MLP Neural Networks in Sex Classification of a High Commercial Fish on the Artisanal Fisheries in Brazil

Tailon Carvalho de Cerqueira[1*], Iramaia de Santana[1]
[1]UNEB, PPGMSB; Alagoinhas, Bahia, Brazil

The scarcity of data on artisanal fishing, particularly in tropical South American countries, presents an additional hurdle for fisheries management and marine conservation. While artificial intelligence (AI) has found applications in various domains, including marine sciences, most AI models primarily focus on species identification. Unfortunately, these models often overlook data-limited scenarios. This article examines the potential of an MLP-type ANN model in addressing this gap. The model aims to classify the sex of a dioecious (separate sexes) fish species of high commercial importance, shedding light on its implications for fisheries management. Evaluation of the model's classification performance using precision, recall, f1-score, and accuracy reveals promising results exceeding 80% for both sexes across both training and testing phases. These findings underscore the potential of MLP models in aiding Brazil's fishing sector management in grappling with challenges stemming from data scarcity. By providing efficient information essential for decision-making regarding the management of specific fishing stocks, such models offer valuable insights into effective fisheries management strategies.
Keywords: Deep Learning. Classification Models. Lack of Data. Fisheries Management. Fishing Models.

Introduction

Artificial neural networks (ANN), a subset of Artificial Intelligence (AI), are designed to mimic the functionality of the human brain [1]. They consist of a synaptic neural system based on logical-mathematical structures that abstract the functioning of neurons in a simplified manner [2]. AI algorithms have gained widespread popularity in marine sciences due to their versatility and efficiency in solving complex problems faster than humans [3]. With their ability to make predictions and classifications based on data relationships, AI algorithms offer solutions to various problems. They utilize different architectures, such as a single unit for regression or binary classification or multiple units (K units) for K-class classification [4]. One of the critical advantages of deep learning is its capability to identify non-linear relationships between variables [5].

Received on 22 November 2023; revised 17 December 2023.
Address for correspondence: Tailon Carvalho de Cerqueira. BR 110, Km 03, Alagoinhas. Zipcode: 48.000.000. Alagoinhas, Bahia, Brazil. E-mail: tailoncerqueira7@gmail.com.

J Bioeng. Tech. Health 2023;6(Suppl 2):26-32

Applying models based on such architectures to predict future scenarios for species exploited by artisanal fishing in tropical environments shows promise [6]. Artisanal fishing often faces challenges due to a scarcity of biological data and exact reproductive data, which is crucial for decision-making processes aimed at species protection [7]. Understanding the sex ratio of species is vital for conservation, management, and sustainable use practices [8].

This study employs a multilayer perceptron (MLP) artificial neural network to classify the sex of a commercially valuable fish species found in the coastal artisanal fisheries of northeastern Brazil.

Materials and Methods

This study adhered to the workflow stages outlined in Figure 1, which were as follows:

1. Exploratory data analysis involves identifying the dataset's structure, conducting basic statistical tests, and addressing missing data.
2. Analysis of attribute correlations and treatment of outliers.
3. Data division for training and testing purposes.

4. Training and testing of the model.
5. Evaluation of performance metrics. Each stage was meticulously followed to ensure a systematic and comprehensive approach to the analysis.

Exploratory Analysis

Females and 430 males, ensuring a balanced dataset regarding the predictor attribute, sex. As such, there was no requirement to implement methods to address data imbalance. Pearson's correlation matrix was employed to identify attributes strongly correlated with each other for network construction. This method facilitated the selection of relevant attributes. Furthermore, all selected attributes exhibited a normal distribution, as confirmed by the Shapiro-Wilk test ($p>0.05$).

Dataset Division: Training and Testing

The ANN model's dataset for training and testing was split by year, with 457 records allocated for training and 446 for testing. This method was chosen to maintain the inherent characteristics of natural temporal periods (such as day, lunar cycle, month, and seasons) as much as possible within a year. Consequently, the training dataset encompassed the timeframe from June to December 2008, while the testing dataset covered the period from January to May 2009. This division ensured that the model was trained and evaluated on distinct temporal periods, facilitating a comprehensive assessment of its performance.

Model Training

A classification model was constructed to predict the sex attribute, with 0 representing males and 1 denoting females. This model utilized biological attributes characteristic of the population dynamics of species commonly targeted by artisanal fishing and circadian and annual seasonality attributes to address data scarcity scenarios. The architecture of the classification model comprised five layers (Figure 2). It included an input layer with 11 attributes (biological and environmental factors) and three hidden tanh layers, as described by Equation 1.

