Machine Learning - Based Prediction of Genotoxicity in Peripheral Erythrocytes of Fish (*Labeo catla*): A Comparative Analysis of Algorithms

Kousik Seal[1], Soumendra Nath Talapatra[2]

[1]Ph. D Scholar, Food and Nutrition, School of Life Sciences, Secom Skills University, Kendradangal, Birbhum, West Bengal, India
Corresponding Author Email: *physiology.kousik[at]gmail.com*
Phone: +91 – 9903040481

[2]School of Life Sciences, Secom Skills University, Kendradangal, Birbhum, West Bengal, India

Abstract: ***This study utilizes machine learning ML algorithms to predict the accuracy of a genotoxicity dataset, focusing on nuclear abnormalities in peripheral erythrocytes of Labeo catla. Eight ML algorithms, including Logistic Regression, K - nearest neighbour, Lazy. KStar, DecisionStump, Hoeffding Tree, RandomForest, and RandomTree, were tested using the WEKA tool. Among these, RandomForest demonstrated the highest predictive accuracy with an area under the ROC curve of 91%. These results indicate that ML algorithms, particularly RandomForest, provide an effective approach for predicting genotoxicity in fish species.***

Keywords: Edible fish, Genotoxicity dataset, Machine learning, MN & NA parameters, *Labeo catla*

1. Introduction

The heavy metal lead (Pb) cause genotoxicity in the different cell types of fish, which may lead to an impact on the fish population and may be found to endanger. Several studies have shown that metals cause genotoxicity in fish species living in metal contaminated water. [1 - 9]

Interestingly, a recent study revealed that post - immersion of idol cause genotoxicity in the peripheral erythrocytes of fish species (*Labeo catla* and *Labeo rohita*) inhabiting pond ecosystem. [10]

Moreover, the dataset usage to evaluate the prediction accuracy through machine learning (ML) algorithm - based technique is suitable in classification and regression methods. Some earlier studies revealed that different algorithms of ML methods progressively classify the dataset of normal and abnormal shape of nuclei where the achievement was better related to different statistical interpretation. [11 - 14]

The present study was attempted to predict accuracy of a genotoxicity dataset focussing especially nuclear abnormalities in the peripheral erythrocytes of fish species (*Labeo catla*) by using ML algorithms.

2. Materials and Methods

In this study, the dataset was created from the image used in an earlier study by Seal and Talapatra. [10] The image was processed in the image analysis tool (CellProfiler, 2.1.1) as per the earlier protocol by Carpenter et al. [15] and Talapatra et al. [12] and the dataset of cell, cytoplasm, nucleus and class effect viz. normal and abnormal were used. The ML modelling was performed by using WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.5) developed by Frank et al. [16] The WEKA explorer was developed with data pre - processing, classification, regression, and association rules. [17]

In this study, the prediction accuracy was obtained through 5 machine learning (ML) algorithms especially The “predictive accuracy” of big dataset of RBCs for fish specimens on the “shape of cells, cytoplasm, and nuclei” after using ML algorithms especially different 8 classifiers viz. “Logistic regression (LR), K - nearest neighbour (IBK), Lazy. KStar (K*); DecisionStump (DS), Hoeffding Tree (HT), RandomForest (RF), and RandomTree (RT); ” based on 4 attributes viz. “cells, cytoplasm, nuclei, effect class (normal and abnormal cellular feature) ” to evaluate the overall predictive accuracy from the comparative dataset of pre and post idol immersion by using ML tool as per cross - validation (CV) test for normal and abnormal class. All the statistical summary results were retrieved for “F - measure, Matthew’s correlation coefficient (MCC), receiver operating characteristic (ROC) curve and Precision - recall curve (PRC) ”.

3. Results

Table 1 evaluates the summary results of correctly and incorrectly classified instances of studied models. In the case of algorithm classification, the highest values were observed in RF and RT (85.00% and 84.00%) followed by K* (81.00%), DS (78.00%), IBK and DTJ48 (77.00%), LR (73.00%), while lowest value of HT (62.00%) as per 10 - fold CV.

