Predicting Aquatic Environment For Fish Species Survival Using Random Forest Model

Sumon Ahemed
dept. Internet of Things and Robotics Engineering
Bangabandhu Sheikh Mujibur Rahman
Digital University, Bangladesh
Mohish bathan, Kaliakor, Gazipur, Bangladesh
1801042@iot.bdu.ac.bd

Bitta Boibhov Barmon
dept. Internet of Things and Robotics Engineering
Bangabandhu Sheikh Mujibur Rahman
Digital University, Bangladesh
Mohish bathan, Kaliakor, Gazipur, Bangladesh
1901050@iot.bdu.ac.bd

***Abstract*—The primary objective of this research is to develop a machine learning model capable of recommending the most suitable fish species for aquaculture within a specific aquatic environment. Our approach employs a Random Forest (RF) model and is validated using a dataset that includes information on 11 different fish species. The prediction of fish species is based on various environmental characteristics, such as pH, temperature, and turbidity. Performance evaluation is carried out using metrics including accuracy, true positive (TP) rate, and kappa statistics. Our experimental results indicate that the RF-based prediction model achieves an accuracy of 98.31% for balanced dataset and 81.08% for unbalanced dataset, a kappa statistic of 98.14%, when applied to the test dataset. Furthermore, we conducted a comparison with other Logistic Regression,Support Vector Machines, Gaussian Naive Bayes,KNN ,Decision Tree. We also developed a web application for predict fish species and also predict aquatic environment. Our proposed model outperforms these existing models, exhibiting higher accuracy, and kappa statistics.**

***Index Terms*—Random forest model, Classification, Python, Flask etc.**

.

I. Introduction

Aquaculture, the practice of farming aquatic organisms for food, encompasses the breeding, nurturing, and harvesting of fish, mollusks, crustaceans, and aquatic plants in both freshwater and saltwater environments. Its origins date back around 4,000 years to China, and today, China and other Asian countries continue to dominate global aquaculture production. Aquaculture plays a crucial role in providing sustenance for impoverished communities worldwide, as well as being a significant industry for major corporations. It now supplies over half of the world's seafood consumption, a percentage that is continually increasing as the global population grows. According to the Food and Agricultural Organization (FAO), aquaculture production has risen from 3 million tons in the 1970s to over 80 million tons in 2017.

Manual fish classification is a complex and laborious task, particularly for those who are not experts in the field. Fish species are vital in various industrial, agricultural, and food production sectors and serve as a significant food source for humans. Marine biologists traditionally classify fish based on their characteristics and utilize classification trees, which has led to the adoption of machine learning and data systems, saving time, effort, and improving the speed of fish classification.

Fish classification involves identifying fish species based on their characteristics or similarities, aiding in various aspects such as pattern recognition, subsistence matching, feature extraction, identification of physical or behavioral traits, statistical analysis, and quality assessment for different types of fish. Additionally, fish classification is critical for fisheries management and population assessments.

Automated fish classification, using machine learning models like the decision tree classifier (J48), random forest (RF), k-nearest neighbor (k-NN), and classification and regression tree (CART), can expedite the process and enhance the accuracy of species identification. While traditional rule-based algorithms lack predictive capabilities for unknown datasets, machine learning models provide prediction features. Evaluation measures like the confusion matrix help assess prediction accuracy, which rule-based algorithms cannot provide. Although deep learning models like Convolutional Neural Networks (CNN) are available, their computational complexity is higher compared to traditional machine learning models, necessitating longer training times.

This paper presents a fish survival prediction model in an aquatic environment based on the RF model. The paper is organized as follows: Section 2 provides a literature review, Section 3 discusses the proposed model, Section 4 details the experimental setup and analysis results, and Section 5 concludes with the findings of this research.

.

II. Literature Review

The existing literature discusses a range of decision support systems implemented in aquaculture garden operations. Some of these systems employ machine learning techniques, while others do not. One such system is proposed for automatic fish identification, where features related to shadow and texture are extracted from fish images [1]. Another system introduces a structure that utilizes real-time water quality indicators and operational data to assess their impact on the survival rate, biomass, and production outcomes of aquaculture species [2]. In addition, a prediction model is presented that focuses on

a single water feature, dissolved oxygen (DO), for aquatic creatures [3]. Furthermore, hardware has been developed to monitor water quality factors, including pH, temperature, and dissolved oxygen [4]. An IoT device is also proposed for detecting and controlling water factors like pH and temperature, although data analysis is not included. Additionally, a regression model is employed for predicting water quality in fish cultivation, albeit without emphasis on prediction accuracy [6].

