Computational Biodiversity

Moumita Ghosh[1], Anirban Roy[2] *, and Kartick Chandra Mondal[1]

[1] Department of Information Technology
Jadavpur University
Kolkata, India
moumita4989@gmail.com
kartickjgec@gmail.com
[2] West Bengal Biodiversity Board, Govt. of West Bengal, India
dr.anirbanroy@yahoo.co.in

Abstract. The term computational biodiversity is a new phrase. Since the last few decades, biodiversity is declining globally and its shrinking rate poses threat to many species. Therefore, there is an evolving need to recognize and evaluate complex ecological problems. Along with the statistical measures securing by the ecologists, computer-science researchers have perceived the prospect of algorithmic solutions in coming up against the adverse environmental issues. Thus the application of various computational methodologies in biodiversity may coin the term computational biodiversity. In this paper, we perform a comprehensive study on recent progress made towards the protection of biodiversity and thus highlight the importance of the collaboration between ecologists and computer-science researchers. We found that the recent computational approaches have broadened the data-driven modeling capability where algorithmic developments can extrapolate the behavior of the environmental variables and find relationships among them. Therefore, we may conclude that the computational approaches can infer holistic solutions towards ecological resilience.

Keywords: Computational Biodiversity, Ecological Preservation, Machine Learning, Data Mining, Computational Approach

1 Introduction

Background: Biodiversity describes the whole range of the different varieties of living organisms. It is the single most important factor behind the equilibrium of the earth. Preservation of biodiversity is needed in order to maintain a stable ecosystem that consists of a biological community of living organisms and their nonliving components, together in a balanced form. Conservation of biodiversity offers a healthy, nutritious, and diverse ecosystem, feasible populations of species, genetic wealth, and sustainable use of biological resources. Presently, maintaining biodiversity is a crucial challenge, as it is difficult to have proper monitoring of different biological components, their changes over time including

* Corresponding author

driving factors. Data science can meet this necessity affording a lot of computational approaches. Different computational approaches that we consider here are artificial intelligence (AI), machine learning(ML), data mining(DM), deep learning(DL), and statistical measures. All these domains are interrelated. AI is fairly recognizable from the rest. It behaves like an intelligent agent by controlling/ monitoring a circumstance. AI forms a superset for ML and DL. AI utilizes the models built by ML. A more powerful version of ML is DL. Both AI and ML use DM techniques and other learning algorithms for model development. DM exploits statistics for finding patterns and phenomena. Figure 1 illustrates their relationships.

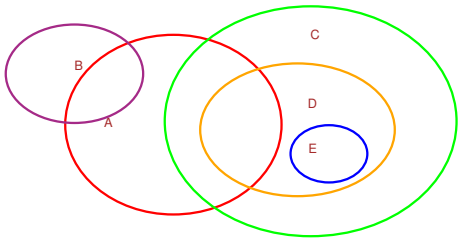

Fig. 1: Illustration using venn diagram : Regions A,B,C,D,E representing the domain of data mining, statistics, artificial intelligence, machine learning, deep learning respectively

The motivation behind the study on computational biodiversity: A report on the case study [19] of the Ganges river basin has shown that more than 10% of the World's human population depends on the Ganges river basin. Despite this, the loss of biodiversity of the river is at an increasing rate. Some fish species like Hilsa (Tenuous is), Tiger prawns (Macro brachium Rosenberg), etc. are at the extinction edge. Having assessed the total economic worth of benefits, achieved by river Halda, Bangladesh, the authors recommend the urgent need to identify the endangered fish habitat, particular reasons for their steady disappearance, suitable environment for their growth, and conservation strategies [9]. In respect of the forest ecosystem, a case study [17] of Odisha, states that Odisha has covered 7.1% of the total forest of India (FSI, 2011), and it is severely affected as it loses its green cover gradually. Forest canopy closure data, fragmentation pattern, forest fire distribution, and impact of biological invasions help in measuring the degradation of the forest ecosystem. Conservation of Western Ghats freshwater biodiversity (A case study on decapod crustaceans) is studied in [15] where the distribution pattern analysis of freshwater invertebrates in the Western Ghats helps to find out the prior areas for conservation.

