IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 March 2024, accepted 28 March 2024, date of publication 3 April 2024, date of current version 26 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3384263

RESEARCH ARTICLE

IoT-Based Smart Biofloc Monitoring System for Fish Farming Using Machine Learning

MUHAMMAD ADEEL ABID[1], MADIHA AMJAD[2], KASHIF MUNIR[2], HAFEEZ UR REHMAN SIDDIQUE[1], AND ANCA DELIA JURCUT[3], (Member, IEEE)

[1]Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[2]Institute of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[3]UCD School of Computer Science, University College Dublin, Dublin 4, D04 C1P1 Ireland

Corresponding authors: Anca Delia Jurcut (Anca.jurcut@ucd.ie) and Muhammad Adeel Abid (adeel.abid@kfueit.edu.pk)

This work was supported in part by University College Dublin and in part by the Irish Research Council Project CHIST-ERA-22-SPiDDS-07.

ABSTRACT Biofloc technology assists in increasing the sustainability of fish farming by reusing and recycling waste water. However, its sophisticated operation makes it very sensitive to environmental conditions. A slight disturbance in one or more parameters can lead to high fish mortality and loss. IoT systems provide an efficient way of closely monitoring the biofloc to avoid catastrophe. The best aqua conditions vary depending on the fish. Therefore, there is a strong need to explore ideal conditions for different fishes. In this work, we have focused on Tilapi fish in the southern Punjab region to find the most suitable parameters. We have developed an IoT solution for monitoring Biofloc and gathering data. We have used low-cost sensors in our product to make it feasible for poor fish farmers. Multiple machine learning algorithms such as decision trees, random forest, support vector machine, logistic regression, Gaussian naive Bayes, XGBoost and ensemble learning are applied to the collected dataset to effectively predict mortality. Our analysis exhibits that the random forest and XGBoost achieved 98% accuracy in estimating mortality. The union of IoT, machine learning, and affordability positions our study at the forefront of advancing sustainable aquaculture practices in southern Punjab, Pakistan.

INDEX TERMS IoT automation, Biofloc, machine learning, mortality of fish, prediction.

I. INTRODUCTION

Malnutrition is becoming a more and more challenging issue with the growing population. There is a major thrust towards improving crops and livestock to increase their productivity. An increase in crop production is achieved through gene mutation and improved insect killer production. Meat production is increased through improved poultry and animal farming in an advanced research-oriented environment. Fish is another significant source of meat [1] that cannot be neglected. Gifted tilapia (gene mutation), shrimps, stinging catfish, dombra, rahu, malli, thella, and pangasius are primarily produced in the south Punjab region of Pakistan. Traditional methods like a simple pond and cage farming are adopted for fish production. Common problems in fish farming include a large amount of space, water and feed measurement, PH levels, nitrogen, dissolved oxygen, and ammonia amounts in water.

In contrast to traditional farming, tank culture fish farming [2] provides an opportunity to grow a large number of fish in a limited space. Figure 1 shows a Biofloc water tank. However, fish farming has a major challenge of environmental impacts and excessive water requirements.

Biofloc fish farming [3] is the latest technology for low-cost and sustainable fish production. One acre pond fish can be produced using a 16-diameter and 6 feet depth Biofloc tank. A large amount of fish production is made in a small Biofloc tank, and feed cost is saved through using probiotics that convert fish waste and remaining fish feed into protein, which is undoubtedly helpful in minimizing the cost of fish feed [4]. A flow-in and flow-out pipe are installed to change the water of the Biofloc tank. For oxygen, an oxygen pump and other techniques are used to maintain the oxygen level of the Biofloc tank. About 10 liters (gradually increasing to

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva.

20 liters with the growth of fish) of water is needed for a fish for proper growth in a Biofloc tank. Different Probiotics [5] are available that convert fish waste into protein that can be reused.

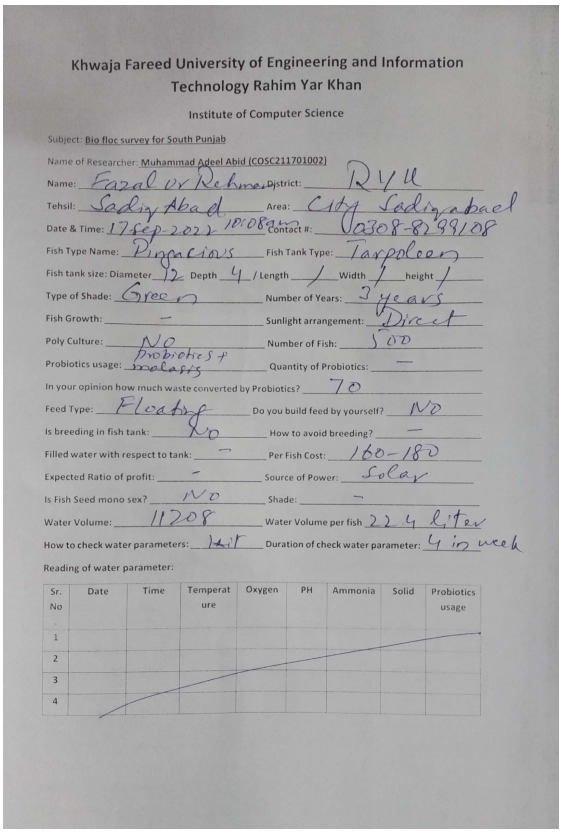
A kit is available to measure the Biofloc tank's PH, dissolved oxygen, solid waste, ammonia, and nitrogen water level [6]. Manual work is required to measure these values. A person is needed to control and check the Biofloc after short intervals and make adjustments accordingly. Maintaining the appropriate quality of water is very crucial in Biofloc fish farming.

FIGURE 1. Biofloc Water tank.

Water quality is disturbed due to the disturbance/ change in the PH level of water, a decrease of dissolved oxygen (DO), an increase of solid waste, an increase in the ammonia level of the tank, and an increase in the nitrogen level of water [7]. Furthermore, if any of the pumps, such as the water or oxygen pump, stop working or the probiotic's recommended quantity is not used. Water quality badly affects fish growth and may lead to mortality of the fish on a large scale.

South Punjab is a backward area of Punjab province and needs help with the availability of necessities of life like food, drinking water and health facilities, etc. According to Mr. Arshad (Assistant Director, Fisheries Department Rahim Yar Khan), about 14 Biofloc units are installed in district Rahim Yar Khan. The life cycle starts by putting fish seed in the month of February and harvest in November and December.The water quality of Biofloc tanks is maintained through the workforce, and some persons are deputed permanently to Biofloc tanks to check the water quality. A survey on Biofloc owner's problems is also performed, and a survey sample is shown in Figure 2. The survey concluded that there is need of an automatic system is needed that can predict bad water quality before 1-2 hours of expected fish mortality based on PH, DO, solid waste, ammonia, and nitrogen level. This is due to the fish farmers faced difficulty in managing the water quality parameters. The manpower is need to check the water quality parameter every time. In case of disturbance in water quality parameters, mortality of the fish is reported on a large scale. This will help minimize the mortality rate of fish in the Biofloc fish tank and beneficial for fish owners. It will also decrease the loss factor and increase the fish production per annum.

IoT (Internet of Things) provides best-automated solutions to real-world problems [8], [9], [10]. IoT contributes to


Khwaja Fareed University of Engineering and Information Technology Rahim Yar Khan

Institute of Computer Science

Subject: Bio floc survey for South Punjab

Name of Researcher: Muhammad Adeel Abid (COSC211701002)

Name: Fazal ur Rehman District: RYK

Tehsil: Sadiq Abad Area: City Sadiqabad

Date & Time: 17 Sep 2022 10:08 Contact #: 0308-8299108

Fish Type Name: Pangasious Fish Tank Type: Tarpolean

Fish tank size: Diameter 12 Depth 4 / Length / Width / height /

Type of Shade: Green Number of Years: 3 years

Fish Growth: – Sunlight arrangement: Direct

Poly Culture: NO Number of Fish: 500

Probiotics usage: Probiotics + molasis Quantity of Probiotics: –

In your opinion how much waste converted by Probiotics? 70

Feed Type: Floating Do you build feed by yourself? NO

Is breeding in fish tank: No How to avoid breeding? –

Filled water with respect to tank: – Per Fish Cost: 160-180

Expected Ratio of profit: – Source of Power: Solar

Is Fish Seed mono sex? NO Shade: –

Water Volume: 11208 Water Volume per fish 22.4 liter

How to check water parameters: kit Duration of check water parameter: 4 in week

Reading of water parameter:

Sr. No	Date	Time	Temperature	Oxygen	PH	Ammonia	Solid	Probiotics usage
1								
2								
3								
4								

FIGURE 2. Survey sample.

fields like home automation, smart cities, smart manufacturing, industrial automation, Agricultural automation, and wearable. IoT comprises communication technologies like RFID, ZIG-BEE, BARCODE, Bluetooth, NFC, WIFI, and many sensors [11]. Biofloc fish farming is productive as well as sensitive. It needs to be concisely monitored, since a minor mistake may lead to the mortality of fish on a large scale, and the setup may undergo loss. IoT can help with real-life problems, and it will be beneficial in solving this problem as well [12]. For developing countries like Pakistan, a cost-effective local IoT-based biofloc monitoring solution is required. The feed requirements and sensitivity of fish to different levels of nitrogen, ammonia, and other parameters vary for different species of fish, therefore, it is important to detect optimal values of different parameters for the local fish and calibrate the system to alarm accordingly before the mortality of the fish. Therefore, in this work, we have focused on designing a solution for southern Punjb region of Pakistan for Tilapi fish. In particular our contributions in this paper are as following.

