Prediction on Bird Diversity from Datasets of Rural and Industrial Area of Asansol, West Bengal: A Machine Learning Approach

Debdyuti Sengupta[1], Soumendra Nath Talapatra[2]

[1]PhD Scholar, Department of Environmental Science, Seacom Skills University, Kendradangal, Shantiniketan, Birbhum – 731236, West Bengal, India
Corresponding Author Email: *debdyuti[at]bccollegeasansol.ac.in*

[2]School of Life Sciences, Seacom Skills University, Kendradangal, Shantiniketan, Birbhum – 731236, West Bengal, India

Abstract: ***The objective of the present study was to predict the accuracy of datasets of bird diversity in rural and industrial area of Asansol, West Bengal during monsoon season. This study was performed through machine learning (ML) algorithms such as BayesNet (BN), NaiveBayes (NB), logistic regression (LR) and Random Tree (RT) by using WEKA tool, version 3.8.6. The study was separately based on 3 attributes such as Order, Numbers, and Effects (High, Moderate and Low) to know overall prediction accuracy of dataset as per 10-fold cross validation (CV) test for each algorithm. In the present study, the weighted average of precision recall curve (PRC) values obtained 100.0% for rural and industrial area on these algorithms. It is concluded that these ML algorithms especially BN and RT predicted accurately from the dataset and obtained rich information with statistical interpretation, which confirmed lower diversity of avifauna.***

Keywords: Asansol, Bird diversity, Machine learning algorithms, Industrial area, Prediction accuracy of dataset, Rural area

1. Introduction

The definition of biodiversity is the variety of life in each ecosystem or ecological complex where the living organisms are inhabiting. [1] Besides, the diversity of several species the study of bird diversity is very important because this indicates air quality of particular area. An international study reported that decreasing of bird population in USA due to air pollutant especially ozone pollution. [2] The emissions from industries and automobiles cause air pollution. In this context, Bhola et al. proposed area-based conservation, which provides the key perspectives to reach post-2020 global biodiversity and sustainability targets. [3]

Generally, the bird species are categorized into different classes as per the physical characteristics, colour, and shape of species. Avifauna inhabiting in any environments and occupy nearly every niche, which are widely used as monitoring targets. [4]

Many studies reported that diversity of avifauna could be achieved through machine learning (ML) modelling by Nanni et al. [5] studied a combination of classifiers as per several investigators, which included as AlexNet, [6] GoogleNet, [7] VGGNet, [8] ResNet and InceptionV3 for the identification of the bird species by processing its audio signal. The audio images such as spectrograms, ScatNet. [9.10] Gawali et al. evaluated and observed the result of "Deep Learning based artificial intelligence (AI) Model", which is useful for identifying birds using their images. [11] Chandra et al. studied ML algorithm especially support vector machine (SVM) for classifying bird species from images. [11] Gavali & Banu explored on assessing the health of bird species in India as they respond to climate-related challenges, which are employing advanced image analysis techniques. They developed and trained a "deep convolution neural network (DCNN)" for classifying health indicators around a range of species. [12] Gavali & Banu also used traditional ML algorithms such as "deep convolutional neural network (DCNN)" and "short-term memory (LSTM) network" after extracting features of bird images, manual feature extraction and training model formed for an automated bird classification system. [13]

This study was attempted to predict the accuracy through machine learning algorithms for bird diversity datasets from rural and industrial area of Asansol, West Bengal.

2. Materials and Methods

2.1 Study area

The study site was selected as per rural area comparatively lower vehicular movements, far from industrial vicinity as A3 site (Latitude = 23° 83′ N and Longitude = 87° 00′ E) and industrial area as A1 site (Latitude = 23° 41′ N and Longitude = 86° 57′ E), Asansol, West Bengal.

2.2 Dataset for the study

As per the earlier study by Sengupta & Talapatra [15] and in recent study, the datasets were prepared on classifier accuracy through specific ML algorithms. The dataset was order, total number of species and effects as high, moderate and low in terms of numbers in which the values were ≥25, 11-24 and <10 nos., respectively of species observed.