Biodiversity measure like socio-cultural measure basically is aiming at surveys and interviews for the ecosystem and environment management, economic evaluation quantifies the impact of biodiversity loss in monetary terms, ecological indicators are generally used to measure the species richness and identify the endangered species. Though all of these help in measuring biodiversity from different perspectives, significant studies are still in need that use computational techniques.

Summarizing the discussion, we can say that the use of computational approaches in the biodiversity domain is highly appreciable and is in demand as it has the capability of treating heterogeneous and voluminous data. Also, it can

handle a scalable dataset. Algorithm-based approaches give accurate, reliable prediction and the capability to handle a huge amount of data.
Methods and Objective: The aim of this research is to present a synopsis of the state-of-the-art of computational approaches attempted in this field. We present a hierarchical classification of different computational approaches that have used in this domain. We identify multiple research groups and their contributions. Both qualitative analysis (based upon the problem and objective) and quantitative analysis (statistical study) of the collected articles are performed here. Our intention is to assist the researchers in gaining insight into the different computational approaches and their associated techniques to come up with new studies in biodiversity research for conservation purpose.
Organization of the paper: The whole paper is consisting of the following sections. In section 2, we have delineated a brief overview of different computational approaches used in biodiversity study and their respective applications in this field. Different research groups are enumerated based upon their followed computational methods in section 3. Analysis of the present scenario and the trends in biodiversity research work are described in section 4. Next, section 5 highlights a few existing biodiversity information systems. Section 6 concludes our study revealing the future scope.

2 Computational approaches in biodiversity

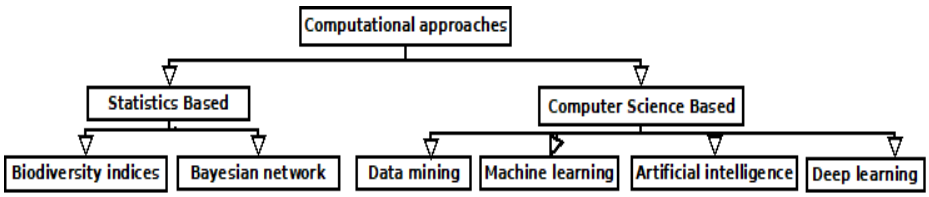

Fig. 2: Different computational approaches

Different computational approaches followed in the biodiversity study are briefly explained in this section (Table 1). Figure 2 segregates the methodologies that fall under different approaches.

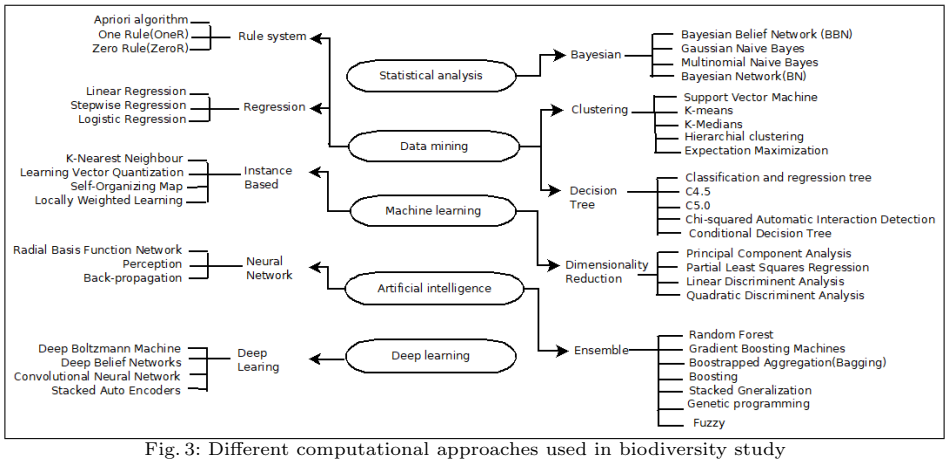

Fig. 3: Different computational approaches used in biodiversity study

We have categorized the commonly applied computational methods broadly in two categories - based on computer science and based on statistical analysis. Our main focus, here, is to analyze the application of different algorithms developed in computer science, specifically in the field of machine learning, data mining, artificial intelligence, and deep learning. In Figure 3, we have summarized the details of different approaches mainly used in the research on biodiversity.