- Solar powered low cost IoT-based automation solution is designed and implemented.
- Extensive data related to water quality parameters have been observed and recorded for over 1.5 months (data collected in May, June 2023-considered most crucial time period in fish growth life cycle).
- Multiple machine learning algorithms such as decision trees, random forest, support vector machine, logistic

regression, Gaussian naive Bayes, and XGBoost have been calibrated for the given data and compared to find the best model that provides highest accuracy in early prediction of mortality depending upon water quality parameters.
- correlation between different water quality parameters are identified.

The challenges faced related to IoT,Energy efficiencies, data rate, outage probability and any other challenges of this research are as follows.

- Minimum cost effective solution is needed and this is handled by using minimum number of sensors so that cost will be decrease.
- The exact calibration of the different sensors according to the environment is required and is achieved by checking and measuring with different instruments.
- Due to the limitation of Thingspeak cloud server storage (8200/hour entries), 2 minutes interval of the data storage is selected.
- The data collection for experiments is required for that period that is most crucial for fish growth/mortality and period selected for data collection is May and June.

It is the most cost effective solution for the sothern punjab region. as it cost nearabout $60 while the cost of the fish seed cost (500 Pangasius) is $174. The union of IoT and machine learning results in an effective system that can be used over time and can contribute in reducing the mortality of the fish in a cost effective way.

The rest of the paper is organized as follows. Section II illustrates a literature review of related work done in the field of Biofloc automation. Section III depicts the experimental setup of the IoT solution for automating Biofloc. This section describes the hardware, software, and methodology adopted for the completion of this research. Section IV presents results and finally we conclude our paper in Section V.

II. LITERATURE REVIEW

There are a number of researchers who focused research on Biofloc fish farming. Most of them researched and proposed solutions for their local Biofloc fish farming industry. Most significant researchers' work are discussed here. Goswami et al. [13] presented a complete description of the equipment and things that are needed to build an ideal Biofloc System. The Authors also suggest a cost-saving devices description to build a Biofloc water tank. Nikhita Rosaline et al. [14] presented an IoT-based solution for Aquaculture monitoring. They used a temperature sensor, Ammonia sensor, Dissolved Oxygen sensor, Salinity sensor, water level sensor, Nodemcu ESP12E controller, Wi-Fi controller, and web server for sending SMS. The research focuses on the growth of shrimp, fish, and other water animals. Saha et al. [15] focus on the automation of BioFloc using IoT. They used their respective sensors to focus on pH, dissolved oxygen, temperature, and water level monitoring. They used Arduino Mega as the main board and displayed all the results on LCD. The acid solution, base solution, air pump, raw salt solution, distilled water, and water pump for controlling the environment. MQTT protocol was used for this research, and a mobile app was also built to display results.

Islam [16] proposed a prediction system based on IoT and KNN. He targeted eight fish species: Silver Carp, Tilapia, Pangas, Sing, Koi, Rui, Prawn, and Katla. They used Arduino UNO and ethernet shields with the internet. Thingspeak server stores the sensor value so that KNN is applied to the collected data for training and testing the proposed model. The proposed model predicts the worst conditions. The Mq7 sensor for observing CO (carbon monoxide) was also used in the experimental setup. Mahmuda et al. [17] presented an IoT-based water quality system for Biofloc water tanks based on image processing. They used Raspberry Pi, a camera, a chemical pump, a water pump, a temperature sensor, a humidity sensor, and a pH sensor. Furthermore, image processing was used to measure the dissolved oxygen and ammonia level. IFTTT was used in this research for sending emails. The main aim of this research is to predict water quality using low costs sensors. Arefin et al. [18] proposed an IoT-based water quality management system based on a regression tree. The results were displayed on the mobile app. Dissolved Oxygen, Nitrogen, pH, temperature, Nitrate, Ammonia, and Carbon Dioxide are observed through sensors. Google Firebase cloud service was used for storing readings obtained from sensors. The proposed system attained an accuracy of 79%, which is quite acceptable. Rahid [19] provided a prediction of the water quality of Biofloc tanks using IoT and machine learning models for Bangladesh Biofloc fish farming. The author used Arduino UNO, pH sensor, temperature sensor, TDS sensor, and Neural network for training the proposed model. Three months of data are used for this purpose, and data contains Date, pH, temperature, TDS, NH3, and Floc quantity. Python 3.8, with the support of tensor flow, Keras, Pandas, Matplot, Numpy, and Boxplot, was used for programming. The author obtained 77.3 accuracy of the proposed model regarding the IoT Biofloc solution. Blancaflor et al. [20] worked on remote water quality management. They divided the solution into three layers. Application layer (mobile app), middleware layer (API, cloud database), and physical layer (pH sensor, temperature sensor, dissolved oxygen sensor, feed feeder, heater, fan, motor). Ten respondents were involved in validating the results. He obtained non-functional and functional mean ratio scores of 3.65 and 3.82, respectively regarding water quality management.

Blancaflor et al. [21] assessed an automated IoT-based Biofloc water quality system for the growth and mortality of Litopenaeus vannamei. The author concluded that water quality had a 10% higher survival rate. He divided that solution into three layers: Application layer, Middleware layer, and physical layers. For the experimental purpose, he uses Arduino ATMEGA 2560, Wi-Fi routers, cloud services, pH sensor, temperature sensor, DO sensor, feeder, solar power, heater, fan, motor, etc. Ahmed et al. [22]

provided a real-time water quality system for the Biofloc water tank. They observed values and noted the readings in a dataset with a 1-hour gap. They used an ESP32 IoT Wi-Fi controller board, temperature sensor, pH sensor, TDS sensor, water level sensor, servo motor, water pump, heater, more fantastic fan, oxygen pump, water filter, and LCD for displaying results. They also observed that if the temperature increases, then Dissolved oxygen decreases. Furthermore, if pH increases, then ammonia rises, and if pH decreases, then ammonia ion is converted into NH4+ (ammonia ion) and OH- (hydroxyl ion). Ahammad et al. [6] suggested a feeding system for Biofloc technology with the support of a GSM device. The proposed method is wholly based on the pH and temperature values of the Biofloc water tank. The researchers used Arduino, a servo motor (for feeding), LED (for display results), GSM, pH sensor, and temperature sensor.

Goswami et al. [13] proposed an intelligent system for Biofloc fish tanks in Bangladesh. They used a pH sensor, TDS sensor, Electrical conductivity, temperature sensor, Recirculation Aquaculture System (RAS), and solar system as a power source. An image processing technique was applied to calculate the weight of the fish. Thermal scanner app for scanning the fish and calculating the approximate weight. Prakosa et al. [23] focused on the acidity level monitoring system of Biofloc tank water. They used a pH sensor, Arduino, Database, LCD, and Thermo hydrometer to monitor the Biofloc acidity. Blancaflor et al. [24] focused on cost and profit comparison. The researcher explored the economic impact of solar power water quality management systems for Biofloc water tanks. The authors found that ROI (Return on Investment) for the target investors and fish farmers is 112% and 103.47%, respectively. He performed a market survey for Biofloc, and 83.3% of respondents thought that water quality plays a vital role in the survival and growth of the fish.

Due to the different levels of sensitivity and feed requirements of the various species of fish, it is necessary to analyze and determine critical values of different water-related parameters specific to particular fish. There is no significant work related to automating biofloc in Pakistan, especially in southern Punjab has been reported. Moreover, the previous work on determining the impact of different values on mortality achieves less accuracy. Our research in this paper focuses on these research gaps.

