VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum-590018

A PROJECT REPORT **(15CSP85)** ON

"WATER QUALITY PREDICTION SYSTEM FOR AQUACULTURE"

Submitted in Partial fulfillment of the Requirements for the Degree of

Bachelor of Engineering in Computer Science & Engineering

By

MEGHANA S (1CR16CS089)

NEHA RANI SINGH(1CR16CS103)

NETHRAVATHI A S(1CR16CS104)

PALLAVI LAKSHMI H L(1CR16CS108)

Under the Guidance of,

Mr. KIRAN BABU T S

Associate Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"Water Quality Prediction System for Aquaculture"** carried out by Ms. Meghana S, USN 1CR16CS089, Ms. Neha Rani Singh, USN 1CR16CS103, Ms. Nethravathi A S, USN 1CR16CS104, Ms. Pallavi Lakshmi H L, USN 1CR16CS108, bonafide students of CMR Institute of Technology, in partial fulfillment for the award of **Bachelor of Engineering** in Computer Science and Engineering of the Visveswaraiah Technological University, Belgaum during the year 2019-2020. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library.

The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Mr. Kiran Babu T S	**Dr. Prem Kumar Ramesh**	**Dr. Sanjay Jain**
Associate Professor	**Professor & Head**	**Principal**
Dept. of CSE, CMRIT	**Dept. of CSE, CMRIT**	**CMRIT**

External Viva

Name of the examiners	Signature with date
1.	________________
2.	________________

DECLARATION

We, the students of Computer Science and Engineering, CMR Institute of Technology, Bangalore declare that the work entitled "**WATER QUALITY PREDICTION SYSTEM FOR AQUACULTURE**" has been successfully completed under the guidance of Prof. Kiran Babu T S, Computer Science and Engineering Department, CMR Institute of technology, Bangalore. This dissertation work is submitted in partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering in Computer Science and Engineering during the academic year 2019 - 2020. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place: Bengaluru

Date: 5/6/2020

Team members:

MEGHANA S (1CR16CS089) ____________________

NEHA RANI SINGH (1CR16CS103) ____________________

NETHRAVATHI A S (1CR16CS104) ____________________

PALLAVI LAKSHMI H L(1CR16CS108) ____________________

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany a successful completion of any task would be incomplete without the mention of people who made it possible, success is the epitome of hard work and perseverance, but steadfast of all is encouraging guidance.

We take this opportunity to express our sincere gratitude and respect to **CMR Institute of Technology, Bengaluru** for providing us a platform to pursue our studies and carry out our final year project.

We have a great pleasure in expressing our deep sense of gratitude to **Dr. Sanjay Jain,** Principal, CMRIT, Bangalore, for his constant encouragement.

We would like to thank **Dr. Prem Kumar Ramesh**, Professor and HOD, Department of Computer Science and Engineering, who shared his opinion and experience through which we received the required information crucial for the seminar.

We consider it a privilege and honour to express our sincere gratitude to our guide **Mr. Kiran Babu T S,** Associate Professor, Department of Computer Science & Engineering, for his valuable guidance throughout the tenure of this review.

Finally, we would like to thank all the faculty of Computer Science and Engineering my parents and friends for all their moral support they have given me during the completion of this work.

MEGHANA S (1CR16CS089)

NEHA RANI SINGH(1CR16CS103)

NETHRAVATHI A S (1CR16CS104)

PALLAVI LAKSHMI H L (1CR16CS108)

ABSTRACT

Indian fisheries and aquaculture are an important sector of food production providing nutritional security, besides livelihood support and gainful employment to more than 14 million people and contributing to agricultural export. Aquatic organisms have specific tolerant range of various environmental parameters, thus fish farming of specific types of fish species requires certain conditions that must be met. Water quality is a critical factor while culturing aquatic organisms. In this project, we have built an application using ML for predicting water parameters and in turn monitor the fish farming ponds. The quality of water is predicted in an hourly manner to ensure growth and survival of aquatic life. A web application is built using Flask to alert the user in critical situations. The impact of water parameter change can be handled productively if the information is examined and water quality is anticipated in advance.