Table 1: Multiple approaches and their application in biodiversity data analysis

Statistical approaches	
Biodiversity indices	Shannon index, Simpson's diversity index have used for geographic distribution of micro-invertebrates species, tree species diversity analysis [5]
Bayesian Belief Network	Maximize the native fish outcome study of invasive alien species in the aquatic region [3], study on rare species in the forest region
Data Mining: Automated prediction and decision-making capability	
Decision tree	Classification tree, regression tree are used for classify deciduous vegetation, spatial modeling of tree diversity [1], species presence/ absence analysis, flow regime in ecology
Clustering	Support vector machine (SVM), k-means and hierarchical clustering are most well-known for species presence-absence analysis [14], analyzing the effect of environmental factor, study on alien invasive species
Rule system-based approach	[21] uses the database of ichthyoplankton and investigating relationship present between the biotic and abiotic factors
Machine learning: Performing a specific task by the machine itself	
Instance-based	Self-organizing map (SOM), K-Nearest Neighbor (KNN), Discriminant Analysis (DA). [2] are used species presence/ absence or distribution analysis
Principal Component Analysis	Used in high-dimensional data and reduces it in the simplest form for easier analysis[4]
Artificial intelligence: Enables a system to perform tasks like an intelligent agent	
Ensemble learning	Genetic programming is highly used in species diversity analysis, species distribution modeling [7]
Artificial neural networks	Perception and radial basis network (multilayered-perception) are used for species classification and prediction problems in [10]
Deep learning: Advanced neural network architecture with multiple hidden layers	
Deep convolution neural networks	Automatic semantic extraction of features from a massive volume of a large set of image[20], species classification from the data available in the form of image, audio or video. Tree species classification from remote sensing data [11]

3 Research groups and their efforts

Table 2: A few examples for contributions made towards biodiversity data analysis

Author	**Methodology followed**	**Contribution**	**Data Used**
Edwin E Herrick and group	Machine learning and data mining	Water quality, resource management and environmental impact assessment on ecology [22]	Fish biodiversity
Peter L. M. Goethals and group	Artificial intelligence algorithms	Effect of micro-invertebrate on water quality in their presence and richness [7], establishing habitat suitability model	Micro-invertebrate datasets
Uttam K. Sarkar, and the group	Statistical analysis	Physiological biodiversity: length, weight, habitat, age, structure, growth pattern, etc. and fish stock identification, [18]	Fish species
Urs G. Kormann, and the group	Hierarchical modelling and statistical techniques	BIOFRAG database [13]: Specifically used for storing the complex results of biodiversity data in forest fragmentation studies, analyzing fragmented habitats	Forest biodiversity data
Nicolas Pasquier and group	Data mining	BioKET: [8] Biodiversity data warehouse, environmental effect on biodiversity	Plant species
Falk Huettmann and group	Data mining and machine learning	Use GMBA[3] portal for analyzing the Himalayan plant database [12], species habitat modeling in Alaska, species distribution model	Ecological wildlife

This section highlights different groups of researchers working in the areas of various computational methodologies which are mentioned in section 2. A few research groups (Table 2) are mentioned here, primarily working with biodiversity data of different ecosystems:

4 Analysis of present scenario

In this paper, we have reviewed different computational approaches that have used in the field of biodiversity for the purpose of accounting the need and present status of conservation activities. The research articles for performing the review are retrieved from google scholar: a freely accessible web search engine. We obtain the full texts of the scholarly literature on the basis of the keywords given till October 2020. Aiming at analyzing the trend of computational biodiversity research, we categorize the collected articles based on both the approaches and applications. We have gathered and separated the research articles broadly into three categories (aquatic, forest, mountain) for addressing the computational approaches to biodiversity field. Below, a brief discussion is made related to our findings.

4.1 A comprehensive analysis on computational biodiversity approaches and applications

Figure 4 is depicting different algorithmic approaches those have taken for applying in different directions. We classify the different research articles according to their main research interest. We find nine such directions like Rare species identification, Invasive alien species, Species Distribution model, Species diversity analysis, Species presence-absence analysis, Species classification, Effect of environmental factors, Flow regime for ecology, Effect of human intervention. Figure 4 highlights the commonly used methods used in meeting different solutions. For example, it can be seen that the researchers have generally used linear regression and other statistical measures to identify rare species.