III. SMART BIOFLOC MONITORING SYSTEM

We have built a custom IoT-based smart biofloc system. For this, we have used cheap sensors so that the system is affordable for poor fish farmers. We have used this system to collect the dataset specifically for fish that are local to the south Punjab region of Pakistan. We have used sensors to monitor the quality of water.

Figure 3 describes the IoT automation [25] of the water quality prediction system of the Biofloc tank and is initiated by capturing values via sensors from the Biofloc water tank. Microcontroller boards and sensors like mq-7, mq-135,pH, TDS, DHT11, and temperature are used to accomplish the task. Gas sensors like mq-7 and mq-135 sensors are used to measure ammonia and carbon monoxide gas in the biofloc water tank [26]. PH sensor is used to detect the PH of the water because PH of the water is the most sensitive and dominant factor of the water quality system. TDS sensor identifies the Dissolved solids in the Biofloc water tank. DHT11 sensor is used to measure the external temperature and humidity of the environment. The temperature sensor measures the temperature of the biofloc water tank as fish growth is greatly affected by poor temperature (too hot water or too cold). All these sensors are connected to Arduino UNO.

Further, for power supply, solar with backup is used to power the sensors, boards, and others. This is done because we need to continuously monitor the water quality of the Biofloc water tank. If power fails, then our automation system will be turned off, which is not bearable at any cost. So, a separate power supply that works on solar power in the daytime and can have 16 hours of backup time at night is used as well [27].

An ESP8266 NodeMCU WIFI microcontroller is connected to Arduino UNO because of the availability of the internet everywhere nowadays. This WiFi module connects the Arduino UNO and sensors to the Cloud to save values periodically. All sensor values are collected and treated as raw data. Then preprocessing techniques are applied to raw data to clean the data. Feature extraction is performed on the data obtained after preprocessing. The obtained dataset is further sliced into training and testing sets. Varioumachine algorithms were trained using training data and compared. Further, different parameters are also identified that affect the fish's mortality much more than other parameters. The accuracy of machine learning algorithms has been evaluated using different evaluation parameters with the help of testing data. Data visualization and early alarming systems are done via Web applications and Android apps. Early prediction is made before the mortality of the fish. Complete details of hardware architecture, cloud server, and applied machine learning on the sensors data are discussed below.

A. HARDWARE ARCHITECTURE

Water Quality parameters of the Biofloc fish tank are collected through different sensors connected to the Arduino UNO microcontroller board and NodeMCU ESP8266 WIFI board [28]. The sensors used in our system are the MQ-7 sensor, MQ-135 sensor, TDS sensor, Turbidity sensor, DHT11 sensor, water temperature sensor, and DHT11 sensor. A description of the microcontroller board and sensors is given below.

1) ARDUINO UNO

Arduino UNO is the most used microcontroller board used in IoT projects, and it is based on the ATmega328P microcontroller board. It contains six analog pins and

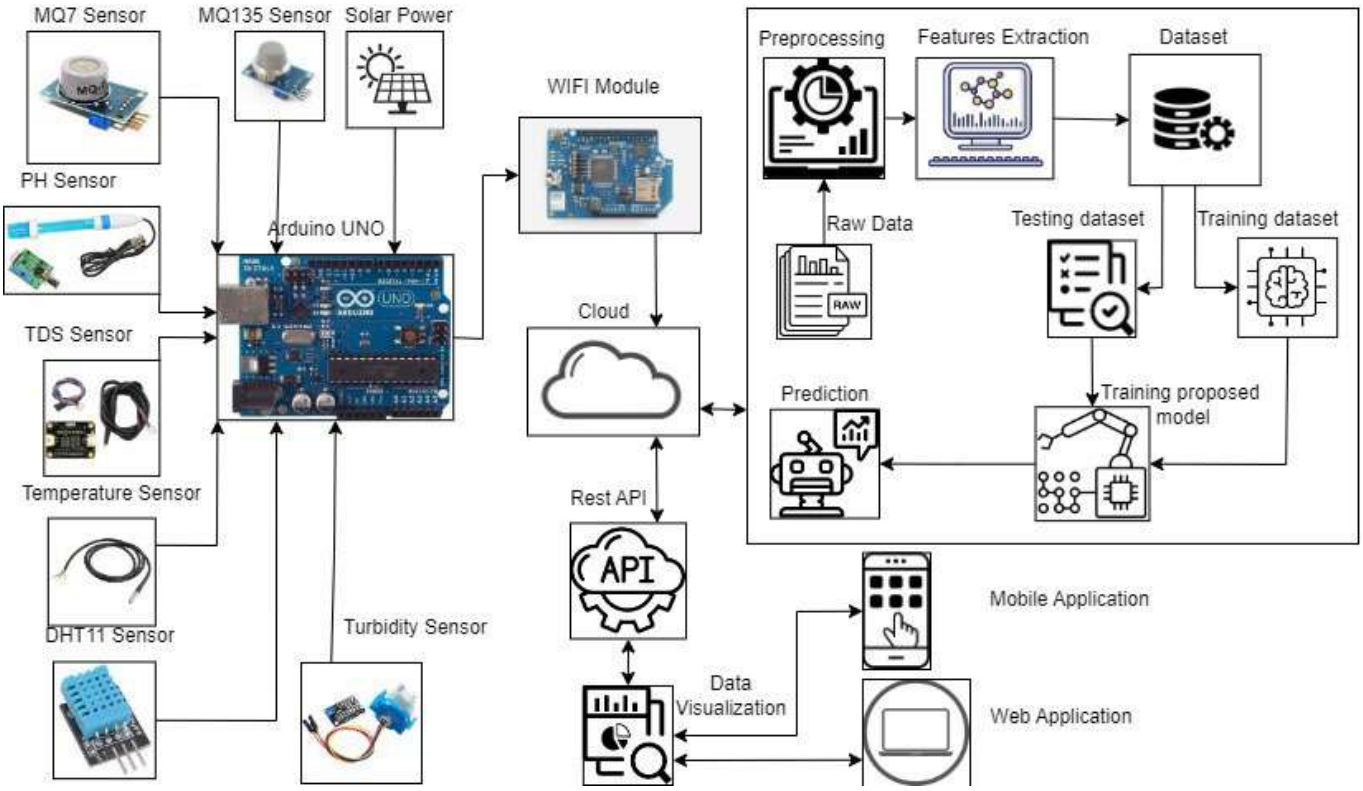


FIGURE 3. Smart Biofloc Monitoring System.

14 analog pins. It works on 5V. Furthermore, it has a USB connection, a reset button, and a power jack.

2) NODEMCU ESP8266 WIFI

NodeMCU ESP8266 WIFI microcontroller is a small board with a WIFI module. It operates on 3.3V and is connected via micro-USB. It contains one analog pin and 11 digital pins.

3) MQ-7 SENSOR

MQ-7 sensor is used to record the CO gas emitted in the surroundings. It helps maintain the air quality. Its measurement ranges from 10-10,000 ppm. This sensor takes 5V for operation and is connected via an Analog pin to the microcontroller board.

4) MQ-135 SENSOR

MQ-135 sensor is used to check the air quality of the environment [29]. MQ-135 is used to detect Ammonia, sulfur, Benzene, CO2, smoke, and other harmful gases. In this research, this sensor is used to measure Ammonia gas. Its measurement ranges from 10-1000ppm (Hydrogen, smoke, ammonia gas, and toluene).

5) TDS SENSOR

TDS stands for Total Dissolved Solid, which calculates the dissolved solids in water. TDS sensors are widely used in aquaculture environments because they can accurately measure TDS. Its measurement ranges from 0 to 1000 PPM.

6) TURBIDITY SENSOR

A turbidity sensor is used for checking water quality. It measures the light intensity scattered by suspended particles in water. The turbidity level of water increases with the increase of TSS (total suspended solids). Its measurement ranges from 0 to 4000 NTU.

7) DHT11 SENSOR

DHT11 sensor is handy as it measures the temperature and humidity in the environment. The temperature ranges of the DHT11 sensor are $-20°C$ – $60°C$ and humidity ranges 5 – 95% RH.

8) WATER TEMPERATURE SENSOR

The water temperature sensor is used to measure the temperature of the water. It measures the behavior and response of aquatic animals concerning temperature. The water temperature sensor ranges from $-5°C$ – $+50°C$.

9) PH SENSOR

PH sensor is used to calculate the value of the pH of water [30]. It ranges from 0 to 14, where seven is considered neutral/good drinking water.

Figure 4 shows the practical implementation of proposed system.