Furthermore, an automated strategy has been devised for fish identification primarily through the utilization of support vector machines and the k-means clustering algorithm [7]. A computerized robust classification approach for Nile-Tilapia fish is introduced in, where scale-invariant features of fish's physical changes are extracted and utilized as input for a support vector machine.

The management of hatchery production is centered around the application of rules and calculations involving physical, chemical, and biological processes. A scientific model has been developed to assess environmental impacts [5]. Domain experts have crafted specific rules for the system. A machine learning method is presented to strike a balance between farm closure and opening events. A feature ranking algorithm is demonstrated for identifying the most influential factors contributing to farm closures. Time series machine learning techniques such as principal component analysis (PCA) and autocorrelation function (ACF) are employed to predict closure events [8]. Furthermore, a set of rules is derived from sensor network data to establish associations between environmental variables and algae growth. An ensemble method is designed to identify relevant environmental variables responsible for algae growth and predict its development. A machine learning method is also devised to forecast the propagation of algae patches along waterways [7].

.

III. Proposed Model

Figure 1 illustrates a comprehensive block diagram of the proposed model. The initial step involves importing our dataset. In the preprocessing phase, we apply filtering and resampling techniques to prepare our dataset. Subsequently, we opt for RF classifiers as our model of choice within the classification section. This is where we perform classification using various machine learning models. Following the classification process, we generate predictions based on the classifier's output.

.

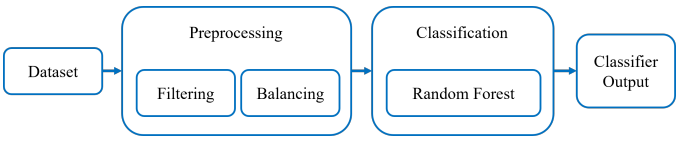

Fig. 1. Block diagram of proposed model

A. Description of dataset

This study relies on data collected from the Faculty of Fisheries at the University of Dhaka, Bangladesh, encompassing 191 instances and four attributes. These attributes include pH, temperature, turbidity, and fish species. The dataset is divided into two parts: one containing information about the aquatic environment, and the other focusing on fish species. The target attribute consists of 11 different fish species, including katla (14 instances), shing (17 instances), prawn (14 instances), rui (19 instances), koi (15 instances), pangas (22 instances), tilapia (25 instances), silver carp (7 instances), karpio (33 instances), magur (11 instances), and shrimp (14 instances).

Regarding the aquatic environment characteristics, we specifically consider pH, temperature, and turbidity as the key parameters of interest.

- pH: pH is necessary for aquaculture as a measure of the acidity of the water or soil. The optimal pH for fish is between 6.5 and 9. Fish will grow poorly, and reproduction will be affected at consistently greater or lower pH tiers [25]. The pH level for warm-water pond fish is 4 for acid death point, 4 to 5 for no reproduction, 5 to 6.5 for slow growth, 6.5 to 8.5 for desirable ranges, 9 to 10 for slow growth, and 11 for alkaline death point.
- Temperature: The increase and endeavor of the fish rely on their physique temperature. The body temperature of the fish is about the same as the water temperature and varies with it. Each fish species is tailored to develop and reproduce inside well-defined stages of water temperatures, but the most useful boom and replica take area within narrower tiers of temperature. It is important, therefore, to understand the water temperatures reachable at your fish farm nicely to pick out the right species of fish and to graph its management as a result.
- Turbidity: The ability of water to transmit the light that restricts light penetration and limit photosynthesis is termed as turbidity and is the resultant impact of several elements such as suspended clay particles, dispersion of plankton organisms, particulate natural things and also the pigments caused with the aid of the decomposition of organic matter. Acceptable turbidity varies from 30-80 cm is properly for fish health.
- Fish species: In our dataset, we utilized a total of 11 fish species as the target variable. The fish species in our dataset are presented in Figure 2; where carpio fish is shown in Figure 2(a), katla fish is in Figure 2(b), rui fish is in Figure 2(c), koi fish is in Figure 2(d), magur fish is in Figure 2(e), pangas fish is in Figure 2(f), prawn fish is in Figure 2(g), silver carp fish is in Figure 2(h), tilapia fish is in Figure 2(i), and shing fish is in Figure 2(j).