Keywords: Aquaculture, Machine Learning, LSTM, Root Mean Square Error, Flask.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

ANN	**Artificial Neural Network**
ARIMA	**Auto Regressive Integrated Moving Average**
DO	**Dissolved oxygen**
LSTM	**Long Short Term Memory**
ML	**Machine Learning**
MSE	**Mean Square Error**
pH	**Potential of Hydrogen**
RMSE	**Root Mean Square Error**
RNN	**Recurrent Neural Network**
SRU	**Simple Recurrent Unit**
SSVM	**Smooth Support Vector Machine**

CHAPTER 1

INTRODUCTION

1.1 Aquaculture

Research in aquaculture is an input to increase stabilized production. In last decade various scientists have made sustained efforts that resulted in development of modern production technologies that have revolutionized farm production. Fish farming have been used for more than three decades. Fish farming refers to farming variety of marine species such as shellfish, sport fish, bait fish, ornamental fish, crustaceans, mollusks, algae, sea vegetables, and fish eggs to breed, rear and harvest in different water environments such as ponds, rivers, lakes, and ocean. Fish are cold-blooded animals, regulating their body temperature directly by the water environment. Changes in water temperature affect the amount of dissolved oxygen in the water and fish oxygen consumption. Although the fish can withstand a broad water temperature range, any sudden, extreme changes in water temperature will have a considerable impact on fish physiology. A chilling injury will cause the fish to rush into, paralysis with a loss of balance, leading to death. The reason may be the respiratory center, or osmotic regulation is affected at high temperatures. As the water temperature increases the fish suffer respiratory arrest.

Fish World magazine found that the amount of dissolved oxygen in water increases or decreases with the seasons. When the water temperature rises, fish metabolic rate will be increased and results in less dissolved oxygen in the water. Low water temperature decreases fish metabolic rate and increases amount of dissolved oxygen in the water. If the amount of dissolved oxygen in water is reduced to below a certain limit fish growth will be hindered. When the amount of dissolved oxygen becomes lower than the fish survival conditions the fish will die. In general fish farming the acidity and alkaline of the water should be maintained between 6 to 8. Too acidic or alkaline will cause adverse effects, acid erosion of the gill tissue, tissue coagulation necrosis, increased mucus secretion, abdominal congestion and inflammation. If the PH value is less than 4.5, the fish will die. Table 1.1 shows the ranges of parameter tolerance of the species.

Table 1.1 Water Quality Tolerance by Species

Species	Temp °F	Dissolved Oxygen mg/L	pH	Alkalinity mg/L	Ammonia %	Nitrite mg/L
Baitfish	60-75	4-10	6-8	50-250	0-0.03	0-0.6
Catfish/Carp	65-80	3-10	6-8	50-250	0-0.03	0-0.6
Hybrid Striped Bass	70-85	4-10	6-8	50-250	0-0.03	0-0.6
Perch/Walleye	50-65	5-10	6-8	50-250	0-0.03	0-0.6
Salmon/Trout	45-68	5-12	6-8	50-250	0-0.03	0-0.6

Aquaculture is the fastest growing food production sector according to research, hence, to consolidate its efficiency and its sustainability. Water quality monitoring is the key to realize the success of Aquaculture. Aqua farmers are relying on manual testing for knowing the parameters of water, which is time consuming and inaccurate as parameters may alter with time.

To overcome this problem, modern technologies like IOT and ML should be brought to aquaculture. Thus, increases productivity and minimizes the losses by constant monitoring of water quality parameters. Water parameters that is needed to by monitored continuously are Temperature, Dissolved Oxygen, PH, turbidity, Salt, etc.

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies or analyze the impact of machine learning processes.

1.2 Problem Statement

The goal is to predict the water quality accurately and efficiently in an hourly manner, and to alert the user in advance. Regression analysis and prediction method through a large number of data samples for correlation analysis to obtain the correlation, and establish a regression equation, on the basis of considering the prediction error to determine the future water quality prediction value. Its model is complex, and the distribution of data and samples requires higher requirements.

In spite of the fact that ARIMA models are very adaptable to time series, their major impediment is that no non-linear can be captured by the ARIMA model. LSTM model is picked so as to build up a far-reaching approach for effective water quality prediction and analysis. LSTM can dynamically memorize and retain the historical water quality parameters information while learning new information.

1.3 Objectives

Our main objective is to increase the productivity of aquatic fishes and save its life.

- To increase the productivity of fishes, efficiently.
- To predict the water quality accurately, ML Algorithms are used.
- Predict the water parameters in an hourly manner. i.e., predict next hour values based on previous data.
- When predicted water quality seems to exceed the critical conditions alert is sent to user immediately to take the precautionary actions required.
- Build a web application using Flask to show alerts to the users as well as data visualization of the various water parameters considered.

1.4 Scope of the project

- The project focusses on the importance of water quality for aquaculture. It helps the aqua farmers to produce good quality of fishes, which in turn helps the economy of aquaculture sector.
- Machine Learning helps provide a better, accurate and faster prediction of quality of water based on the previous data collected.
- Predictive analysis can help to capture relationships among numerous variables that can help to assess risk with a particular set of conditions. LSTM NN is considered better and gives increasingly precise data to foresee and assess the water quality.
- In India, freshwater fishes like Rohu and Catla are predominantly grown through Aquaculture.

