ACCURACY PREDICTION FOR VISITING INSECT DIVERSITY ON RICE CROP: A MACHINE LEARNING APPROACH

Sukanya Bandyopadhyay[1]*, Tapan Kumar Chatterjee[1], Soumendra Nath Talapatra[2]

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

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

*Corresponding author: Dr. Sukanya Bandyopadhyay, Guest Lecturer, Ph.D.,

Abstract: *The present study was predicted the accuracy of datasets of insect diversity on mustard flower. This study was performed through machine learning (ML) especially tree algorithms by using WEKA tool, version 3.8.6. Different tree algorithm such as Naïve Bayes (NB), Decision Tree J48 (DT J48) and Logistic Model Tree (LMT) were studied separately based on 3 attributes such as Species, Numbers, and Effect class (High, Moderate and Low) to know overall prediction accuracy as per 10-fold cross validation. In the present study, NB model was obtained 83.3% on Precession recall curve value for Low effect. Moreover, ROC curve for low effect in the dataset for three algorithms in which area under curve obtained higher in LMT (93.8%) followed by DT J48 (87.5%) while lower in NB (75.0%) were recorded. It is concluded that the ML algorithms performed accurately from the dataset of visiting insect diversity on rice crop, which was based on statistical interpretations.*

Keywords: Accuracy prediction; Data mining; Insect diversity; Machine learning algorithms; WEKA tool

1. INTRODUCTION

Modern agricultural practices depend on highest yield of crops. To produce maximum yield of crops, the pest control is necessary approach in the agricultural fields [1]. Among several crops, rice crop yield was reported about 99 million tonnes in 2012-2013 in Indian context [2,3].

In addition, increasing crop yield and improving quality of the fruits that are employed through pollination by insects [4]. Some studies indicated that there are mutual relationships between insect pollinators and flowering plants [5,6].

Siregar et al. [7] reported that the diversity of inhabiting and visiting insects such as Homoptera, Hymenoptera, Coleoptera, Orthoptera, Odonata, Lepidoptera, Hemiptera and Diptera on the paddy crops.

In recent research interest, the big data mining through machine learning (ML) modelling is prioritized where the predictive information can obtain from dataset. It can easily be undertaken through ML algorithms that predict the accuracy of the dataset [8-12].

The objective was to predict the accuracy of dataset on insect diversity on rice crop through machine learning algorithms especially tree models.

2. MATERIALS AND METHODS

Firstly the dataset were prepared as per the number of species under family of insects. The effect values viz. >100 numbers (High), 50-100 numbers (Moderate) and <50 numbers (Low) were categorized. In the present study, we used data mining tool namely WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.5) developed by Frank et al. [13] in which prediction accuracy might be observed through ML modelling algorithms. The WEKA explorer was developed with data pre-processing, classification, regression, and association rules [14]. In pre-processing, all the data were made through unsupervised instance and 10-fold cross validation (CV) dataset was used.

The predictive accuracy of dataset (insect diversity) on two attributes viz. species and numbers as well as their effect as high, moderate and low diversity through ML algorithms especially tree models such as Naïve Bayes (NB), decision tree (DT) J48, and Logistic model trees (LMT), separately studied from dataset to predict the overall prediction accuracy from the dataset of our earlier study of Talapatra et al. [15].

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 earlier study by Talapatra et al. [15] and Bhattacharya et al. [16]. As per Bouckaert et al. [17], the results for each algorithm model summary were retrieved from WEKA tool. The prediction accuracy of studied ML models as per 10-fold cross validation test was retrieved from summary results and the statistical methods as per earlier study by Bandyopadhyay et al. [11].

3. RESULTS AND DISCUSSION

In our study, qualitative and quantitative assessment revealed that the variety of insect species were observed on paddy (rice crop) in the study sites (Table 1). About 12 types of insect varieties were recorded. A maximum number of specimens under the family Muscidae followed by Stratiomyidae, Bracoinidae and Corridae were recorded in the study sites.