2.3 ML modelling and statistical data interpretation

In this study, the data mining through WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.6) was used, which developed by Frank et al. [16] where prediction accuracy of dataset through ML algorithms easily obtained. The WEKA explorer was developed by Witten et al. in which pre-processing for all the data were made through unsupervised instance and 10-fold cross validation (CV) test. [17]

The predictive accuracy of dataset (bird diversity) on above-mentioned attributes through ML algorithms especially BayesNet (BN), NaiveBayes (NB), logistic regression (LR) and Random Tree (RT) studied separately from dataset to predict the overall prediction accuracy as per earlier study by Talapatra et al. [18]

The prediction accuracy of above-mentioned ML model classifications related to correctly and incorrectly classified instances, Kappa statistics (KS), mean absolute error (MAE) and root mean squared error (RMSE) were studied for 10-fold CV test. As per Bouckaert et al., [19] the results for each algorithm model summary were retrieved. The prediction accuracy of studied ML models was retrieved from summary results and the statistical methods such as F-value, Matthew's correlation coefficient (MCC), receiver operating characteristic (ROC) curve and Precision-recall curve (PRC) area, respectively were obtained.

3. Results

In Table 1, the dataset of bird diversity in rural area of Asansol was prepared for the prediction accuracy in which the bird species under order Columbiformes and Passeriformes were found in higher values in A3 while only Columbiformes was obtained in higher values in A1 sites.

Table 1: Dataset for bird diversity in rural and industrial area of Asansol

A3 site			A1 site		
Order	Total Numbers	Class	Order	Total Numbers	Class
Columbiformes	60	High	Passeriformes	15	Moderate
Columbiformes	18	Moderate	Columbiformes	42	High
Passeriformes	12	Moderate	Passeriformes	12	Moderate
Passeriformes	8	Low	Passeriformes	4	Low
Cuculiformes	1	Low	Columbiformes	2	Low
Accipitriformes	2	Low	Passeriformes	8	Low
Passeriformes	25	High	Suliformes	1	Low
Gruiformes	7	Low	Anseriformes	12	Moderate
Psittaciformes	12	Moderate	Passeriformes	2	Low
Passeriformes	2	Low	Galliformes	15	Moderate
Passeriformes	1	Low	Passeriformes	5	Low
Pelecaniformes	12	Moderate			
Passeriformes	5	Low			
Anseriformes	11	Moderate			
Galiformes	9	Low			
Passeriformes	4	Low			
Passeriformes	8	Low			
Passeriformes	2	Low			
Cuculiformes	1	Low			
Piciformes	2	Low			

The accuracy of studied dataset as per ML algorithm classifications in which correctly and incorrectly classified instances, KS, MAE and RMSE were studied by 10-fold CV test. In the case of algorithm model classification, the correctly classified instances were observed a higher value of about 95.00 for BN and RT followed by 85.00 for LR while lower value of about 80.00 for NB in the dataset. In the case of algorithm model classification, the correctly classified instances were observed a higher value of about 95.00 for BN and RT followed by 85.00 for LR while lower value of about 80.00 for NB in the dataset (Table 2).

Table 2: Results on different classified instances and statistical values for different algorithm models on bird diversity for both sites

Classifier model	Correctly classified instances	Incorrectly classified instances	KS	MAE	RMSE
A3 site					
BN	95.00	5.00	0.90	0.12	0.22
NB	80.00	20.00	0.61	0.16	0.34
LR	85.00	15.00	0.69	0.10	0.32
RT	95.00	5.00	0.90	0.03	0.18
A1 site					
BN	90.91	9.09	0.83	0.16	0.24
NB	90.91	9.09	0.82	0.07	0.20
LR	72.73	27.27	0.51	0.19	0.43
RT	90.91	9.09	0.83	0.09	0.26

BN = BayesNet; NB = NaiveBayes; LR = Logistic regression; RT = Random tree; KS = Kappa Statistics; MAE = Mean Absolute Error; RMSE = Root Mean Squared Error

Table 3 and Table 4 evaluate the detailed accuracy of studied algorithms from the dataset. To evaluate the accuracy of a classifying values for F-measure, MCC, ROC and PRC, the better performances were studied for 3 effects for bot sites. In the present study, BN and RT models were obtained 100.0% Precession recall curve value for the confirmation of lower diversity of bird species.