All the parameter are described that are essential for better aquatic environment like PH, temparature, turbidity. And also In figure 2 all the fish that are used in our project are shown by image. There are 11 species fish and all the fish species image is shown together and labeled.

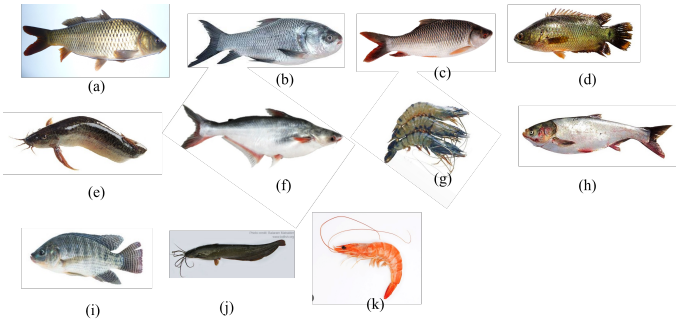

Fig. 2. Sample fishes: (a) carpio fish, (b) katla fish, (c) rui fish, (d) koi fish, (e) magur fish, (f) pangas fish, (g) prawn fish, (h) silver carp fish, (i) tilapia fish, (j) shing fish and (k) shrimp fish

B. Visualization

In figure 3 we see the max abd min values of all parameter like ph, temparatue, turbidity. In the plot we also observed those values that is more in dataset.

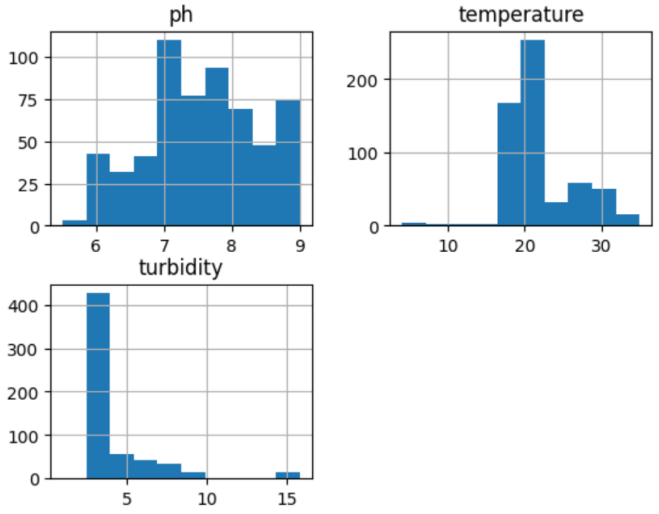

Fig. 3. Histogram Plot of numerical value

C. Preprocessing

In the preprocessing step, we filtered our dataset using a resampling option for observing the current relation of instances and attributes of the dataset. In the attribute selection window, we can check the missing, unique, and distinct value of each attribute. All attributes show 0% missing and pH has 28 unique values, temperature has 22 unique values, turbidity has 56 unique and fish has 11 distinct values.

.

D. Classification

In the classification section, we classified our dataset using 5 various classifiers model. RF outperforms the other described model.

1) Random forest: RF is a supervised learning method that is a decision tree-based algorithm. As the name proposes as forest the RF classifier is an ensemble of decision trees wherever a random vector sample produce each classifier from the input vector [28] and every tree cast a unit vote for the most popular class to classify an input vector, nearly all of the time trained with a bagging method.

The preparation calculation for RF applies the overall strategy of bootstrap collecting, or packing, to tree students. Given a preparation set $\mathrm{X} = \mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}$ with reactions $\mathrm{Y} = \mathrm{y}_1, \ldots, \mathrm{y}_\mathrm{n}$, stowing more than once (A times) chooses an irregular example with substitution of the preparation set and fits trees to these examples. For $\mathrm{a} = 1, \ldots\ldots, \mathrm{A}$: - Test, with substitution, n preparing models from X, Y; call these X_a, Y_a. - Train a characterization or relapse tree $\mathrm{f_a}$ on $\mathrm{X_a}, \mathrm{Y_a}$.