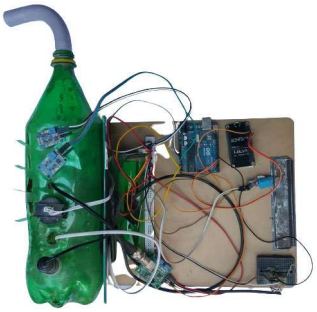


FIGURE 4. Practical Implementation.

B. CLOUD SERVER

Cloud servers like Thingspeak, Microsoft's Azure, Google's Cloud, and Amazon's AWS are available. For the proposed system, the Thingspeak cloud server is used as it is free and supports 8200 values of storage per day in the free version. The channel named "BioFloc Data Collection" with Channel ID "2081173" is created to store the sensor's values. Sensors are set up to take readings from the Biofloc water tank after 2 minutes, and then these values are stored on the cloud server via NodeMCU ESP8266 WIFI module. There are several Cloud services [31] available for storing the sensor data, but the Thingspeak Server is used to accomplish the storage of sensors. Figure 5 shows the IDE of the Thingspeak cloud server.

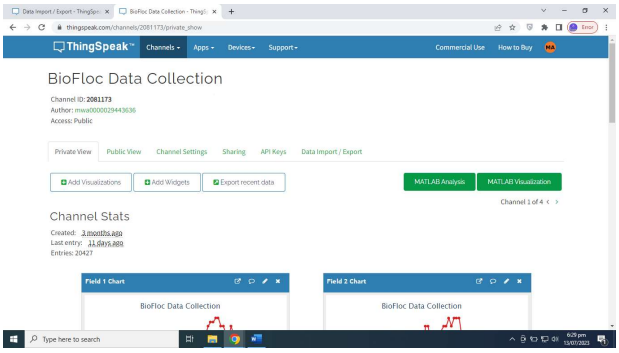


FIGURE 5. Thingspeak cloud server IDE.

C. DATA COLLECTION

The first setup of the Biofloc water tank is installed to collect datasets used for this study. Tilapia is the most common fish in the south punjab region and is used for this research. Gas sensors, PH sensors, TDS sensors, Temperature sensors, and other supporting devices/microcontrollers are installed on the Biofloc water tank to collect data [16], [22]. Arduino UNO board is interfaced to the WIFI module to send data to a cloud server. Data is collected for 1.5 months (13-May-2023 to 02-July-2023) with an interval of 2 minutes, and 19878 values are collected against each sensor.

Table 1 shows the statistics of the dataset used to accomplish this research article. There is a collection of values under different categories. Date & Time is the date and time at which the value from the sensor is received by

TABLE 1. Data Collection Summary.

Sr. No.	Category	Total Values collected	Values after Preprocessing
1	Date & Time	18978	18342
2	CO level	18978	18342
3	Ammonia level	18978	18342
4	Humidity	18978	18342
5	Turbidity	18978	18342
6	External Temperature	18978	18342
7	Total Dissolved Solid	18978	18342
8	pH value	18978	18342
9	Water Temperature	18978	18342
10	Mortality	18978	18342

the Thingspeak cloud server [32]. CO stands for Carbon monoxide present in water. Ammonia level is the value obtained from the MQ135 sensor that measures the amount of ammonia present. Humidity and external temperature of the environment are recorded via the DHT11 sensor as the Biofloc fishpond is also affected by the surrounding environment [33]. Turbidity refers to the cleanliness of water and is collected via the turbidity sensor. TDS stands for Total Dissolved Solid that is present in water and managed using a TDS sensor [34]. The pH of the water ranges from 1 to 14, and 7 is considered the most suitable water for drinking. pH value of the water of the Biofloc fish tank is collected via a pH sensor.

D. DATA VISUALIZATION

This section illustrates the dataset in visual form collected for this research. Overall, 18978 values against each sensor are collected on the Thingspeak cloud server. After preprocessing, 18342 values remain for further experiments, and 3% values are omitted during preprocessing. Figure 6 shows the dataset statistics after applying preprocessing.

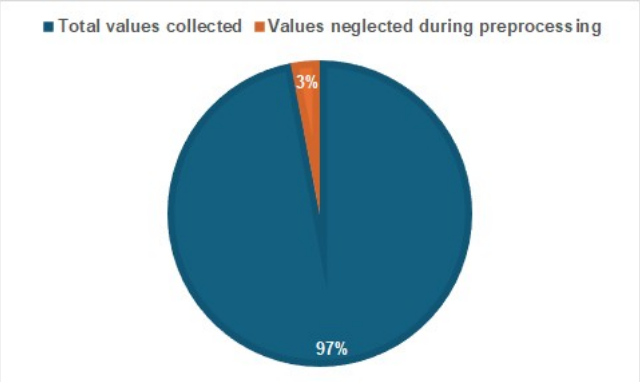


FIGURE 6. Statistics of dataset used for experiment purpose.

E. DATA PREPROCESSING

The dataset collected for experimenting contains numeric values. Still, some missing values or outliers in the dataset need to be sorted out before the experiments are done [35].

Figure 7 Shows the complete steps that are included in preprocessing. *Import Libraries:* The library in any language contains the functions and methods that support different operations in that language. The first step of preprocessing is importing libraries that are necessary to perform necessary procedures related to preprocessing of the dataset. Without importing libraries, we can't complete the task whose definition is defined in the specific library. *Read Dataset:* The next step is to load the dataset upon which experiments would be performed. Pandas' library supports loading the dataset on which investigations would be completed. *Checking for Missing Values:* The library in any language contains the functions and methods that support different operations in that language. The first step of preprocessing is importing libraries that are necessary to perform necessary procedures related to preprocessing of the dataset. Without importing libraries, we can't complete the task whose definition is defined in the specific library. *Checking of Categorical Data:* The next step is checking the categorical column whether they are present in the dataset. If some flat column is current, then encoding is performed. All the columns in the dataset are numeric are contains values that are not categorized except mortality, which is organized and needs to be determined based on the importance of the other parameters. *Standardize the Data:* Standardizing the data is the next step in preprocessing, in which scaling is performed so that all dataset features are well-tuned and the accuracy of machine learning algorithms is not affected due to different columns contain different ranges of values and need to be standardized before splitting the data. In the case of the dataset used for this research, columns other than "created_at," "entry_id," and "Mortality" are numeric, and scaling is applied to the columns in preprocessing. *Data Splitting:* Data splitting of the dataset is significant because it is needed to train the machine learning algorithms, and then testing data is used to evaluate the results. The dataset under observation is split into different ratios like 60%-40%, 70%-30%, 80%-20% and 90%-10% is used to test the machine learning algorithms [36]. *Implementation of Preprocessing:* All the experiments on the dataset are performed on a Core i7 10th generation laptop with Windows 10 as the operating System. Jupiter Notebook and Python are used as IDE and language to accomplish the experiments.

Table 2 shows the the sample values of the dataset. Here some deals are missing in columns due to the calibration time of the sensors, and Water Temperature Column consists of `''\r\n''` and some special characters that need to be cleaned. So, the first null values from all the datasets are dropped. Then each column is checked thoroughly, and if some special characters are there and were removed. Mainly Water Temperature column is focused as it contains `''\r\n''` in each cell that needs to be removed. After removing and cleaning the data, columns name CO, Ammonia, Humidity, Turbidity, External Temperature, TDS, pH, and Water Temperature are converted to float data type. The mortality column contains values of "Yes" and "No." The "Yes" is converted to 1, and the "No" value is converted to 0. Then Mortality column is converted to int data type. Scaling is the next step that is performed next. Table 3 shows the sample values of the dataset after preprocessing. *ADASYN:* ADASYN stands for Adaptive Synthetic Sampling, that is the modification of SMOTE for dealing and adjusting the weights for minority class [37]. It is beneficial as it deals with the minority class as equity with major class. This helps in better training the proposed model and facilitates in obtaining the desired results.