Table 1: List of inhabiting and visiting insect species on rice crop in the study sites

Sl. No.	Scientific Name	Common name	Total No.
Araneidae			
1.	*Cyclosa spirifera*	Weaver Spider	61
Salticidae			
2.	*Carrhotus viduus*	Jumping spider	72
3.	*Phintella vittata*	Jumping spider	59
4.	*Hyllus semicupreus*	Jumping spider	49
Carabidae			
5.	*Casnoidea indica*	Ground beetle	36
Anthicidae			
6.	*Micrapsis crocea*	Lady beetle	39
Miridae			
7.	*Cyrtorhinus lividipennis*	Rice bug	37
Corridae			
8.	*Limnogonus fossarum*	Water bug	101
Muscidae			
9.	*Musca domestica*	Housefly	123
Stratiomyidae			
10.	*Hermetia illucens*	Black soldier fly	112
Bracoinidae			
11.	*Macrocentrus philippinensis*	Wasp	103
12.	*Telenomus rowani*	Wasp	71

The performance of model accuracy of studied ML algorithm classifications as per correctly and incorrectly classified instances, KS, MAE and RMSE were studied as per 10-fold CV test. In the case of algorithm model classification, the correctly classified instances were observed a higher value of about 75.00 for NB while lower value of about 58.33 for DT J48 and LMT in the dataset (Table 2).

Table 3 evaluates the detailed accuracy of studied tree models for the studied dataset. To evaluate the accuracy of a classifying values for F-measure, MCC, ROC and PRC, the better performances were studied for 3 classes. In the present study, NB model was obtained 83.3% on Precession recall curve value for Low effect.

Table 2: Results on different classified instances and statistical values for different algorithm models on insect diversity on rice crop

Classifier model	Correctly classified instances	Incorrectly classified instances	KS	MAE	RMSE
NB	75.00	25.00	0.61	0.23	0.43
DT J48	58.33	41.67	0.35	0.27	0.45
LMT	58.33	41.67	0.35	0.29	0.43

NB = Naïve Bayes; DT J48 = Pruned and unpruned decision tree C4; LMT = Logistic model trees; KS = Kappa Statistics; MAE = Mean Absolute Error; RMSE = Root Mean Squared Error

Table 3: Statistical data for prediction accuracy of studied algorithms on insect diversity on rice crop

Classifier model	Effects	F-value	MCC	ROC area	PRC area
NB	High	0.667	0.556	0.630	0.480
	Moderate	0.727	0.507	0.571	0.503
	Low	0.857	0.816	0.750	0.833
DT J48	High	0.333	0.111	0.630	0.397
	Moderate	0.545	0.169	0.643	0.683
	Low	0.857	0.816	0.875	0.833
LMT	High	0.333	0.111	0.815	0.533
	Moderate	0.545	0.169	0.543	0.578
	Low	0.857	0.816	0.938	0.917

NB = Naïve Bayes; DT J48 = Pruned and unpruned decision tree C4; LMT = Logistic model trees; MCC = Matthew's correlation coefficient; ROC = Receiver operating characteristic; PRC = Precision-recall curve

Fig 1 to 3 depicted ROC curve for low effect in the dataset for three algorithms in which area under curve obtained higher in LMT (93.8%) followed by DT J48 (87.5%) while lower in NB (75.0%).