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{tan\,h}(\,\boldsymbol{x}) = \frac{2}{1+\boldsymbol{e}^{-2\boldsymbol{x}}} - 1 \qquad (1)$$

and a sigmoid-type output (Equation 2),

$$\mathbf{Sigmoid} = \mathbf{f}(\mathbf{x}) = \frac{2}{1+\mathbf{e}^{-\mathbf{x}}} - 1 \qquad (2)$$

The model's output represents the probability of belonging to each respective class. Utilizing a threshold of 0.5, values equal to or below this threshold were classified into class 0, while values above it were classified into class 1.

The network's predictive capability was evaluated using the following metrics, as described by [9]:

Accuracy: This metric assesses the model's

Figure 1. Workflow: Steps used until the output of the model.

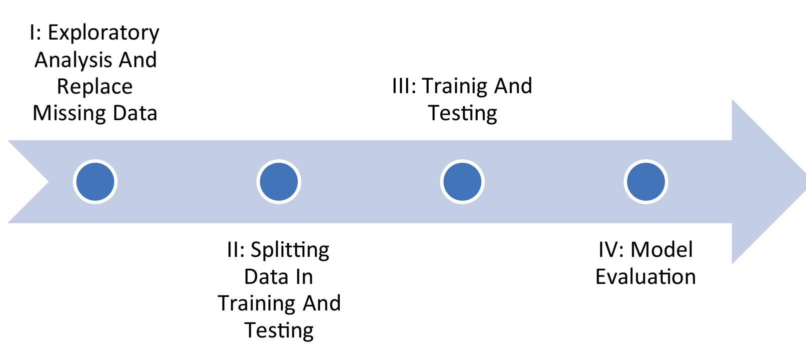
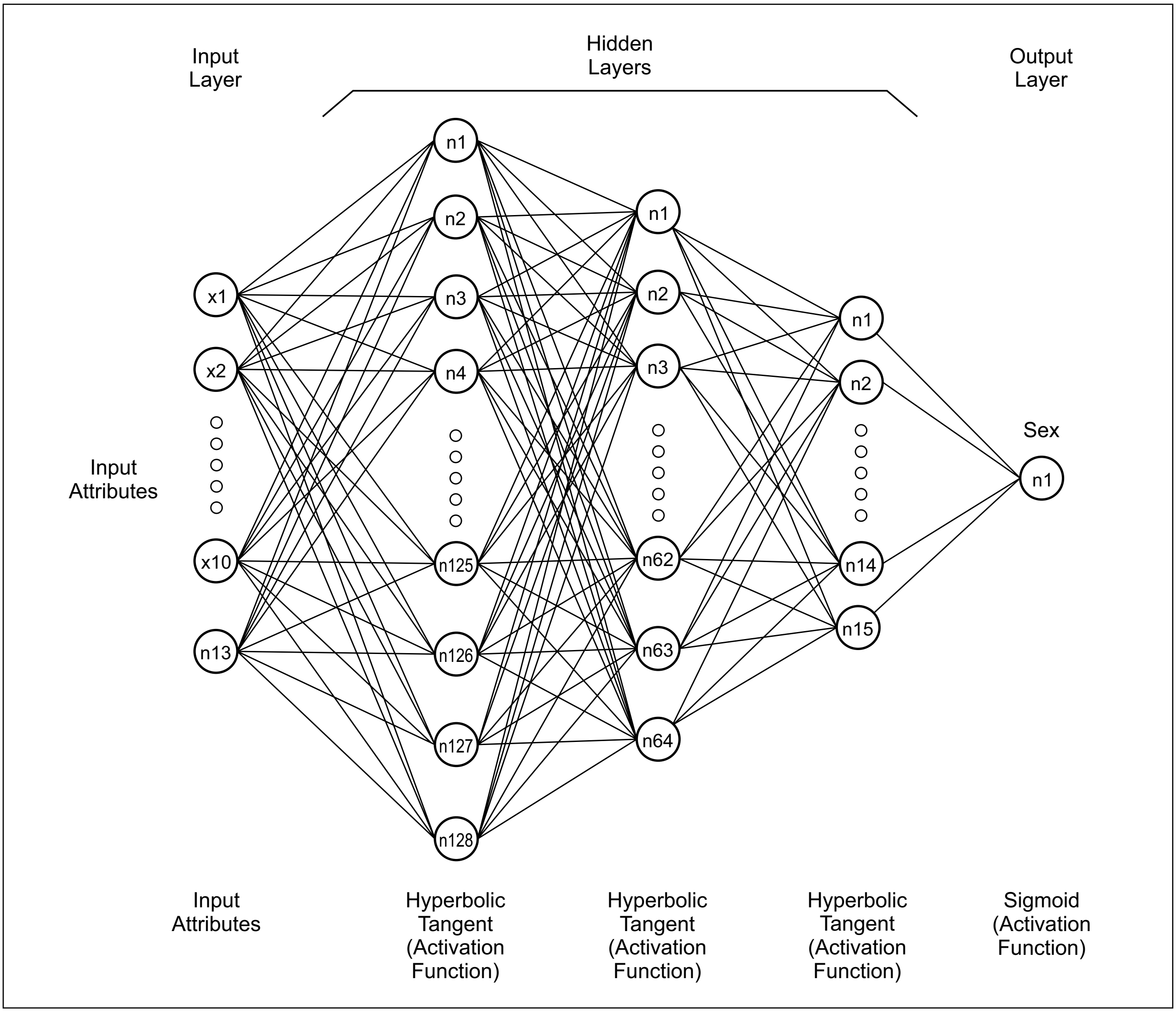