1) MACHINE LEARNING ALGORITHMS

Different numbers of machine learning algorithms are used for classification and clustering purposes. This study uses decision trees, random forests, support vector machines, logistic regression, Gaussian Naïve Bayes and XGBoost. Different train-test splits of the dataset are used to train tmachine learning models [38]. Description of machine learning classifiers along with the calculation formula is given below.

a: DECISION TREE

A decision tree is a famous machine-learning algorithm that makes decisions based on a group of features in a hierarchy [39], [40]. It undergoes a recursive process of splitting the data into different subsets and creating a tree model. On each node of the decision tree, the decision is made based on some specific feature and value, creating a model of the tree. The decision defines which side of the tree would be followed. The process under specific criteria continuous unit maximum depth is achieved. A decision tree is the simplest structure and can be understood and visualized easily. They can handle both numerical and categorical values and are robust against outliers. They used appropriate strategies to address missing values with imputation. When the tree becomes deep, it can be overfit. To solve this issue, Random forests combine different trees to reduce overfitting and improve performance.

b: RANDOM FOREST

Random Forest is an ensemble technique that combines many decision trees [41]. Different Bootstrap samples are used to train ml algorithms. Subsampling of datasets is done to get a bootstrap sample where the sample size is the same as the training dataset size. Random Forest can be calculated as in Equation 1. The Random Forest prediction equation is given by [42]:

$$P = \text{majority}\,(T_1(y), T_2(y), T_3(y), \ldots, T_n(y)) \qquad (1)$$

where,

- p is the calculated prediction based on majority decision trees.
- T1(y), T2(y), T3(y) …, Tn(y) shows the number of decision trees participating in the prediction process.

FIGURE 7. Preprocessing Steps.

TABLE 2. Sample Values of dataset before preprocessing.

created_at	entry_id	CO	Ammonia	Humidity	Turbidity	External Temperature	TDS	pH	Water Temperature	Mortality
2023-06-04T15:44:01+05:00	9372	379.43	0.19	64	33.0	25.0	67.22	6.68	29.19	No
2023-06-09T16:54:58+05:00	14353	367.21	2.80	50	40.0	29.0	64.71	8.29	34.38	No
2023-05-21T11:31:38+05:00	4503	471.08	8.43	47	38.0	13.0	1006.48	13.02	33.75	Yes
2023-06-07T05:20:06+05:00	11323	325.67	0.52	75	27.0	30.0	70.96	9.01	29.13	Yes

TABLE 3. Sample Values of dataset after preprocessing.

created_at	entry_id	CO	Ammonia	Humidity	Turbidity	External Temperature	TDS	pH	Water Temperature	Mortality
2023-06-04 T15:44:01 +05:00	9372	-0.120878	-0.846868	0.444554	-0.298511	0.110959	-0.578350	-1.100943	-0.983682	0
2023-06-09 T16:54:58 +05:00	14353	-0.351051	-0.056054	-0.800198	1.094541	0.702740	-0.584524	-0.411247	1.123773	0
2023-05-21 T11:31:38 +05:00	4503	1.605416	1.649801	-1.066930	0.696526	-1.664385	1.732025	1.615002	0.867955	1
2023-06-07 T05:20:06 +05:00	11323	-1.133487	-0.746880	1.422574	-1.492556	0.850686	-0.569151	-0.102812	-1.008046	1

c: SUPPORT VECTOR MACHINE(SVM)

SVM (Support Vector Machine) is a famous classification and pattern recognition technique [43]. It controls high-dimensional datasets by calculating a hyperplane that maximizes the separation margin between classes, thus minimizing the error in the category. However, when faced with overlapped data, SVM's performance is highly affected in classification. SVM is a supervised machine learning model designed to solve binary classification problems. In most situations, SVM proves to be fruitful and produces better accuracy.

d: LOGISTIC REGRESSION

Logistic Regression works with several independent variables to produce separate values. It calculates the probability of each class present in the dataset. Due to this, it is considered a good classifier that is used for categorical data. It finds and works on the association between dependent and independent variables. Equation 2 illustrates the Logistic Regression formula [44].

$$P(Y = 1) = \frac{1}{1 + e^{-(m(x - v_0)}} \tag{2}$$

where,

- e represents Euler Number.
- vo represents x-value of sigmoid midpoint.
- L represents the maximum value of the curve.
- m represents the curve's steepness.

e: GAUSSIAN NAÏVE BAYES

Gaussian Naïve Bayes is based on the Bayes theorem and changes pretty differently from other algorithms as all the features in this classifier are independent [45]. It is used for classification purposes for objects having customarily distributed data. Due to its characteristics, it is known as Gaussian Naïve Bayes. Equation 3 and 4 shows the formula to calculate the Gaussian Naïve Bayes [46].

$$P(c|x) = \frac{P(c) \cdot P(x|c)}{P(x)} \tag{3}$$

$$P(c|x) \propto P(c) \cdot P(x|c) \tag{4}$$

where,

- P(c | x) represents the target class's posterior probability
- P(c) represents the class's prior probability.
- P(x | c) represents the predictor class's posterior probability.
- P(x) represents predictor's prior probability.

f: XGBOOST

XGBoost or Extreme Gradient Boosting is an efficient gradient boosting algorithm used as a machine learning algorithm [47]. It can handle high and complex dimensional

datasets and perform classification and regression. XGBoost combines multiple weak models and performs an iterative process where each model corrects the mistakes made by its previous model. This certainly improves the accuracy of the proposed model and makes it a more robust model as compared to others. The main strength of XGBoost is dealing with numerical and categorical features and missing values in the dataset. It also uses a regularization technique to avoid overfitting in the training process. It can handle large, big datasets and perform fast activity and predictions.

F. EVALUATION PARAMETERS

Testing data is applied to check whether it is correctly trained. For testing purposes, many evaluation parameters are available. We evaluated precision, recall, and F1 score for various machine learning algorithms to select the most suitable for predicting fish mortality.

1) PRECISION

Precision [48] is used to measure and calculate the correctness of a model. The formula used to calculate the precision is as below [49].

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5}$$

2) RECALL

Recall [50] deals with the completeness of the classifiers. Recall calculated as the total no. of true positives divided by the addition of no. of true positives and no. of false negatives. The formula to calculate the Recall is given below [49].

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

3) F1 SCORE

The F1 score is another evaluation parameter calculated from Recall and Precision value. It takes the Precision and Recall value and then finds the harmonic mean between them. Its values range from 0 to 1. The formula to calculate the F1 score is described below [51].

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

4) ACCURACY

Accuracy is used to measure and calculate the correctness of the target classifiers. Its values range from 0 to 1. Equation 3 illustrates the formula for the calculation of accuracy. Formula of accuracy for Binary Classification is given below [52].

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Samples}} \tag{8}$$

where,

- True Positive = The algorithm predicted mortality and real value also shows mortality.
- True Negative = The algorithm model predicted no mortality and the real value also shows no mortality.
- False Positive = The algorithm model predicted no mortality and the real value also shows mortality.
- False Negative = The algorithm model predicted mortality and the real value also shows no mortality.

IV. RESULTS AND DISCUSSION

This section illustrates the results of the experiment and evaluation parameters results. IoT sensors, along with the micro controller boards, are installed on the Biofloc water tank. Thingspeak cloud server saves the data sent from the sensors via the internet. The dataset is imbalanced, therefore, a sampling technique named ADASYN balances the dataset. Machine Learning algorithms random forest, decision tree, Support Vector Machine (SVM), logistic regression, XGBoost, and Gaussian Naïve Bayes are trained for improved accuracy. Ensemeble learning for all the above mentioned algorithms was also performed. Evaluation parameters precision, recall, f1 score, and accuracy are used to evaluate the accuracy of machine learning models. ADASYN is performed on the dataset to deal with the unbalancing of the class. Machine learning models are trained and tested with the different percentages using 60-40, 70-30, 80-20, and 90-10 train-test split of the dataset to check the performance of the machine learning algorithms at different proportions of the training dataset. Table 4 shows the results of machine learning algorithms on the dataset's 60-40 train test split. Random forest, decision tree, and XGBoost offer better accuracy of 97% compared to others. The support vector machine shows a marginally low accuracy of 94% after the random forest, decision tree, and XGBoost. Gaussian naïve Bayes offers lower accuracy, precision, recall, and f1 score values than other machine learning classifiers [53].

