;
7. Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R. Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W. Mathis, Frank van Langevelde, Tilo Burghardt, Roland Kays, Holger Klinck, Martin Wikelski, Iain D. Couzin, Grant van Horn, Margaret C. Crofoot, Charles V. Stewart &Tanya Berger-Wolf. 2022. Perspectives in machine learning for wildlife conservation. Nat Commun 13, 792. https://doi.org/10.1038/s41467-022-27980-y;