After preparing, expectations for concealed examples x' can be made by averaging the forecasts from all the individual relapse trees on x' :

$$\hat{f} = \frac{1}{A} \sum_{a-1}^{A} f_a\left(\mathrm{x}'\right)$$

also, a gauge of the vulnerability of the forecast can be made as the standard deviation of the expectations from all the individual relapse trees on x :

$$\sigma = \sqrt{\frac{\sum_{a-1}^{A}\left(f_a\left(\mathrm{x}'\right) - \hat{f}\right)^2}{A-1}}$$

The universal thought of the bagging method is that the composing of the learning method increases the overall result. The RF is less sensitive than other streamline machine learning classifiers to overfitting and to the quality of training samples [29]. Figure 3 shows the concept of RF model. Tree 1 and Tree 2 belong to Class A. So, predicted output will be Class A. Majority vote is Class A in Figure 4.

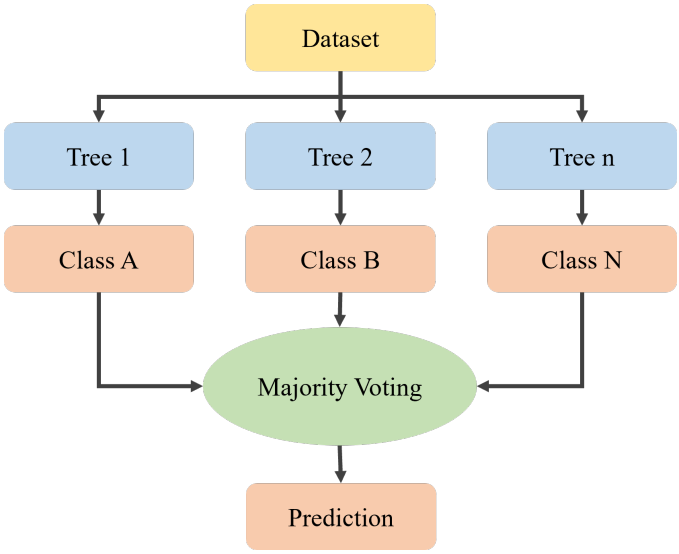

Fig. 4. RF model

E. Classifier output

In the classifier section, we can see the result performance of our model and other state-of-art models. By choosing our described model, we can check results. In this section, we can see detailed accuracy by class. Figure 4 shows these performance results. We did not find any machine learning model for fish environment monitoring using RF. The dataset we have used in our own dataset. Figure 4 showing the accuracy of various fish species classification using a machine learning algorithm. The table has 13 rows and 4 columns. The first column shows the fish species, the second column shows the precision, the third column shows the recall, and the fourth column shows the f1-score.

All of the fish species in the table have an accuracy of 98% or higher, except for katla, which has an accuracy of 93%. This means that the algorithm is very good at identifying all of the fish species, except for katla, which it is sometimes mistaken for other fish species.

The overall accuracy of the algorithm is 98%, which means that it correctly identified 98% of the fish species in the test set. This is a very good result, and it shows that the algorithm is effective at identifying fish species.

.

```
print(classification_report(test_y,y_pred))
```

	precision	recall	f1-score	support
karpio	1.00	1.00	1.00	35
katla	1.00	0.93	0.96	28
koi	0.87	1.00	0.93	27
magur	1.00	1.00	1.00	29
pangas	1.00	0.97	0.99	35
prawn	0.95	1.00	0.97	37
rui	1.00	1.00	1.00	32
shrimp	1.00	1.00	1.00	30
silverCup	1.00	1.00	1.00	35
sing	1.00	0.97	0.99	39
tilapia	1.00	0.93	0.96	28
accuracy			0.98	355
macro avg	0.98	0.98	0.98	355
weighted avg	0.98	0.98	0.98	355

Fig. 5. Classification report

IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

As data analysis, we have used WEKA tool for classifying the proposed model and described other models. The tool is very helpful to analyze and has various techniques embedded in it. We have used 10% images for testing and 90% images for training in each species for all described model.

A. Performance metrics

Performance parameters are the most important metrics to compare among classifier methods to get the best classifier. We have applied 3 performance parameters which are accuracy, true positive (TP) rate and kappa statistics. The parameter is calculated from a confusion matrix which is situated in every step of classification. Accuracy is measured by dividing the total number of correctly classified instances by the total number of instances and also it is measured by confusion matrix). TP rate is another performance metric of our study and it is calculated by (1). And kappa statistic is the last metric of our paper which is computed by (3). The higher the kappa statistics, the better the model accuracy level. A general view of the confusion matrix is illustrated in Figure 6 and Figure 7.