TABLE 4. Results of machine learning classifiers using 60–40 train-test split of dataset.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Random Forest	97	95	99	97
Decision Tree	97	95	99	97
Support Vector Machine	94	89	99	94
Logistic Regression	93	93	89	97
XGBoost	97	95	99	97
Gaussian Naïve Bayes	91	89	92	90
Ensemble Learning	96	94	99	96

Table 5 illustrates the results of machine learning classifiers on a 70-30 train-test split of the dataset. Random forest shows a significant accuracy of 98% among all the machine learning classifiers. The random forest offers a better performance in terms of precision, recall, and f1 score as compared to others. XGBoost secured 2nd place in accuracy percentage among the machine learning classifiers. Support vector machine and logistic regression show marginally lower accuracy than XGBoost. Gaussian naïve Bayes shows the

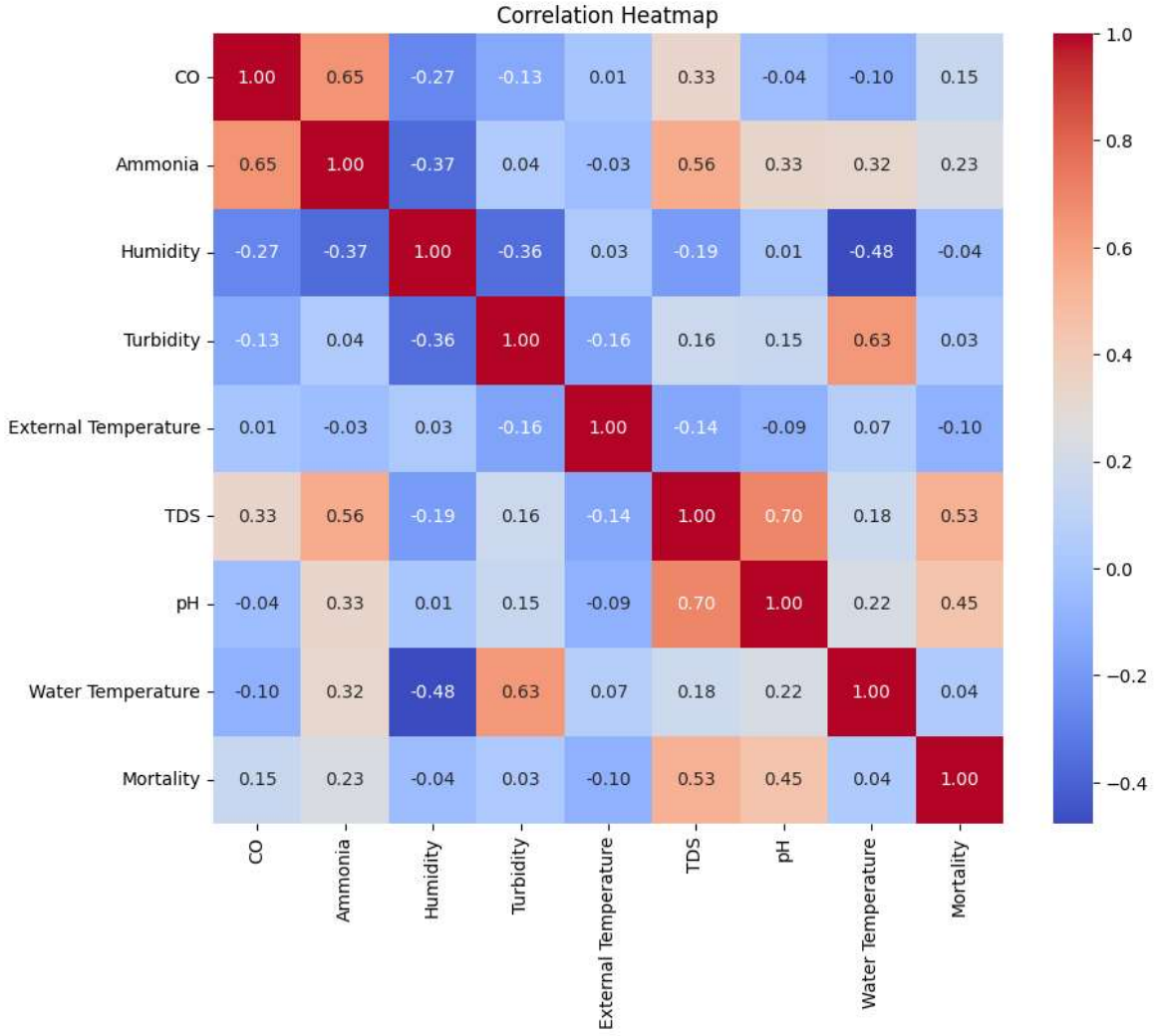


FIGURE 8. Correlation matrix between water quality parameters and mortality.

weakest results in accuracy, precision, recall, and f1 score among all the classifiers.

TABLE 5. Results of machine learning classifiers using 70–30 train-test split of dataset.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Random Forest	98	97	99	98
Decision Tree	96	96	96	96
Support Vector Machine	95	91	99	95
Logistic Regression	94	94	91	97
XGBoost	97	96	99	97
Gaussian Naïve Bayes	91	90	92	91
Ensemble Learning	97	95	99	97

Table 6 shows the evaluation parameters results of machine learning classifiers on the dataset's 80-20 train-test split. Random forest and XGBoost outperformed with 97% accuracy. A marginal low accuracy is observed in the y decision tree and support vector machine classifier. Logistic regression stands second last among all the machine learning classifiers regarding accuracy. Gaussian naïve Bayes shows the lowest 90% accuracy for the dataset's 80-20 train-test split.

TABLE 6. Results of machine learning classifiers using 80–20 train-test split of dataset.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Random Forest	97	96	98	97
Decision Tree	95	95	95	95
Support Vector Machine	95	93	98	95
Logistic Regression	93	93	91	94
XGBoost	97	96	98	97
Gaussian Naïve Bayes	90	91	89	90
Ensemble Learning	97	95	98	97

Table 7 shows the results of machine learning classifiers for the dataset's 90-10 train-test break. Random forest and XGBoost offer an extraordinary 97% accuracy among all the machine learning classifiers. The decision tree classifier shows marginally low accuracy compared to random forest and XGBoost. Gaussian naïve Bayes offers 89% accuracy, recorded as minimum accuracy among all the machine learning classifiers.

TABLE 7. Results of machine learning classifiers using 90–10 train-test split of dataset.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Random Forest	97	97	97	97
Decision Tree	96	95	97	96
Support Vector Machine	93	90	97	93
Logistic Regression	92	92	89	95
XGBoost	97	97	97	97
Gaussian Naïve Bayes	89	87	91	89
Ensemble Learning	94	90	99	94

Random forest and XGBoost machine learning classifiers show better accuracy than others in all four train-test splits of the dataset. The percentage of train-test breaks that are usually considered good is 70-30 and 80-20 because both contain a significant amount of training data to train a model and have enough datasets to perform and test the accuracy of a model.

Figure 9. shows the accuracy of machine learning classifiers using different train-test splits of the dataset. XGBoost offers a consistent accuracy of 97% in all the train-test divisions of the dataset. The random forest also shows better accuracy and remains equal to XGBoost. The decision tree shows marginally low accuracy compared to random forest and XGBoost.

Gaussian Naive Bayes offers the most insufficient accuracy in all dataset train-tests. Evaluating which water quality parameter affects the fish's mortality is also essential. Because there are various water quality parameters, some are sensitive while others are not sensitive enough to be cared for. The correlation matrix is essential as it clearly shows the correlation between different parameters and gives a better idea of dependent variables. Figure 8 shows the correlation [54] between water quality parameters and the mortality of the fish. Carbon monoxide has a strong association with the Ammonia level of water and is correlated with the TDS of the water as well. Ammonia is strongly associated with CO and TDS, pH, and water temperature, and it also impacts fish mortality. Humidity in the environment has a loose relationship with external temperature. Turbidity of the water relates to water temperature and has a weak association with pH, TDS, and mortality. TDS is strongly associated with ammonia, pH, mortality, and CO. pH of the water is highly correlated with TDS and mortality. The mortality of

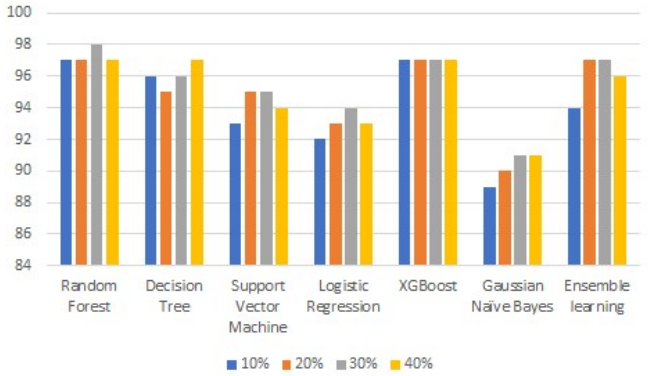


FIGURE 9. Accuracy of machine learning classifiers on different train-test splits of the dataset.

the fish strongly depends on the value of TDS and pH and slightly lower association with the ammonia level of the water. Furthermore, mortality is weakly associated with the CO, turbidity, and water temperature of the Biofloc water tank. A new fish farmer can adopt the proposed IoT based solution that focuses on Ammonia level, TDS and pH level that are considered most important factors in the fish growth. The sensors installed on the Biofloc water tank continuously taking readings. It can help a fish farmer by predicting the worst/bad water conditions before 1-2 hours that can lead to the mortality of the fish.