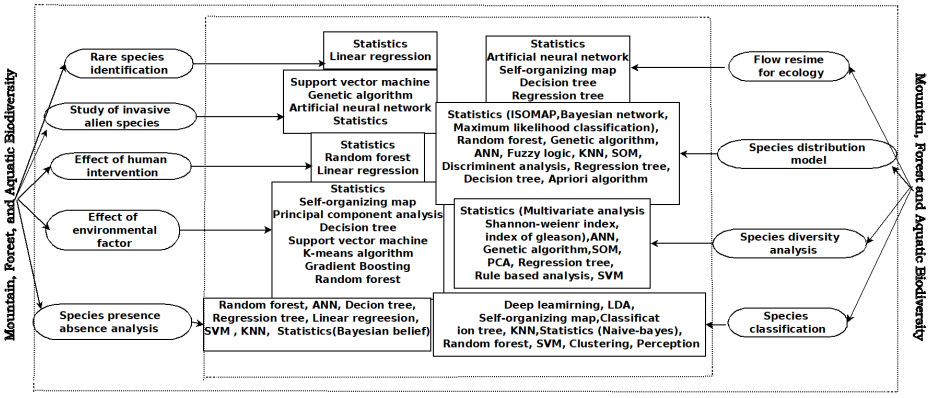

Fig. 4: Listing of different algorithms used in different applications

4.2 Computational biodiversity approaches in different domain

The number of proposed works in different domains (i.e., aquatic, forest, mountain) using different computational approaches viz, artificial intelligence, machine learning, data mining, deep learning, and statistical analysis - is depicted by a graph in Figure 5. This figure shows that most of the research attempts to focus on aquatic biodiversity compared to forest and mountain biodiversity.

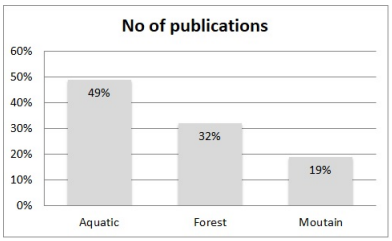

Fig. 5: Percentage of publications in different domains

Figure 6 is showing the rate of use of different algorithms in aquatic biodiversity, mountain biodiversity, and forest biodiversity. In all cases, statistics and data mining are the leading methods followed by most of the researchers. The overall use of different algorithms in all domains is depicting in Figure 7.

4.3 Analysis based on different applications

From this study, we visualize the trend and use of different methodologies in percentage which is depicted in Figure 8(left side). Here, we have shown the employment of different approaches, i.e AI, ML, DL, DM, and Statistics in percentage measure.

The figure says that the use of statistical measure is high enough compared to the other computational approaches. All approaches except deep learning are prevailing for the last two decades. Thus, though deep learning is showing a minimum number of applications, it is a promising approach for biodiversity study. Figure 8(right side) is presenting focused applications by different researchers. In most of cases, finding species diversity in a region and its distribution model, are of main concern. This trend of research is presenting the fact that the effect of human intervention and environmental factors, rare species identification, and the effect of invasive alien species require more research attention.

5 Development towards biodiversity information system and future scopes

Due to voluminous and heterogeneous biodiversity data, it is quite challenging to maintain a unique database that will keep records of all species closed to a particular domain or region. A few initiatives by the government and others as well have already made (Table 3).