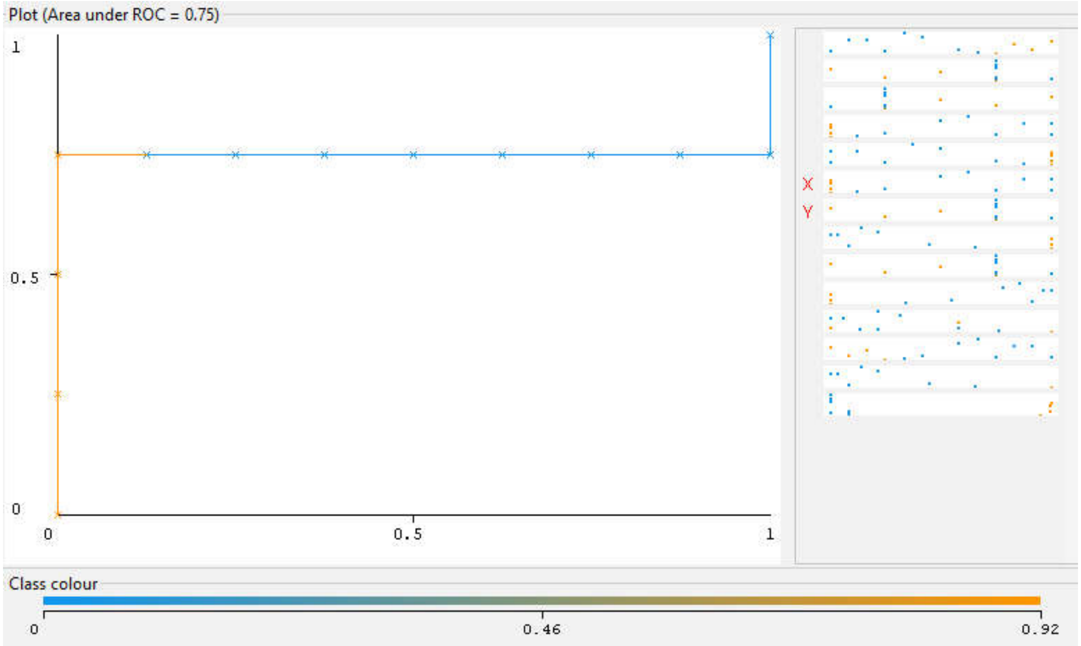

Fig 1: ROC curve for low effect on NB algorithm

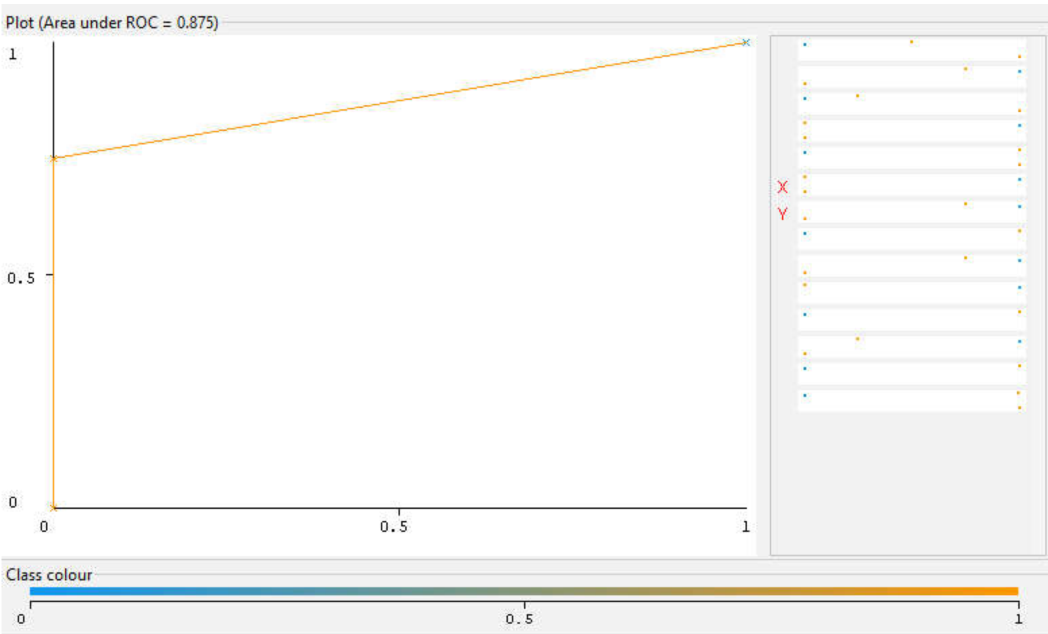

Fig 2: ROC curve for low effect on DT J48 algorithm

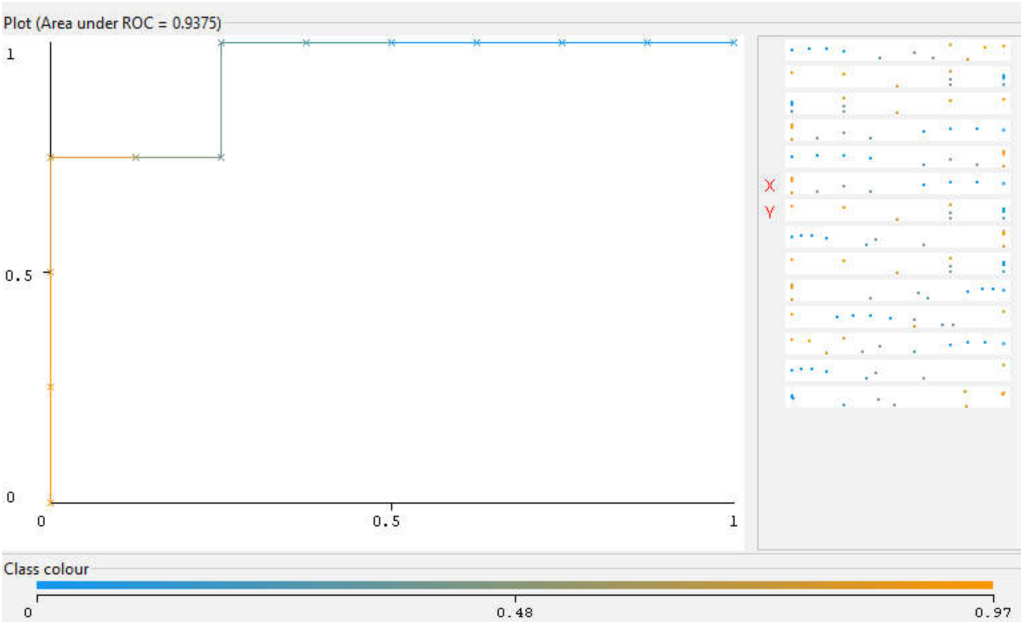

Fig 3: ROC curve for low effect on LMT algorithm

Many investigators predicted the dataset on biological science through ML algorithms [11,12,16,18-21]. In our earlier study, the PRC values were obtained higher (100.0%) compared to the present study (83.3%), which emphasized that lower diversity of insects on rice crop [11].

Xi et al. [19] studied a deep learning method with improved model accuracy and target object detection for Chinese butterfly specimens in which the training model accuracy of about 98% recorded.

Almryad & Kutucu [21] used Convolutional Neural Networks (CNNs) for the identification of butterfly species. Their results obtained approximately 80% success for both test and training dataset of three different network structures.

The predictive value of abnormal cellular features was indicated highest prediction accuracy of 98% as per different statistical values when studied neural network especially Multilayer perceptron [MLP] algorithm [16].

Interestingly, ML algorithms have been successfully utilized to a range of biodiversity issues [22], particularly to species richness as well as distribution [12, 23] and species-diversity modelling [24].

4. CONCLUSION