V. CONCLUSION AND FUTURE WORK

Biofloc with tank culture is an attractive solution to increase fish production in a limited space. Recycling and reusing of water in Biofloc results in decreased environmental impact and water consumption. Besides the benefit of Biofloc, due to its sophisticated operation, Biofloc is very sensitive to the changes in the quality of water. A slight change in the quality of water may lead to the mortality of fish on a large scale. Moreover, these parameters depends on the type of fish in the water tank. Therefore, it is important to design an IoT Solution along with the machine learning algorithm for the specific fish to warn the alarming situation. In this article, we have designed and implemented an IoT system for an effective continuous monitoring of Biofloc. Using the system, data is collected for Tilapia fish in southern Punjab, Pakistan. Multiple machine learning algorithms such as decision trees, random forest, support vector machine, logistic regression, Gaussian Naïve Bayes, XGBoost and Ensemble learning are applied. Evaluation parameters named precision, recall, f1 score, and accuracy are evaluated to estimate mortality. Among all the models, random forest and XGBoost show better accuracy of up to (98%). This significance of this accuracy is that the system is 98 percent accurate in identifying the conditions of fish mortality and this is effective in building an early alarming system for fish farmers before the mortality of fish occurs. In the future, we intend to use our system to collect data for different species of fish and train model for their efficient monitoring.

We will also intend to evaluate ensemble learning to train the model and get better results.

REFERENCES

[1] D. J. McClements and L. Grossmann, "The science of plant-based foods: Constructing next-generation meat, fish, milk, and EGG analogs," *Comprehensive Rev. Food Sci. Food Saf.*, vol. 20, no. 4, pp. 4049–4100, Jul. 2021.

[2] Y. I. Chu, C. M. Wang, J. C. Park, and P. F. Lader, "Review of cage and containment tank designs for offshore fish farming," *Aquaculture*, vol. 519, Mar. 2020, Art. no. 734928.

[3] M. H. Khanjani, M. Sharifinia, and S. Hajirezaee, "Recent progress towards the application of biofloc technology for tilapia farming," *Aquaculture*, vol. 552, Apr. 2022, Art. no. 738021.

[4] M. H. Khanjani and M. Sharifinia, "Biofloc technology as a promising tool to improve aquaculture production," *Rev. Aquaculture*, vol. 12, no. 3, pp. 1836–1850, Aug. 2020.

[5] A. Zabidi, F. M. Yusoff, N. Amin, N. J. M. Yaminudin, P. Puvanasundram, and M. M. A. Karim, "Effects of probiotics on growth, survival, water quality and disease resistance of red hybrid tilapia (*Oreochromis* spp.) fingerlings in a biofloc system," *Animals*, vol. 11, no. 12, p. 3514, Dec. 2021.

[6] M. B. Ahammed, S. Sultana, A. Sarkar, and A. Momin, "pH and temperature monitoring with a GSM-based auto feeding system of a biofloc technology," *Int. J. Sci. Eng. Res.*, vol. 13, no. 4, pp. 270–274, 2022. [Online]. Available: http://www.ijser.org

[7] E. O. Ogello, N. O. Outa, K. O. Obiero, D. N. Kyule, and J. M. Munguti, "The prospects of biofloc technology (BFT) for sustainable aquaculture development," *Sci. Afr.*, vol. 14, Nov. 2021, Art. no. e01053.

[8] R. P. Singh, M. Javaid, A. Haleem, and R. Suman, "Internet of Things (IoT) applications to fight against COVID-19 pandemic," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, no. 4, pp. 521–524, Jul. 2020.

[9] A. Khanna and S. Kaur, "Internet of Things (IoT), applications and challenges: A comprehensive review," *Wireless Pers. Commun.*, vol. 114, no. 2, pp. 1687–1762, Sep. 2020.

[10] S. Y. Y. Tun, S. Madanian, and F. Mirza, "Internet of Things (IoT) applications for elderly care: A reflective review," *Aging Clin. Experim. Res.*, vol. 33, no. 4, pp. 855–867, Apr. 2021.

[11] H. Landaluce, L. Arjona, A. Perallos, F. Falcone, I. Angulo, and F. Muralter, "A review of IoT sensing applications and challenges using RFID and wireless sensor networks," *Sensors*, vol. 20, no. 9, p. 2495, Apr. 2020.

[12] A. Badshah, A. Ghani, A. Daud, A. Jalal, M. Bilal, and J. Crowcroft, "Towards smart education through Internet of Things: A survey," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–33, Feb. 2024.

[13] N. Goswami, S. A. Sufian, M. S. Khandakar, K. Z. H. Shihab, and M. S. R. Zishan, "Design and development of smart system for biofloc fish farming in Bangladesh," in *Proc. 7th Int. Conf. Commun. Electron. Syst. (ICCES)*, Coimbatore, India, Jun. 2022, pp. 1424–1432.

[14] N. Rosaline and S. Sathyalakshimi, "IoT based aquaculture monitoring and control system," *J. Phys., Conf. Ser.*, vol. 1362, Nov. 2019, Art. no. 012071.

[15] K. K. Saha, A. Islam, S. S. Joy, I. Writwik, and K. Shikder, "Bio-floc monitoring and automatic controlling system using IoT," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2021, pp. 15–21.

[16] M. M. Islam, J. Uddin, M. A. Kashem, F. Rabbi, and M. W. Hasnat, "Design and implementation of an IoT system for predicting Aqua fisheries using Arduino and KNN," in *Proc. Int. Conf. Intell. Hum. Comput. Interact.*, vol. 12616, 2021, pp. 108–118.

[17] B. Mahmuda, E. Haque, A. Al Noman, and F. Ahmed, "Image processing based water quality monitoring system for biofloc fish farming," in *Proc. Emerg. Technol. Comput., Commun. Electron. (ETCCE)*, Dec. 2021, pp. 1–6.

[18] S. A. Mozumder and S. Sagar, "Smart IoT-biofloc water management system using decision regression tree," 2021, *arXiv:2112.02577*.

[19] M. M. Rashid, A.-A. Nayan, S. A. Simi, J. Saha, M. O. Rahman, and M. G. Kibria, "IoT based smart water quality prediction for biofloc aquaculture," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 470–477, 2021. [Online]. Available: www.ijacsa.thesai.org

[20] E. Blancaflor and M. Baccay, "Design of a solar powered IoT (Internet of Things) remote water quality management system for a biofloc aquaculture technology," in *Proc. 3rd Blockchain Internet Things Conf.*, Jul. 2021, pp. 24–31.

[21] E. B. Blancaflor and M. Baccay, "Assessment of an automated IoT-biofloc water quality management system in the litopenaeus vannamei's mortality and growth rate," *Automatika*, vol. 63, no. 2, pp. 259–274, Apr. 2022.

[22] I. Ahamed and A. Ahmed, "Design of smart biofloc for real-time water quality management system," in *Proc. 2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST)*, Jan. 2021, pp. 298–302.

[23] J. A. Prakosa, "Development of monitoring techniques and validation of the acidity level of biofloc pond water for optimizing tilapia aquaculture," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 1017, no. 1, Apr. 2022, Art. no. 012006.

[24] E. Blancaflor and M. Baccay, "Economic & operational impact analysis of a solar powered remote water quality management system designed for an indoor biofloc aquaculture setup," in *Proc. 5th Int. Conf. E-Soc., E-Educ. E-Technol.*, Aug. 2021, pp. 81–86.

[25] W.-S. Kim, W.-S. Lee, and Y.-J. Kim, "A review of the applications of the Internet of Things (IoT) for agricultural automation," *J. Biosyst. Eng.*, vol. 45, no. 4, pp. 385–400, Dec. 2020.

[26] K. B. K. Sai, S. Mukherjee, and H. P. Sultana, "Low cost IoT based air quality monitoring setup using Arduino and MQ series sensors with dataset analysis," *Proc. Comput. Sci.*, vol. 165, pp. 322–327, Jan. 2019.

[27] H. Agrawal, R. Dhall, K. S. S. Iyer, and V. Chetlapalli, "An improved energy efficient system for IoT enabled precision agriculture," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 6, pp. 2337–2348, Jun. 2020.

[28] D. R. Prapti, A. R. M. Shariff, H. C. Man, N. M. Ramli, T. Perumal, and M. Shariff, "Internet of Things (IoT)-based aquaculture: An overview of IoT application on water quality monitoring," *Rev. Aquaculture*, vol. 14, no. 2, pp. 979–992, Mar. 2022.