Table 1: Summary results of different models (correctly and incorrectly classified instances)

Classifier models	Correctly classified instances (%)	Incorrectly classified instances (%)
LR	73.00	27.00
IBK	77.00	23.00
K*	81.00	19.00
DS	78.00	22.00
HT	62.00	38.00
DTJ48	77.00	23.00
RF	85.00	15.00
RT	84.00	16.00

LR = Logistic regression; IBK = K - nearest neighbour; K* = Lazy. KStar; DS = DecisionStump; HT = Hoeffding Tree; RF = RandomForest; RT = RandomTree

Table 2 evaluates the summary results of Kappa statistic (KS), mean absolute error (MAE) and root mean squared error (RMSE) of studied models related to 10 - fold CV. In case of prediction accuracy of the class of KS values, the highest values were observed in RF and lowest value was observed for HT while MAE and RMSE values were declined as per following manner.

Table 2: Model summary (Kappa statistic, mean absolute error and root mean squared error) results

Classifier models	KS	MAE	RMSE
LR	0.47	0.40	0.45
IBK	0.54	0.23	0.47
K*	0.62	0.26	0.37
DS	0.54	0.31	0.41
HT	0.27	0.41	0.52
DTJ48	0.52	0.31	0.42
RF	0.70	0.23	0.34
RT	0.68	0.16	0.40

KS = Kappa statistic; MAE = Mean absolute error; RMSE = root mean squared error

Table 3 evaluates the statistical data of F - measure, MCC, ROC area and PRC area of studied models as per 10 - fold CV test. In case of prediction accuracy of the algorithms, the MCC, ROC area and PRC ranges between 70.0% - 50.0%, 61.0% - 91.0% and 67.0% - 91.0%, respectively observed. The area under curve (AUC) as per ROC curve value, the algorithm RF was predicted better as 91%.

Table 3: Statistical data for prediction accuracy of studied algorithms

Classifier models	F - measure	MCC	ROC area	PRC area	Class
LR	0.703	0.500	0.865	0.831	Normal
	0.752	0.500	0.865	0.853	Abnormal
IBK	0.777	0.543	0.773	0.745	Normal
	0.763	0.543	0.773	0.674	Abnormal
K*	0.822	0.618	0.888	0.897	Normal
	0.796	0.618	0.888	0.875	Abnormal
DS	0.823	0.578	0.744	0.720	Normal
	0.711	0.578	0.744	0.750	Abnormal
HT	0.513	0.331	0.613	0.705	Normal
	0.689	0.331	0.612	0.509	Abnormal
DTJ48	0.816	0.560	0.736	0.676	Normal
	0.693	0.560	0.736	0.748	Abnormal
RF	0.865	0.698	0.907	0.876	Normal
	0.831	0.698	0.907	0.914	Abnormal
RT	0.849	0.680	0.841	0.811	Normal
	0.830	0.680	0.841	0.759	Abnormal

MCC = Matthew’s correlation coefficient; ROC = Receiver operating characteristic curve; PRC = Precision - recall curve

4. Discussion

This study holds significance as it applies modern ML techniques to predict environmental genotoxicity, which could aid in monitoring aquatic ecosystem health and managing pollution impacts. Moreover, the genotoxicity dataset of fish species (*Labeo catla*) related to MN and NAs in the peripheral erythrocytes post idol immersion predicted better for an algorithm viz. RF (91%) in which the area under curve (AUC) as per ROC curve value. A similarity was obtained related to RF algorithm from previous studies by Abass [9] and Talapatra et al. [12] Moreover, Mondal et al. [14] predicted the accuracy of dataset related to the shape of normal and abnormal features of fish erythrocytes through 11 ML algorithms in which RF was also predicted better.

5. Conclusion

In conclusion, this study demonstrates that the RandomForest algorithm outperforms other ML models in predicting genotoxicity in *Labeo catla*. In the present study, the area under curve (AUC) as per ROC curve value, the RF algorithm recorded of about 91%, which predicted the mutagenic risk among this fish species.