Table 3: Statistical data for prediction accuracy of studied algorithms on bird diversity in A3 sites

Classifier model	Effects	F-value	MCC	ROC area	PRC area
BN	High	0.667	0.688	0.750	0.591
	Moderate	0.909	0.882	0.973	0.927
	Low	1.000	1.000	1.000	1.000
NB	High	0.000	0.076	0.528	0.306
	Moderate	0.667	0.545	0.787	0.530
	Low	0.960	0.899	0.945	0.979
LR	High	0.500	0.444	0.819	0.375
	Moderate	0.667	0.577	0.813	0.588
	Low	0.963	0.892	0.989	0.995
RT	High	0.667	0.688	0.750	0.550
	Moderate	0.909	0.882	0.967	0.833
	Low	1.000	1.000	1.000	1.000

BN = BayesNet; NB = NaiveBayes; LR = Logistic regression; RT = Random tree; MCC = Matthew's correlation coefficient; ROC = Receiver operating characteristic; PRC = Precision-recall curve

Table 4: Statistical data for prediction accuracy of studied algorithms on bird diversity in A1 site

Classifier model	Effects	F-value	MCC	ROC area	PRC area
BN	High	0.000	0.000	0.700	0.250
	Moderate	0.889	0.828	0.893	0.733
	Low	1.000	1.000	1.000	1.000
NB	High	0.000	0.000	1.000	1.000
	Moderate	1.000	1.000	1.000	1.000
	Low	0.923	0.828	1.000	1.000
LR	High	0.000	0.100	0.000	0.091
	Moderate	0.750	0.607	0.821	0.646
	Low	0.833	0.633	0.833	0.906
RT	High	0.000	0.000	0.400	0.091
	Moderate	0.889	0.828	0.893	0.733
	Low	1.000	1.000	1.000	1.000

BN = BayesNet; NB = NaiveBayes; LR = Logistic regression; RT = Random tree; MCC = Matthew's correlation coefficient; ROC = Receiver operating characteristic; PRC = Precision-recall curve

4. Discussion

In the present study, BN and RT models were obtained 100.0% Precession recall curve value for the confirmation of lower diversity of bird species during monsoon season. A contradictory study by Roopha et al., [20] which was reported that during the monsoon months, the avian diversity was higher compared to pre-monsoon and post-monsoon months in Tamil Nadu. But a similar finding with the present study that Passeriformes dominated among avifaunal species richness during monsoon period. Another study fully supported that the richness as well as the diversity of birds was low in monsoon season due to the higher rainfall, which decreases the activity of birds and the nesting behaviour in 8 wetlands of Odisha. [21]

Many studies reported that diversity of avifauna could be achieved through ML modelling through suitable algorithms. [5,11] Interestingly, WEKA platform used to predict and data-mine for body weight dataset of turkeys. [22] They used some of the algorithms viz. J48 decision tree (C4.5), Random Forest, Naive Bayes, Sequential Minimal Optimization (SMO) of support vector machine and multilayered perceptron artificial neural network (MLP). Talapatra et al. evaluated machine learning (ML) algorithm models viz. BayesNet (BN), NaiveBayes (NB), logistic regression (LR), Lazy.KStar (K*), decision tree (DT) J48, Random forest (RF) and Random tree (RT) in the WEKA tool (version 3.8.5) for the prediction of the accuracy of the dataset generated from an image. [18] Still, industrial and rural area are predicted lower diversity of avifauna may be due to air quality index and downwind direction during post-monsoon season. In this study, more species were obtained in A3 site but the number of species were lower quantity while in A1 site the species types and numbers were lower, which predicted and confirmed the lower avifaunal diversity.

5. Conclusion

In the present study, the weighted average of precision recall curve (PRC) values obtained 100.0% for these algorithms like BN and RT. It is concluded that these ML algorithms especially BN and RT predicted accurately from the dataset of bird diversity and obtained rich information with statistical interpretation, which confirmed the lower diversity during post- season. The future study in WEKA tool can easily be analysed with more dataset to predict classifier accuracy related to bird diversity related to pre-monsoon seasons for big data mining.