Here, TP signifies the number of properly classified positive occurrences. Equation 1:

$$\text{TP Rate} = \frac{TP}{FN + TP}$$

It is also known as the recall. It tells us what percentage of positive instances have been correctly identified. - FP signifies the number of misclassified positive occurrences. - FN signifies the number of misclassified negative occurrences. - TN signifies the number of properly classified negative occurrences. Equation 2:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is also represented by total accuracy. Equation 3:

$$\text{Kappa statistic} = \frac{\text{Total accuracy} - \text{random accuracy}}{1 - \text{random accuracy}}$$

We have used Waikato environment for knowledge analysis (WEKA) for processing data. The proposed model, RF shows the accuracy as the value 88.4817%, the average TP rate as the weight of 88.5% and kappa statistic as the standard of 87.11%. We can say, these three metrics give a better result. We have compared the performance metrics with our proposed model and other state-art-models. We utilized 5 models in our experimental work. They are RF, J48, Naïve Bayes, k-NN, and CART. Table 3 depicted a detailed comparison with all model each other.

The confusion matrix in figure 6 in the image is for a classification problem with 10 classes. The accuracy of the classifier is 98.31%, meaning that it correctly predicted 98.31% of the samples in the test set. The precision of the classifier is 98.48%, meaning that 98.48% of the samples that the classifier predicted as positive were actually positive. The recall of the classifier is also 98.31%, meaning that 98.31% of the positive samples in the test set were correctly predicted as positive.

These performance metrics are shown in Figure 7 graphically. The fihure you sent is a heatmap of the correlation between three variables: pH, temperature, and turbidity. The heatmap shows that the temperature and turbidity are highly correlated, with a correlation coefficient of 0.27. The pH is

```
Confusion Matrix:
[[35  0  0  0  0  0  0  0  0  0  0]
 [ 0 26  2  0  0  0  0  0  0  0  0]
 [ 0  0 27  0  0  0  0  0  0  0  0]
 [ 0  0  0 29  0  0  0  0  0  0  0]
 [ 0  0  0  0 34  1  0  0  0  0  0]
 [ 0  0  0  0  0 37  0  0  0  0  0]
 [ 0  0  0  0  0  0 32  0  0  0  0]
 [ 0  0  0  0  0  0  0 30  0  0  0]
 [ 0  0  0  0  0  0  0  0 35  0  0]
 [ 0  0  0  0  0  1  0  0  0 38  0]
 [ 0  0  2  0  0  0  0  0  0  0 26]]
Accuracy: 98.31 %
Precision: 98.48 %
Recall: 98.31 %
F1-Score: 98.33 %
Matthews Correlation Coefficient: 98.15 %
Kappa Statistic: 98.14 %
```

Fig. 6. Confution matrix

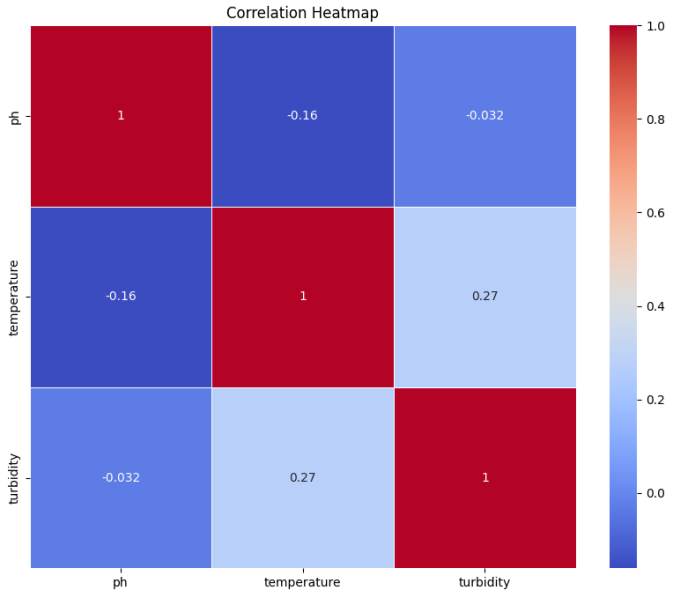

Fig. 7. Confution matrix in Hitmap

weakly correlated with both the temperature and turbidity, with correlation coefficients of 0.2 and 0.032, respectively.