Acknowledgement

Authors convey thanks to the developer of the tools used in this study.

Conflict of interest

Authors declare no conflict of interest in the present study.

References

[1] Talapatra SN, Banerjee SK. Detection of micronucleus and abnormal nucleus in erythrocytes from the gill and kidney of *Labeo bata* cultivated in sewage - fed fish farms. Food and Chemical Toxicology.2007; 45 (2): 210 - 5.

[2] Omar WA, Zaghloul KH, Abdel - Khaleka AA, Abo - Hegaba S. Genotoxic effects of metal pollution in two fish species, *Oreochromis niloticus* and *Mugil cephalus*, from highly degraded aquatic habitats. Mutation Research.2012; 746: 7 - 14.

[3] Nagpure NS, Srivastava R, Kumar R, Dabas A, Kushwaha B, Kumar P. Assessment of pollution of river Ganges by tannery effluents using genotoxicity biomarkers in murrel fish, *Channa punctatus* (Bloch). Indian Journal of Experimental Biology.2015; 53: 476 - 83.

[4] Nagpure NS, Srivastava R, Kumar R, Dabas A, Kushwaha B, Kumar P. Mutagenic, genotoxic and bioaccumulative potentials of tannery effluents in freshwater fishes of river Ganga. Human and Ecological

Risk Assessment: An International Journal.2016; 23 (1): 98 - 111.

[5] Igbo JK, Chukwu LO, Oyewo EO, Zelikoff JT, Jason BL. Micronucleus assay and heavy metals characterization of e - waste dumpsites in Lagos and Osun states, Southwest Nigeria. Journal of Applied Sciences and Environmental Management.2018; 22 (3): 329 - 37.

[6] Mandal M. Assessment of lead accumulation in muscle and abnormal nucleation in the peripheral erythrocytes of fish (*Mystus cavisus* HAM. - BUCH.) of Hooghly river downstream. Journal of Advanced Scientific Research.2020; 11 (1): 202 - 7.

[7] Mondal B, Bhattarcharya K, Swarnakar S, Talapatra SN. Assessment of nuclear abnormalities in the peripheral erythrocytes of fish specimen of Sundarbans coastal zone, West Bengal, India. Pollution Research.2021; 40 (4): 233 - 7.

[8] Gupta A, Talapatra SN. Assessment of genotoxicity in two fish species from East Kolkata Wetlands: Impacts and insights amidst shifting ecological dynamics. International Journal of Science and Research.2023; 12 (8): 1372 - 5.

[9] Abass NY. The influence of heavy metals on cytotoxicity in *Tilapia zillii*. Aquaculture International.2024.

[10] Seal K, Talapatra SN. A comparative assessment of genotoxicity in the peripheral erythrocytes of two fish species, *Labeo catla* and *Labeo rohita* Hamilton after pre - and post - immersion of idol in pond of Hooghly Area, West Bengal. International Journal of Science and Research.2024; 13 (4): 1063 - 6.

[11] Ratra R, Gulia P. Experimental evaluation of open source data mining tools (WEKA and Orange). International Journal of Engineering Trends and Technology.2020; 68 (8): 30 - 5.

[12] 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 - 16.

[13] Loebens N, Crispim B, Lima N, Tetila E, Costa C, Amorim W, et al. Fish erythrocytes nuclear abnormalities classification using machine learning. Anais do Workshop de Visão Computacional (WVC).2023; 96 - 101.

[14] Mondal B, Bhattacharya K, Talapatra SN. Image analysis to obtain dataset and machine learning for prediction accuracy on abnormal features of peripheral erythrocytes of fish specimen. Journal Of Electronics Information Technology Science and Management.2023; 13 (1): 125 - 35.

[15] Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology.2006; 7 (10): R100.

[16] Frank E, Hall MA, Witten I H. The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.

[17] Witten IH, Frank E, Hall MA. Data Mining: Practical machine learning tools and techniques.3rd edn, Morgan Kaufmann, Burlington, MA, 2011.