Acknowledgment

Authors thankful to the developer for the present tool used this study.

Conflict of interest

No conflict of interest.

References

[1] Stork NE. Biodiversity. In: Encyclopedia of Insects. Resh VH, Ring T, Cardé, RT. (eds.). Chapter 21, 2nd edition, Academic Press, pp. 75-80.

[2] Liang Y, Rudika I, Zou EY, Johnston A, Rodewald AD, Klinga CL. Conservation cobenefits from air pollution regulation: Evidence from birds. Proceedings of the National Academy of Sciences USA. 2020;117(49):30900-6.

[3] Bhola N, Klimmek H, Kingston N, Burgess ND, van Soesbergen A, Corrigan C, et al. Perspectives on area-based conservation and its meaning for future biodiversity policy. Conservation Biology. 2021;35(1):168-178.

[4] Kahl S, Wood CM, Eibl M, Klinck H. BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics. 2021; 61: 101236.

[5] Nanni L, Maguolo G, Brahnam S, Paci M. An ensemble of convolutional neural networks for audio classification. Applied Science. 2021; 11:5796.

[6] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012;25(2):1097–1105.

[7] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015; pp. 1-9.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations (ICLR 2015), 2015; pp. 1-14.

[9] Sifre L, & Mallat S. Combined scattering for rotation invariant texture analysis. ESANN. 2012; 44:68-81.

[10] Bruna J, Mallat S. Invariant scattering convolution networks. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1872-1886.

[11] Gawali NV, Agavile TV, Dalvi AG, Joshi SS. Bird species image identification using transfer learning. Journal of Emerging Technologies and Innovative Research. 2022;9(2): d567-d570.

[12] Chandra B, Raja S, Gujjar R, Varunkumar J, Sudharsan A. Automated bird species recognition system based on image processing and SVM classifier. Turkish Journal of Computer and Mathematics Education. 2021; 12:351-6.

[13] Gavali P, Banu JS. Integrating health informatics and deep learning to evaluate climate change effects on Indian bird health. Frontiers in Health Informatics. 2024;13(3):3959-3975.

[14] Gavali P, Banu JS. A novel approach to Indian bird species identification: employing visual-acoustic fusion techniques for improved classification accuracy. Frontiers in Artificial Intelligence. 2025; 8:1527299.

[15] Sengupta D, Talapatra SN. Study of bird diversity during monsoon season related to air quality at Asansol, West Bengal. International Journal of Science and Research. 2023;12(10): 1887-1890.

[16] Frank E, Hall MA, Witten IH. The WEKA workbench, Online appendix for data mining: Practical machine learning tools and techniques. 4th edition, Morgan Kaufmann, 2016.

[17] Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. 3rd edition, Morgan Kaufmann, Burlington, MA, 2011.

[18] Talapatra SN, Chaudhuri R, Ghosh S. CellProfiler and WEKA tools: Image analysis for fish erythrocytes shape and machine learning model algorithm accuracy prediction of dataset. World Scientific News. 2021; 154:101-116.

[19] Bouckaert RR, Frank E, Hall M, et al. WEKA manual for version 3-8-5. University of Waikato, Hamilton, New Zealand, 2020, December 21.

[20] Roopha, P. D., Thatheyus, A. J., Sonia, T., & Kishore, R. (2022). Avifaunal diversity in the tropical thorn forest of Kiluvamalai, Madurai district, Tamil Nadu, India. *Asian Journal of Conservation Biology*. 11(2):274-280.

[21] Panda BP, Das AK, Jena SK, Mahapatra B, Dash AK, Pradhan A, et al. Habitat heterogeneity and seasonal variations influencing avian community structure in wetlands. Journal of Asia-Pacific Biodiversity. 2021;14(1):23-32.

[22] Flores KR, de Carvalho LVFM, Reading BJ, Fahrenholz A, Ferket PR, Grimes JL. Machine learning and data mining methodology to predict nominal and numeric performance body weight values using Large White male turkey datasets. Journal of Applied Poultry Research. 2023;32(4):100366.