Figure 2. Architecture binary classification.



performance by calculating the percentage of correct classifications, representing the total error measure. It is calculated using Equation 3:

$$accuracy = \frac{\Sigma TruePositive + \Sigma TrueNegative}{\mathbf{SampleSize}} \quad (3)$$

Precision: This metric reflects the accuracy of the model by measuring the proportion of correctly identified positive values (TP) over the total number of instances predicted as positive (TP + FP). Precision indicates how many of the predicted positive instances are correctly classified. It is calculated using Equation 4:

$$precision = \frac{\Sigma TruePositive}{\mathbf{\Sigma TruePositive + \Sigma FalsePositive}} \quad (4)$$

Recall: Also referred to as sensitivity or actual positive rate, recall measures the model's ability to predict positive instances correctly or the proportion of actual positive instances correctly identified. It is calculated as the ratio of true positives (TP) to the total number of actual positive instances.

Recall indicates how many positive instances were correctly classified by the model. This metric is calculated using Equation 5:

$$\boldsymbol{recall} = \frac{\boldsymbol{\Sigma TruePositive}}{\mathbf{\Sigma TruePositive + \Sigma FalseNegative}} \quad (5)$$

F1-score: As an F-measure, the F1-score is a function that computes the harmonic mean between precision and recall, providing a balanced assessment of the model's performance. It is calculated using Equation 6:

$$\boldsymbol{f}1_{score} = 2 * \frac{\mathbf{precision} * \mathbf{recall}}{\mathbf{precision} + \mathbf{recall}} \quad (6)$$

The confusion matrix was utilized to visualize errors and successes by class, serving as a reliable indicator of classification performance compared to the expected results (Table 1).

Results and Discussion

The classification model's performance was assessed using the metrics described in the training section of the models (Table 2).

Initially, the cross_val_score function was employed (Figure 3), which generates a cross-validated accuracy score for each data point within our dataset. This method involves splitting the dataset into multiple subsets of training and testing

Table 1. Confusion Matrix for a binary output.

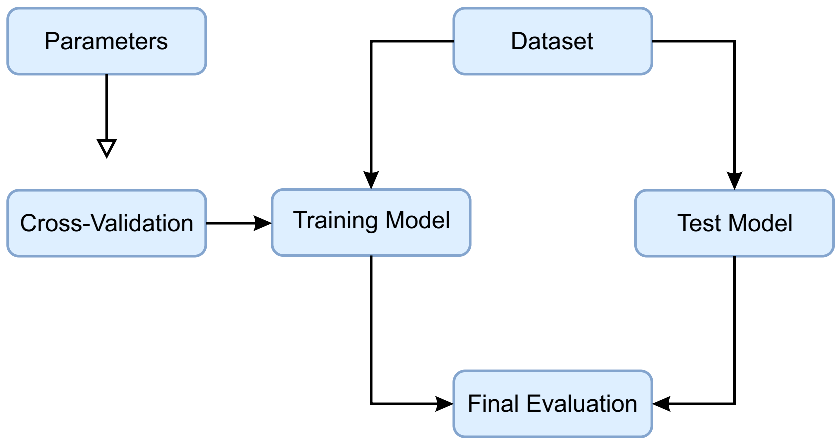
True Value		Predict Value: ŷ= 0	Predict Value: ŷ= 1
	y = 0	True Positive	False Positve
	y = 1	False Negative	True Negative

Table 2. Generalized model input parameter variations.