The color of the cells in the heatmap indicates the strength of the correlation between the two variables. Red cells indicate a strong positive correlation, blue cells indicate a strong negative correlation, and white cells indicate no correlation.

The heatmap also shows that the pH and temperature are negatively correlated, with a correlation coefficient of -0.16. This means that as the pH increases, the temperature tends to decrease.

Overall, the heatmap shows that the temperature and turbidity are the most closely correlated variables, while the pH is weakly correlated with both of them.

The image caption ”Correlation Heatmap” also confirms that the image is a heatmap of the correlation between the three variables.

```
# Balanced dataset -->
                                    Model
Accuracy Score (%)
98.31                         Random Forest
98.03                         Decision Tree
91.55                                   KNN
50.99                   Gaussian Naive Bayes
47.89                 Support Vector Machines
42.82                     Logistic Regression

# Unbalanced dataset -->
                                    Model
Accuracy Score (%)
81.76                         Decision Tree
81.08                         Random Forest
63.51                                   KNN
40.54                   Gaussian Naive Bayes
35.81                 Support Vector Machines
33.78                     Logistic Regression
```

Fig. 8. Comparison among classification model

Figure 8 shows that RF gives the highest score of every model as accuracy 98.31%. The second highest score belongs to the Decision tree model which tells accuracy as 98.03%, There are 6 rows in the table in the image, split into two groups by the header text ”Balanced dataset” and ”Unbalanced dataset”.

The first 3 rows show the accuracy scores of three models on a balanced dataset: Random Forest: 98.03%, Decision Tree: 91.55%, KNN: 50.99% The next 3 rows show the accuracy scores of the same three models on an unbalanced dataset:

Decision Tree: 81.76%, Random Forest: 81.08%, KNN: 40.54% The remaining 4 models (Gaussian Naive Bayes, Support Vector Machines, and Logistic Regression) have accuracy scores below 47% on both datasets.

Overall, the Random Forest and Decision Tree models perform the best on both datasets, with accuracy scores above 81%. The KNN model performs worse, with accuracy scores below 51% on both datasets. The remaining 4 models perform the worst, with accuracy scores below 47% on both datasets.

V. Web Application

A. Used technologies

HTML : or HyperText Markup Language, is a crucial coding language used in web development. It serves as the

Fig. 9. Used technologies

standard markup language for creating and structuring web content. HTML employs a system of tags enclosed in angle brackets to define the structure and layout of web pages. These tags, such as html, body, p, a, and img are used to format text, create paragraphs, add links, and insert images, among other functions.

SCSS : The term SCSS is an acronym for Sassy Cascading Style Sheets. It is basically a more advanced and evolved variant of the CSS language. Natalie Weizenbaum and Chris Eppstein created it, and Hampton Catlin designed it. It comes with more advanced features- thus often called Sassy CSS.

JavaScript : JavaScript (JS) is a versatile and widely-used programming language primarily employed in web development. It enables the creation of interactive and dynamic content on websites. Unlike HTML and CSS, which focus on structure and design, JavaScript is responsible for adding behavior and functionality to web pages. It can be used to respond to user actions, manipulate the Document Object Model (DOM), and make web applications more engaging and responsive.

Python : Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming

Flask : Flask is a lightweight and versatile web framework for building web applications in Python. It is designed to be simple, easy to learn, and provide the essentials for creating web applications while allowing developers the freedom to add components as needed. Flask follows the WSGI (Web Server Gateway Interface) specification, making it compatible with various web servers and other web development tools.

VS code : Visual Studio Code, also commonly referred to as VS Code, is a source-code editor made by Microsoft with the Electron Framework, for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git..

B. Interfaces

In figure 8, in web interface there is a from. We can predict one and possible fish species based on ph, temperature and turbidity values. The output shown into right side of the interface. We use random forest model for processing.

In figure 9, in web interface there is a from. We can predict the aquatic environment like ph, temperature and turbidity values based on one fish species. The output shown into right side of the interface. We use random forest model for processing.