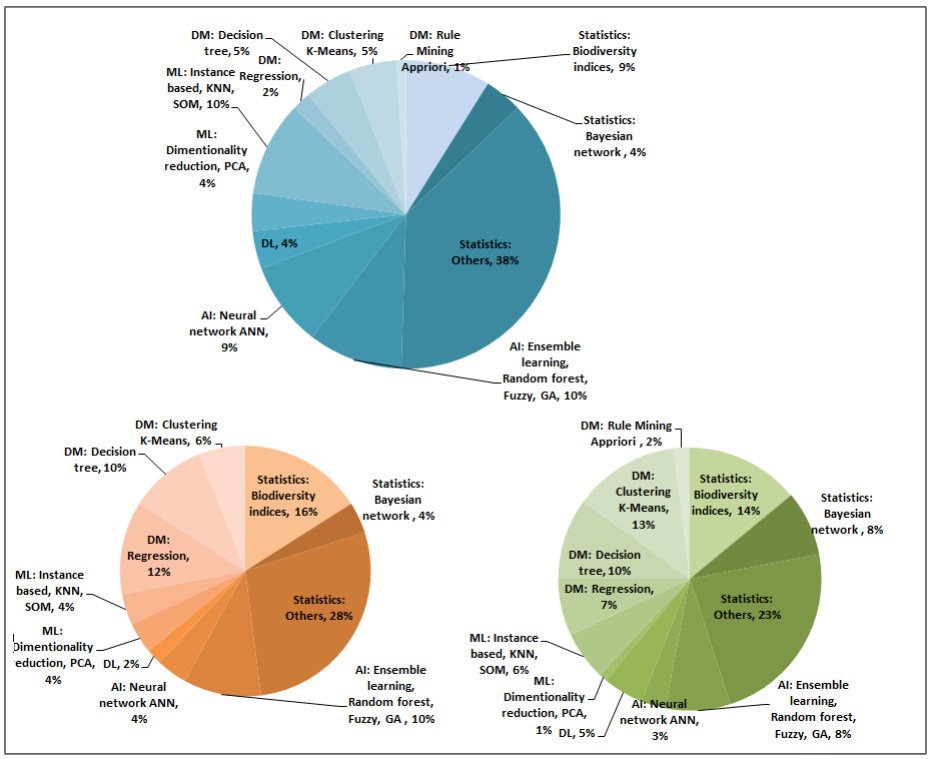

Fig. 6: Comparative use of different algorithms in aquatic biodiversity (Top), mountain biodiversity(Bottom Left) and forest biodiversity (Bottom Right)

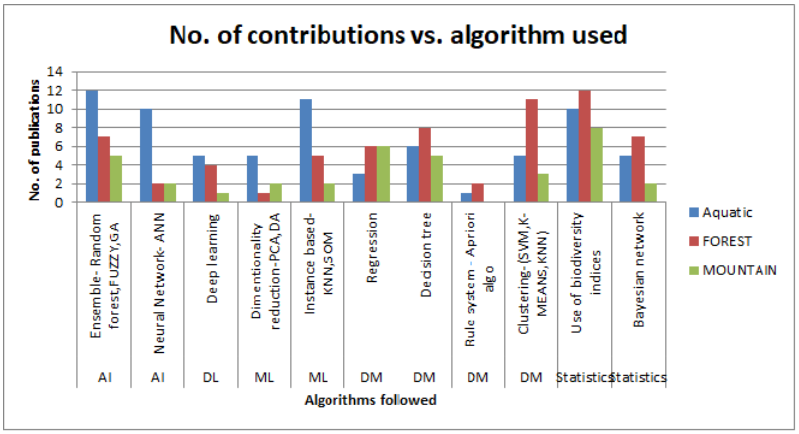

Fig. 7: Percentage of the use of the algorithms

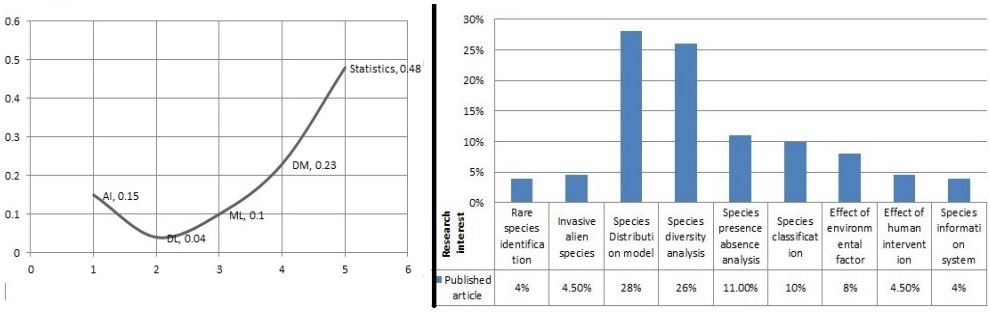

Fig. 8: Use of different approaches showing in percentage(left side); Research interest showing in percentage(Right side)