Model: BDPAF+ EA			
Metric		**Male**	**Femele**
Precision	Tn	0.85%	0.86%
	Ts	0.91%	0.84%
Recall (sensibility)	Tn	0.86%	0.85%
	Ts	0,86%	0.91%
f1-score	Tn	0.85%	0.86%
	Ts	0,89%	0,87%
Accuracy	Tn	0.86%	
	Ts	0,88%	
Cross-Validation		Acuracy	Std
		82.93%	(5.33%)

BDPAF = biological attributes of population dynamics in fisheries; EA = environmental attributes.
Legend: Tn= train, Ts= test, Std = Standard Deviation

Figure 3. Workflow by applying cross-validation before evaluating separate training and testing metrics.



data, training the model on each training subset, making predictions on the testing subset, and outputting the prediction accuracy score for each subset [10]. The data separation technique used was Stratified K-Folds, which provides train/test indices to split data into train/test sets.

This method aimed to provide an initial assessment of the model's performance by considering the entire dataset before conducting training and subsequent separate training and testing evaluations. The preliminary results of the cross-validation closely aligned with those obtained in the training and testing stages, yielding an accuracy of 86.02% and a standard deviation of 2.88% concerning the error. These findings demonstrate a model performance consistent with the training and testing evaluations, as discussed below.

The training data revealed that the model effectively distinguished between the classes with acceptable error rates, amounting to 14% and 12% for training and testing, respectively, for both males and females. This demonstrates a satisfactory performance during training. Incorporating biological parameters of fishing and environmental population dynamics led to a higher prediction accuracy in the testing phase compared to the training phase, achieving 88% and 86%, respectively. This indicates that the model could generalize responses and establish a strong correlation between attributes, effectively distinguishing males from females.

The precision (accuracy) of the model in classifying males and females during the training phase was 85% and 86%, respectively. In the testing phase, the precision was 84% for males and 91% for females. The results combining the recall metric (86% for males in both testing and training and 91% and 85% for females) are auspicious when considering biological attributes amidst temporal variations in the context of commercially exploited species. Environmental fluctuations play a crucial role in determining fishermen's fishing strategies, thereby exerting a deterministic influence on the sex ratio of species. Hence, the model's ability to accurately classify individuals based on these attributes is significant.

Given that the behavior and physiology of living beings exhibit recurring and periodic patterns, such as those conditioned to seasonal variations [11], natural organisms are expected to manifest changes in sex ratio at least qualitatively, as predicted in theory [12]. The promising nature of utilizing MLP to discern a fundamental attribute in population dynamics studies is thus confirmed. With an f1-score of 87% for the male class and 89% for the female class, there exists a balance in the accuracy and sensitivity metrics for predicting both sexes. Models exhibiting high accuracy alongside high sensitivity are particularly suitable for fisheries management scenarios, as this indicates the relevance of the accuracy achieved in classification.

Overall, the confusion matrix yielded a hit/error rate of 188 males (86%) and 194 females (85%) in the training phase and 204 males (86%) and 184 females (91%) in the test phase (Tables 3 and 4). These results demonstrate that the MLP possesses a high capability to confirm classes, both in estimating the probability of correct positive and negative predictions and through its ability to avoid classification errors, i.e., in estimating the probability of error.

In scenarios where data collection capacity is limited, understanding the distribution of individuals across sexes provides valuable insights into the state of fish stocks for a given species, beginning with the implications of the sex ratio. For instance, estimating the number of individuals by sex informs about sperm competition [13] and the success of oocyte fertilization, impacting the spawning stock biomass [14], fishing exploitation practices, and the communities reliant on them. Aquatic ecologists often conduct manual underwater animal counts, a process that is time-consuming, labor-intensive, and costly [3].

By integrating computer vision technologies, supporting models for assessing fish stocks in data-scarce conditions becomes feasible. For instance, by identifying the sex of individuals recorded during the landing process, sampling costs can be optimized and more efficient. This approach holds promise for enhancing the accuracy and efficiency of fisheries management strategies.

Table 3. Confusion matrix showing classification error (Error, 12%) for male *versus* female.