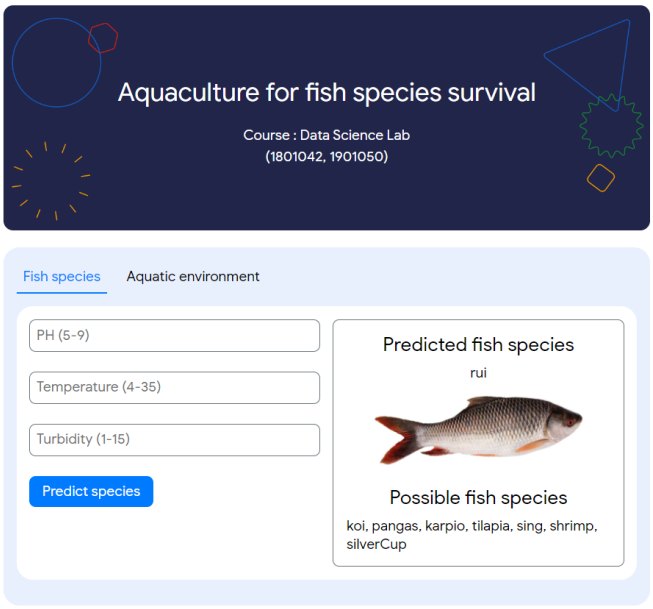

Fig. 10. Fish prediction based on parameter

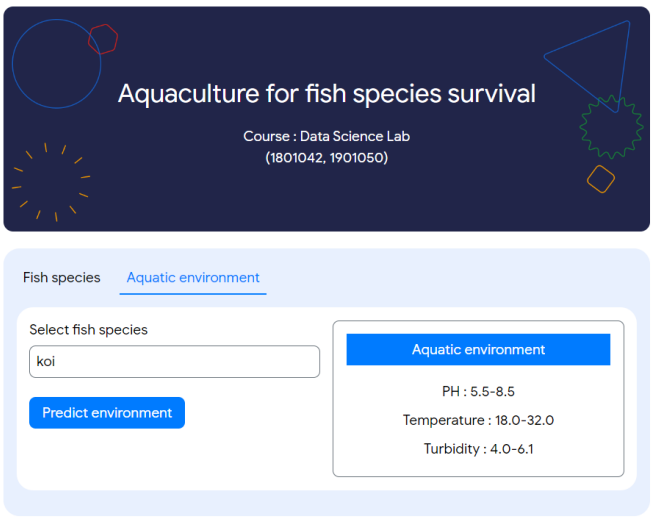

Fig. 11. Aquatic environment prediction based on Fish species

VI. Reference

[1] FAO, The State of World Fisheries and Aquaculture 2020-sustainability in action, 2020, FAO, Italy, 2020, pp. 0-244, http://www.fao.org/3/ca9231en/CA9231EN.pdf.

[2] M. K. Alsmadi, K. B. Omar, S. A. Noah, and A. I. Almarashdeh, "Fish Recognition Based on Robust Features Extraction from Size and Shape Measurements Using Neural Network," Journal of Computer Science, vol. 6, no. 10, pp. 1088-1094, 2010, doi: 10.3844/jcssp.2010.1088.1094.

[3] T. H. Hoang, K. Lock, A. Mouton, and P. L.M.Goethals, “Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam,” Ecological Informatics, vol. 5, no. 2, 2010, pp. 140-146, 2010, doi: 10.1016/j.ecoinf.2009.12.001.

[4] B. Benson, J. Cho, D. Goshorn, and R. kastner, “Field programmable gate array (FPGA) based fish detection using haar classifiers,” American Academy of Underwater Sciences, Georgia, USA, 2009, pp. 1-8.

[5] S. Bermejo, “Fish age classification based on length, weight, sex and otolith morphological features,” Fisheries Research, vol. 84, no. 2, pp. 270-274, 2007, doi: 10.1016/j.fishres.2006.12.007.

[6] A. G. Cabreira, M. Tripode, and A. Madirolas, “Artificial neural networks for fish-species identification,” ICES Journal of Marine Science, vol. 66, no. 6, pp. 1119-1129, 2009, doi: 10.1093/icesjms/fsp009.

[7] I. Khan, X. Zhang, M. Rehman, and R. Ali, “A Literature Survey and Empirical Study of Meta-Learning for Classifier Selection,” IEEE Access, vol. 8, pp. 10262-10281, 2020, doi: 10.1109/ACCESS.2020.2964726.

[8] J. Hu, D. Li, Q. Duan, Y. Han, G. Chen, and X. Si, “Fish species classification by color, texture and multi-class support vector machine using computer vision,” Computers and Electronics in Agriculture, vol. 88, pp. 133-140, 2012, doi: 10.1016/j.compag.2012.07.008.