Table 3: Digitalized biodiversity: Multiple initiatives for portal development

Portal	**Contribution**	**Focused data**
GFBI [4]	Managing world's forest inventory database, policy making for forest science, platforming forest study and research	Tree level forest inventory data and services
GBIF [5]	Global network for providing research infrastructure and openly accessible biodiversity data.	Data on all type of live on earth.
Fishbase [6]	Global database for species. It is an analytical and graphical tool for identifying, managing and restoring depleted fish stock	Fish species data
iDigBio [7]	Aiming at digitization of biodiversity collection	Specimen data
BioGeo-Mancer [8]	Maximize the quality and quantity of biodiversity data by integrating it with geospatial data. Thus support planning, conservation and management in biodiversity data	Biodiversity data along with geographical location

Besides the highlighted portals (Table 3), few more studies are there, e.g. [16] has worked upon Western Ghat ecology, the biodiversity information system [6] on rare species is built on the European dataset. There may have scope for generating a unique interactive database having information regarding synonym, common name, habitat, ecological status, distribution, habit, identification, description of different species of a particular region along the time dimension. It may find the reason for extinction or migration, etc. It also may help in knowing where did some specific species arise? How many of them survive? How are they spreading? Such kind of biodiversity database has a high prospect in making the decision for the conservation purpose only when computational techniques will be incorporated at the back end to deal with the voluminous data.

6 Conclusion and Future work

This paper provides a brief review of the computational approaches attempted in the biodiversity domain which is unavailable in the state-of-the-art. It would be helpful for the research community as a brief integrated scenario on approach and application is emphasized here. It has been noticed that most of the ecologists use statistical analytical tools where hypothesis tests have been performed in

[4] Global Forest Biodiversity Initiative: https://www.gfbinitiative.org
[5] Global Biodiversity Information Facility: https://www.gbif.org/
[6] https://www.worldfishcenter.org/fishbase
[7] Integrated Digitized Biocollections:https://www.idigbio.org/portal
[8] BG: https://sites.google.com/site/biogeomancerworkbench/

finding relationships among the predictor and response variables. But, with the help of computer-science based methodologies, exploratory analysis is possible instead of confirmatory analysis. Thus, exploring the data helps in building more accurate models in order to assist in future research.

In the future, system modeling could be attempted using a computational framework for building holistic solutions for complex environmental and ecological issues, even incorporating big data. Henceforth, automation in a built-in model would assist the ecologists to find feasible solutions with minimum human intervention. Above all, our study would motivate the interdisciplinary research in this domain and help to identify the major research scope in preventing biodiversity resilience.

Acknowledgements

The authors are grateful to the Department of Science & Technology, Government of India, New Delhi, for financial assistance under the scheme of WOS-A (Women Scientist Scheme A) to carry out this Ph.D. research project.

References

1. A. Abdollahnejad, D. Panagiotidis, P. Surov, et al. Investigation of a possibility of spatial modelling of tree diversity using environmental and data mining algorithms. *J. FOR. SCI*, 62(12):562–570, 2016.
2. M. A. Acevedo, C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206–214, 2009.
3. P. Boets, D. Landuyt, G. Everaert, S. Broekx, and P. Goethals. Evaluation and comparison of data-driven and knowledge-supported bayesian belief networks to assess the habitat suitability for alien macroinvertebrates. *Environmental Modelling & Software*, 74:92–103, 2015.
4. Francisco Ronaldo Alves de Oliveira, Carlos Tadeu dos Santos Dias, Henrique Antunes de Souza, Breno Leonan de Carvalho Lima, and Mírian Cristina Gomes Costa. Tree legumes with fertilizer potential: a multivariate approach. *Revista Ciência Agronômica*, 52(1):1–10, 2021.
5. M. K. Gautam, R. K. Manhas, and A. K. Tripathi. Patterns of diversity and regeneration in unmanaged moist deciduous forests in response to disturbance in shiwalik himalayas, india. *Journal of Asia-Pacific Biodiversity*, 9(2):144–151, 2016.
6. P. Genovesi, L. Carnevali, A. Alonzi, and R. Scalera. Alien mammals in europe: updated numbers and trends, and assessment of the effects on biodiversity. *Integrative zoology*, 7(3):247–253, 2012.
7. S. Gobeyn, M. Volk, L. Dominguez-Granda, and P. L. Goethals. Input variable selection with a simple genetic algorithm for conceptual species distribution models: A case study of river pollution in ecuador. *Environmental Modelling & Software*, 92:269–316, 2017.
8. S. Inthasone, N. Pasquier, et al. The bioket biodiversity data warehouse: Data and knowledge integration and extraction. In *International Symposium on Intelligent Data Analysis*, pages 131–142. Springer, 2014.