1.5 Methodology

To complete this project successfully agile methodology as shown in Table 1.2 is followed.

- Sprint 1: Literature survey is done to understand other related works in this domain and hence do the analysis of problem in the existing model.

- Sprint 2: Identifying the project goals, project requirements and formulating the problem statement. Collection of dataset and preprocessing.
- Sprint 3: Developing the models with various ML algorithms. These models are trained, tested and evaluated.
- Sprint 4: Choosing the best model and predicting the water parameters in hourly manner.
- Sprint 5: Developing web application for displaying the predicted values and alert.
- Sprint 6: Implementation of data visualization for better understanding of the parameter chances over the past hours.

Table 1.2 Software Agile Methodology

Story ID	Requirement description	User stories/Task	Description
1	Collection of water data with various parameters.	Data Collection	In .csv format.
2	Determining the relevant parameters needed for the water quality.	Cleaning the dataset.	Removal of noisy data.
3	Selection of Algorithm	Designing the model.	Identifying the best algorithm with the data available.
4	Training the datasets.	Training, testing and evaluating the model	Training the model to predict the water quality.
5	Prediction of water quality on hourly basis.	Predict water quality	Water quality is predicted for next hour based past 6 hours
6	Testing the results.	Accuracy calculation.	Calculating the accuracy of algorithm.
7	Alert system and data visualization	To display alert message and graph.	Web application showing the predicted values and appropriate message. Changes of the parameters are shown in the form of graphs.

The following Figure1.1 shows the stages followed in the project for the successful implementation.

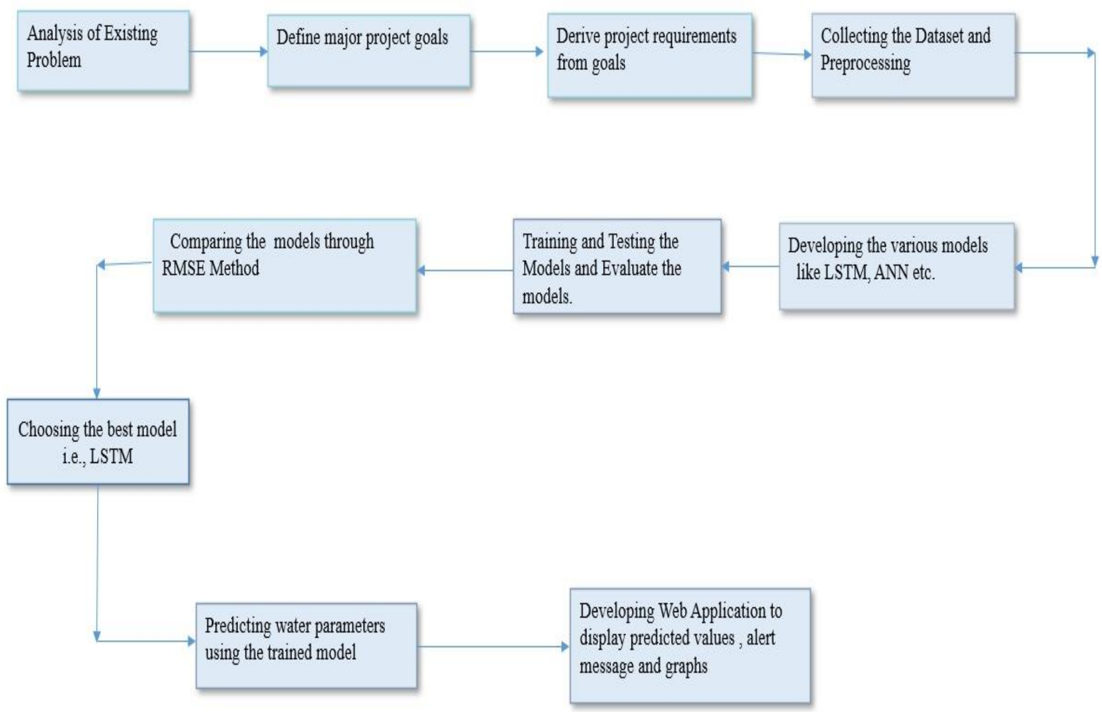

Figure 1.1 Methodology

1.6 Summary

In this chapter, we have discussed about what is aquaculture, its importance in the economy according to research. We have also discussed about the scope and the relevance of our project in the current trends. The main objectives of water quality testing in Aquaculture is discussed in this chapter.