In the present study, NB, DT J48 and LMT models were explored NB model was obtained 83.3% on Precession recall curve value for Low effect. The ML algorithms performed accurately from the dataset of visiting insect diversity on rice crop, which was based on statistical interpretations. In this study, higher prediction accuracy was obtained for lower diversity as per WEKA tool. The future study is suggested with big dataset to predict classifier accuracy related to insect diversity as per big data mining.

Acknowledgment

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

Conflict of interest

No conflict of interest.

REFERENCES

[1] S. K. Mandal, and R. K. Mandal, Comparative efficacy of insecticides against mustard aphid, *Lipaphis erysimi* Kalt. Annals of Plant Protection Science, vol. 18 no. 2, **(2010),** pp. 333-335.

[2] R. Akoijam, and B. Singh, Persistence and metabolism of imidacloprid in rice. Bulletin of Environmental Contamination and Toxicology, vol. 92 no. 5, **(2014),** pp. 609-615.

[3] Anonymous, Agricultural Statistics at a Glance. Directorate of Economic and Statistics, Department of Agriculture and Cooperation, Ministry of Agriculture, Government of India, **(2013)**.

[4] M. A. Bashir, S. Saeed, A. Sajjad, K. A. Khan, H. A. Ghramh, M. A. Shehzad, H. Mubarak, et al. 2019.Insect pollinator diversity in four forested ecosystems of southern Punjab, Pakistan. Saudi Journal of Biological Sciences, vol. 26, **(2019)**, pp. 1835-1842.

[5] S. Bezzi, D. Kessler, C. Diezel, A. Muck, S. Anssour, and I. T. Baldwin, Silencing expression increases nectar germin, nectarins, and hydrogen peroxide levels and inhibits nectar removal from plants in nature. Plant Physiology, vol. 152, **(2010)**, pp. 2232-2242.

[6] A. Arenas, and W. M. Farina, Learned olfactory cues affect pollen-foraging preferences in honeybees, *Apis mellifera*. Animal Behaviour, vol. 83, **(2012)**, 1023-1033.