9. H. Kabir, M. Kibria, M. Jashimuddin, and M. M. Hossain. Conservation of a river for biodiversity and ecosystem services: the case of the halda–the unique river of chittagong, bangladesh. *International Journal of River Basin Management*, 13(3):333–342, 2015.
10. K. López-de Ipiña, M. Iturrate, J. B. Alonso, and B. Rodríguez-Herrera. Automatic acoustic analysis for biodiversity preservation: A multi-environmental approach. In *Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on*, pages 43–48. IEEE, 2015.
11. Janne Mäyrä, Sarita Keski-Saari, Sonja Kivinen, Topi Tanhuanpää, Pekka Hurskainen, Peter Kullberg, Laura Poikolainen, Arto Viinikka, Sakari Tuominen, Timo Kumpula, et al. Tree species classification from airborne hyperspectral and lidar data using 3d convolutional neural networks. *Remote Sensing of Environment*, 256:112322, 2021.
12. D. Nemitz, F. Huettmann, E. M. Spehn, and W. B. Dickoré. Mining the himalayan uplands plant database for a conservation baseline using the public gmba webportal. In *Protection of the Three Poles*, pages 135–158. Springer, 2012.
13. M. Pfeifer, V. Lefebvre, T. A. Gardner, V. Arroyo-Rodriguez, L. Baeten, C. Banks-Leite, J. Barlow, M. G. Betts, J. Brunet, A. Cerezo, et al. Biofrag–a new database for analyzing bio diversity responses to forest frag mentation. *Ecology and Evolution*, 4(9):1524–1537, 2014.
14. R. Pouteau, J. Meyer, R. Taputuarai, and B. Stoll. Support vector machines to map rare and endangered native plants in pacific islands forests. *Ecological Informatics*, 9:37–46, 2012.
15. R. Raghavan, S. Dahanukar, N.and Philip, P. Iyer, B. Kumar, B. Daniel, and S. Molur. The conservation status of decapod crustaceans in the western ghats of india: an exceptional region of freshwater biodiversity. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 25(2):259–275, 2015.
16. T. V. Ramachandra and A. Suja. Sahyadri: Western qhats biodiversity information system http://ces. use, erne t. in/biodiversity. *Biodiversity in Indian Scenarios*, page 1, 2006.
17. C Sudhakar Reddy, Anzar A Khuroo, P Hari Krishna, KRL Saranya, CS Jha, and VK Dadhwal. Threat evaluation for biodiversity conservation of forest ecosystems using geospatial techniques: a case study of odisha, india. *Ecological engineering*, 69:287–303, 2014.
18. Kavitha Mandhir Sandhya, Lianthuamluaia Lianthuamluaia, Gunjan Karnatak, Uttam Kumar Sarkar, Suman Kumari, Puthiyottil Mishal, Vikash Kumar, Debabrata Panda, Yousuf Ali, and Bablu Kumar Naskar. Fish assemblage structure and spatial gradients of diversity in a large tropical reservoir, panchet in the ganges basin, india. *Environmental Science and Pollution Research*, 26(18):18804–18813, 2019.
19. R. S. Shrivastava. Biodiversity of river ganga (india): An environmental economist perspective. *International Institute of Fisheries Economics and Trade*, 2008.
20. S. A. Siddiqui, A. Salman, M. I. Malik, F. Shafait, A. Mian, M. R. Shortis, and E. S. Harvey. Deep learning for microalgae classification. *IEEE*, 2017.
21. M. Silva, D. Q. Trevisan, D. N. Prata, E. E. Marques, M. Lisboa, and M. Prata. Exploring an ichthyoplankton database from a freshwater reservoir in legal amazon. In *International Conference on Advanced Data Mining and Applications*, pages 384–395. Springer, 2013.
22. W. P. Tsai, F. J. Chang, and E. E. Herricks. Exploring the ecological response of fish to flow regime by soft computing techniques. *Ecological Engineering*, 87:9–19, 2016.