[29] K. B. K. Sai, S. R. Subbareddy, and A. K. Luhach, "IoT based air quality monitoring system using MQ135 and MQ7 with machine learning analysis," *Scalable Comput., Pract. Exper.*, vol. 20, no. 4, pp. 599–606, Dec. 2019.

[30] W. Indrasari, E. Budi, Umiatin, S. Rizqy Alayya, and R. Ramli, "Measurement of water polluted quality based on turbidity, pH, magnetic property, and dissolved solid," *J. Phys., Conf. Ser.*, vol. 1317, no. 1, Oct. 2019, Art. no. 012060.

[31] S. Kunal, A. Saha, and R. Amin, "An overview of cloud-fog computing: Architectures, applications with security challenges," *Secur. Privacy*, vol. 2, no. 4, p. e72, Jul. 2019.

[32] F. Khan, M. A. B. Siddiqui, A. U. Rehman, J. Khan, M. T. S. A. Asad, and A. Asad, "IoT based power monitoring system for smart grid applications," in *Proc. Int. Conf. Eng. Emerg. Technol. (ICEET)*, Feb. 2020, pp. 1–5.

[33] M. S. Novelan and M. Amin, "Monitoring system for temperature and humidity measurements with DHT11 sensor using nodeMCU," *Int. J. Innov. Sci. Res. Technol.*, vol. 5, no. 10, pp. 123–128, 2020.

[34] S. U. N. Goparaju, S. S. S. Vaddhiparthy, C. Pradeep, A. Vattem, and D. Gangadharan, "Design of an IoT system for machine learning calibrated TDS measurement in smart campus," in *Proc. IEEE 7th World Forum Internet Things (WF-IoT)*, Jun. 2021, pp. 877–882.

[35] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC Trends Anal. Chem.*, vol. 132, Nov. 2020, Art. no. 116045.

[36] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 15, no. 4, pp. 531–538, Aug. 2022.

[37] C.-C. Chang, Y.-Z. Li, H.-C. Wu, and M.-H. Tseng, "Melanoma detection using XGB classifier combined with feature extraction and K-Means SMOTE techniques," *Diagnostics*, vol. 12, no. 7, p. 1747, Jul. 2022.

[38] M. A. Abid, M. F. Mushtaq, U. Akram, M. A. Abbasi, and F. Rustam, "Comparative analysis of TF-IDF and loglikelihood method for keywords extraction of Twitter data," *Mehran Univ. Res. J. Eng. Technol.*, vol. 42, no. 1, p. 88, Jan. 2023.

[39] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemometrics*, vol. 18, no. 6, pp. 275–285, Jun. 2004.

[40] Y. Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, Apr. 2015.

[41] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[42] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Jun. 1995, pp. 278–282.

[43] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.
[44] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.
[45] S. Jayachitra and A. Prasanth, "Multi-feature analysis for automated brain stroke classification using weighted Gaussian Naïve Bayes classifier," *J. Circuits, Syst. Comput.*, vol. 30, no. 10, Aug. 2021, Art. no. 2150178.
[46] A. H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," in *Proc. Artif. Intell. Signal Process. Conf. (AISP)*, Oct. 2017, pp. 209–212.
[47] O. Sagi and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Inf. Sci.*, vol. 572, pp. 522–542, Sep. 2021.
[48] S. Anantharaj, S. R. Ede, K. Karthick, S. Sam Sankar, K. Sangeetha, P. E. Karthik, and S. Kundu, "Precision and correctness in the evaluation of electrocatalytic water splitting: Revisiting activity parameters with a critical assessment," *Energy Environ. Sci.*, vol. 11, no. 4, pp. 744–771, 2018.
[49] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[50] S. A. Khan and Z. Ali Rana, "Evaluating performance of software defect prediction models using area under precision-recall curve (AUC-PR)," in *Proc. 2nd Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2019, pp. 1–6.
[51] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implications for evaluation," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, Mar. 2005, pp. 345–359.
[52] J. Wahl, J. Freyss, M. von Korff, and T. Sander, "Accuracy evaluation and addition of improved dihedral parameters for the MMFF94s," *J. Cheminformatics*, vol. 11, no. 1, pp. 1–10, Dec. 2019.
[53] J. Heckman, J. L. Tobias, and E. Vytlacil, "Four parameters of interest in the evaluation of social programs," *Southern Econ. J.*, vol. 68, no. 2, p. 210, Oct. 2001.
[54] I. Archakov and P. R. Hansen, "A new parametrization of correlation matrices," *Econometrica*, vol. 89, no. 4, pp. 1699–1715, 2021.

MUHAMMAD ADEEL ABID received the M.C.S. and M.S.C.S. degrees from The Islamia University of Bahawalpur, Bahawalpur, in 2008 and 2015, respectively. He is currently pursuing the Ph.D. degree in computer science with the Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan.

He has been a Lecturer in computer science with the Khwaja Fareed University of Engineering and Information Technology, since April 2018, accumulating nearly 16 years of teaching experience in various educational institutions. His research interests include predictive analysis on numeric and textual data, pattern identification, and the development of data clusters based on dataset characteristics. Additionally, he specializes in IoT-based solutions addressing real-time problems. He received the Gold Medal during the M.C.S. studies and actively takes on various responsibilities assigned by the department and the university.

MADIHA AMJAD received the M.Sc. degree in electronics from Quaid-i-Azam University, Pakistan, in 2008, the M.S. degree in computer engineering from the University of Engineering and Technology, Pakistan, in 2011, and the Ph.D. degree from the National University of Sciences and Technology (NUST), Pakistan, in 2020. She is currently an Assistant Professor with the Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan. Her research interests include the design and optimization MAC layer schemes for hybrid VLC/RF networks, WSNs and molecular nanonetworks, resource optimization, clustering in unmanned vehicular area networks (UAV), and design of IoT-based solutions to solve indigenous problems.

KASHIF MUNIR received the B.Sc. degree in mathematics and physics from The Islamia University of Bahawalpur, Pakistan, in 1999, the M.Sc. degree in information technology from Universiti Sains Malaysia, in 2001, the M.S. degree in software engineering from the University of Malaya, Malaysia, in 2005, and the Ph.D. degree in informatics from the Malaysia University of Science and Technology, Malaysia, in 2015. He has been in the field of higher education, since 2002. After an initial teaching experience in courses with the Binary College, Malaysia, for one semester and with the Stamford College, Malaysia, for around four years, he later relocated to Saudi Arabia. He was with the King Fahd University of Petroleum and Minerals, Saudi Arabia, from September 2006 to December 2014. Then, he moved to the University of Hafr Al-Batin, Saudi Arabia, in January 2015. In July 2021, he joined the Khwaja Fareed University of Engineering and IT, Rahim Yar Khan, where he is currently an Assistant Professor with the IT Department. He has published journal articles, conference papers, book, and book chapters. His research interests include cloud computing security, software engineering, and project management. He has been in the technical program committee of many peer-reviewed conferences and journals, where he has reviewed many research papers.

HAFEEZ UR REHMAN SIDDIQUE received the B.Sc. degree in mathematics from Islamia University Bahawalpur (IUB), Pakistan, in 1998, the M.Sc. degree in computer science from Bahauddin Zakariya University (BZU), Multan, Pakistan, in 2000, and the Ph.D. degree in electronic engineering from London South Bank University, in April 2016. He was a Lecturer of computer science in network with the Institute of Computer Education (NICE), from 2001 to 2006. His research interests include biomedical and energy engineering applications, data recognition, image processing, system embedded programming, and machine learning.

ANCA DELIA JURCUT (Member, IEEE) received the B.Sc. degree in computer science and mathematics from the West University of Timisoara, Romania, in 2007, and the Ph.D. degree in security engineering from the University of Limerick (UL), in 2013. She has been an Assistant Professor with the UCD School of Computer Science, since 2015. She was a Postdoctoral Researcher with UL, a member of the Data Communication Security Laboratory, and a Software Engineer with IBM, Dublin, in the area of data security and formal verification. She has recently acted as an Evaluator of H2020 proposals for the cryptography and cybersecurity call. Her Ph.D. study was funded by the Irish Research Council for Science Engineering and Technology (IRCSET). Her research interests include security protocols design and analysis, mathematical modeling, automated techniques for formal verification, cryptography, computer algorithms, security for Internet of Things, and blockchain security. Much of her work has focused on formal verification techniques for security protocols using deductive reasoning methods (modal logics and theorem proving), automation of logics for formal verification, the development of new logic-based techniques and tools for formal verification, the design and analysis of security protocols, and formalization and modeling of design requirements for security protocols.