[7] A. Z. Siregar, Tulus, and K. S. Lubis, Diversity of pest insects in paddy field cultivation: A case study in Lae Parira, Dairi. International Journal of Trend in Research and Development, vol. 4 no. 5, **(2017)**, pp. 58-62.

[8] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning. Nature, vol. 521 no. 7553, **(2015)**, pp. 436-444.

[9] S. Mishra, A. Dash, and L. Jena, Use of deep learning for disease detection and diagnosis. In: Bio-inspired Neurocomputing. Studies in Computational Intelligence. (eds. Bhoi, A., Mallick, P., Liu, C.M. and Balas, V.), Springer, Singapore, **(2021)**, p. 903.

[10] M. Ebrahim, A. A. H. Sedky, and S. Mesbah, Accuracy assessment of machine learning algorithms used to predict breast cancer. Data, vol. 8 no. 2, **(2023)**, p. 35.

[11] S. Bandyopadhyay, T. K. Chatterjee, and S. N. Talapatra, Prediction of performance accuracy for visiting insect diversity on mustard flower: an approach through machine learning algorithms especially tree modelling. Journal Of Electronics Information Technology Science and Management, vol. 13 no. 5, **(2023)**, pp. 1-8.

[12] Y. Cao, T. E. Schartel, D. C. Houghton, J. Ross, and D. M. Infante, Evaluating machine leaning algorithms for accuracy, stability, and among-predictors discriminability in modeling species-richness across ten datasets. Ecological Informatics, vol. 90, **(2025)**, 103323.

[13] E. Frank, M. A. Hall, and I. H. Witten, The WEKA workbench, Online appendix for data mining: Practical machine learning tools and techniques. Morgan Kaufmann, 4th edition, **(2016)**.

[14] Witten, I. H., Frank, E., Hall, M. A. Data Mining: Practical machine learning tools and techniques. 3rd edn, Morgan Kaufmann, Burlington, MA, **(2011)**.

[15] S. N. Talapatra, R. Chaudhury, and S. Ghosh, CellProfiler and WEKA Tools: Image analysis for fish erythrocytes shape and machine learning model algorithm accuracy prediction of dataset. World Scientific News, vol. 154, **(2021)**, pp. 101-116.

[16] K. Bhattacharya, B. Mondal, and S. N. Talapatra, Multilayer perceptron network of machine learning for prediction accuracy after genetic biomonitoring of estuarine fish specimen. Journal of Electronics Information Technology Science and Management, vol. 12 no. 10, **(2022),** pp. 42-50.

[17] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, WEKA manual for version 3-8-5. University of Waikato, Hamilton, New Zealand, **(2020)**.

[18] I. Chakraborty, A. Choudhury, and T. S. Banerjee, Artificial intelligence in biological data. Journal of Information Technology & Software Engineering, vol. 7 no. 4, **(2017)**, p. 1000207.

[19] T., Xi, J., Wang, Y., Han, T., Wang, and L. Ji, The effect of background on a deep learning model in identifying images of butterfly species. Electrical and Electronics Engineering: An International Journal, vol. 8 no. 1, **(2019),** pp. 1-8.

[20] K. P. S. Attwal, and A. S. Dhiman, Exploring data mining tool - Weka and using Weka to build and evaluate predictive models. Advances and Applications in Mathematical Sciences, vol. 19 no. 6, **(2020)**, pp. 451-469.

[21] A. S. Almryad, and H. Kutucu, Automatic identification for field butterflies by convolutional neural networks. Engineering Science and Technology, an International Journal, vol. 23 no. 1, **(2020),** pp. 189-195.

[22] L. J. Pollock, J. Kitzes, S. Beery, K. M. Gaynor, M. A. Jarzyna, O. Mac Aodha, B. Meyer, D. Rolnick, G.W. Taylor, D. Tuia, and T. Berger-Wolf, Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. Nat Rev Biodiv, vol. 1, **(2025)**, pp. 1-17

[23] R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, J. Elith, Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol Monogr, vol. 92, **(2022)**, e01486.

[24] H. Yu, D.M. Infante, A.R. Cooper, J.A. Ross, Evaluating species richness, turnover, and range shifts under climate change for fluvial fishes in northeastern and Midwestern USA. Ecol Process, vol. 14, **(2025)**, p. 43