True Value \ Predict Value	ŷ = 0	ŷ = 1	Error (%)
y = 0	188	31	14%
y = 1	33	194	15%

Values represent the numbers of fish in each sexo based on the actual classification (rows) *versus* the predicted classification (columns).

Table 4. Confusion matrix showing classification error (14%) for males *versus* females.

True Value \ Predict Value	ŷ = 0	ŷ = 1	Error (%)
y = 0	204	34	14%
y = 1	19	184	9%

Values represent the numbers of fish in each sexo based on the actual classification (rows) *versus* the predicted classification (columns). Legend: 0 = males; 1 = females.

Conclusion

In conclusion, this study examined the efficacy of an MLP in identifying meaningful patterns for sex classification in a species of high commercial value, utilizing biological data on population dynamics and environmental factors relevant to artisanal fisheries in tropical marine ecosystems. However, the significance of these findings extends beyond the model's performance to the size and temporal constraints of the dataset, which is characteristic of information available for artisanal fishing fleets in tropical countries.

The outcomes of this research underscore the considerable potential of MLPs in aiding the conservation and management of fish populations, particularly in scenarios where data availability is limited, such as in artisanal fishing in developing nations. This highlights the practical utility of advanced computational tools in addressing complex challenges in fisheries management and underscores the importance of continued research and application of such methods in marine conservation efforts.

Acknowledgment

We acknowledge PCI-AECID: PCI 2007 Iberoamérica A/8268/07 and PROGRAMA PROFORTE, 14/2008, PPG - UNEB, for financing part of the field collections to compose the dataset used in this work. We also acknowledge the IIM-CSIC for all the expenses covered for the perfect development of the laboratory activities and the stay in its facilities.

References

1. Turing AM. On computable numbers, with an application to the Entscheidungs problem. Proceedings of the London Mathematical Society 1937;(1 s2-42):230–265, 1937. doi: https://doi.org/10.1112/plms/s2-42.1.230.
2. Géron A. Mãos a obra com: Aprendizagem de máquina com Scikit-Learn, Keras & TensorFlow. Rio de Janeiro, 2021.
3. Ditria EM, Lopez-Marcano S, Sievers M, Jinks EL, Brown CJ, Connolly RM. Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. Front Mar Sci 2020;7(Jun). doi: 10.3389/fmars.2020.00429.
4. Grosse R. Lecture 5: Multilayer Perceptrons. 2018.
5. Cavieses-Núñez R, Ojeda-Ruiz MA, Flores-Irigollen A, Marín-Monroy E, Lbañez-Lucero M, Sánchez-Ortíz C. Focused small-scale fisheries as complex systems using deep learning models. Lat Am J Aquat Res 2021;49(2):342–353. doi: 10.3856/vol49-issue2-fulltext-2622.
6. Jiang M, Zhu Z. The role of artificial intelligence algorithms in marine scientific research. Front Mar Sci 2022;9(May). doi: 10.3389/fmars.2022.920994.
7. de Santana I. Ecologia Reprodutiva de Lutjanidae no Litoral Norte da Bahia, Brasil: contribuições ao manejo sustentável da pesca artesanal. Universidade de Vigo, 2015.
8. Selvaraj JJ, Rosero-Henao LV, Cifuentes-Ossa MA. Projecting future changes in distributions of small-scale pelagic fisheries of the southern Colombian Pacific Ocean. Heliyon 2022;8(2). doi: 10.1016/j.heliyon.2022.e08975.
9. Bruce P, Bruce A. Estatística Básica para Cientistas de Dados. Rio de Janeiro, 2019.
10. Educative. How to implement cross_val_score in sklearn. https://www.educative.io/answers/how-to-implement-crossvalscore-in-sklearn.
11. Araújo JF. Cronobiologia: uma multidisciplinaridade necessária. Marques N 2022: 95–112.
12. Jeger J, Stubblefield J. Models of sex ratio evolution. 2002:2–25. doi: 10.1017/CBO9780511542053.002.
13. Klemme JAEI et al. Reproductive success of male bank voles (*Clethrionomys glareolus*): the effect of operational sex ratio and body size. Behav Ecol Sociobiol 2007;61:1911–1918.
14. Brown-Peterson NJ, Leaf RT, Leontiou AJ. Importance of depth and artificial structure as predictors of female red snapper reproductive parameters. Trans Am Fish Soc 2021;150(1):115–129. doi: https://doi.org/10.1002/tafs.10277.