CHAPTER 2

LITERATURE SURVEY

2.1 Related Work

Water quality prediction provides a significant reference for dynamic regulation of water quality and sudden events. After reading through the following research papers, we learnt about the importance of water quality in the aquaculture domain and its need to be monitored. The various parameters that will affect the water quality and their ideal ranges suitable for survival of fishes. The different methodologies used for water quality prediction.

In terms of water quality prediction, the main water quality prediction methods include time series method, Markov method, grey system theory method and support vector regression machine method. However, these methods have some problems, such as, the poor adaptive ability, low computational efficiency and inaccurate prediction results, which cannot meet the requirement in precision agriculture. In recent years, the methods based on artificial neural network have been proposed, which have the advantages of good robustness, high fault tolerance and sufficient fitting of complex nonlinear relations.

Juntao Liu et al.[1] focuses on the prediction of pH and water temperature parameters in key water quality parameters. Firstly, the water quality parameters are pre-processed by improved method. Then, the Pearson correlation coefficient method is used to find the correlation between the water quality parameters. Finally, the SRU (Simple Recurrent Unit) deep learning model is used to establish a prediction model for the key water quality parameters, so as to achieve accurate prediction. Meanwhile, we also evaluate the prediction effect of prediction model built by RNN (Recurrent Neural Network) deep learning network. This paper compares the working SRU and RNN and brings how SRU is better in terms of accuracy and how well SRU will fit the data compared to RNN. Figure 2.1 shows the flow of method followed for prediction.

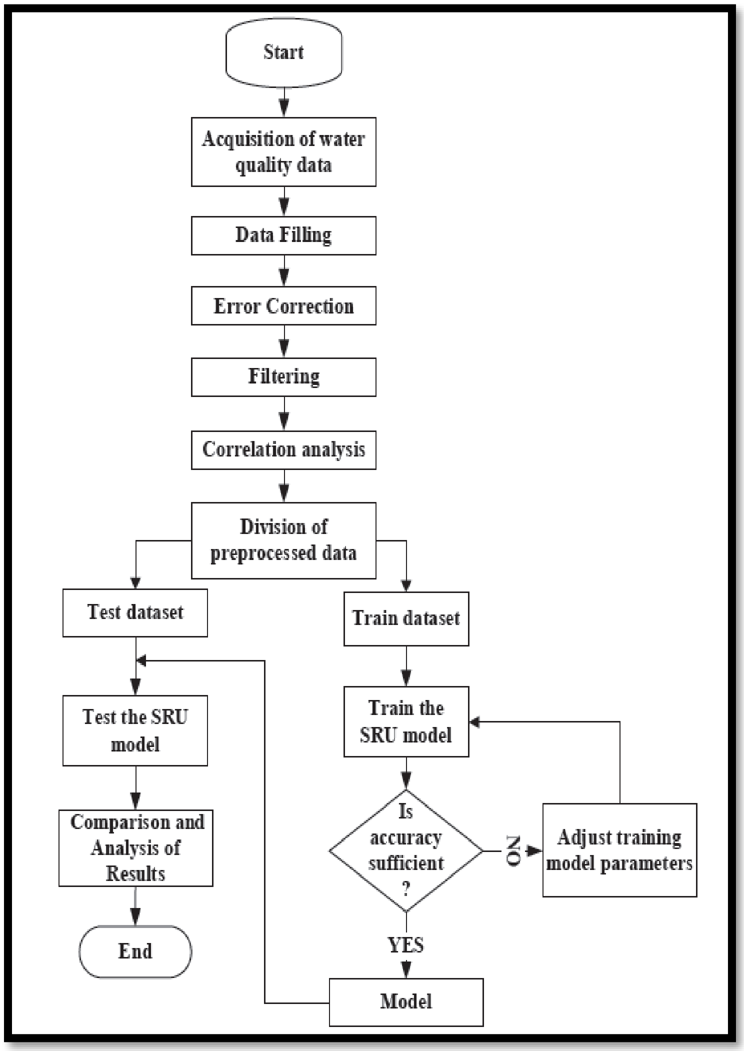

Figure 2.1 The Flowchart of SRU Model

Both models are close to the true value in prediction as seen in Figure 2.2, but it is obvious that the SRU performances a higher degree of fitting with the real value, so the SRU water quality parameter prediction model has better prediction result. The SRU adds a new unit states based on the structure of the RNN to control information in the network, and it is slightly redundant in the training time. In the pH prediction model, SRU consumed 16.929% more time than RNN. While in the water temperature prediction model, SRU takes 18.494% more time than RNN. It can be concluded that under the same conditions, SRU training is more time-expensive than RNN.

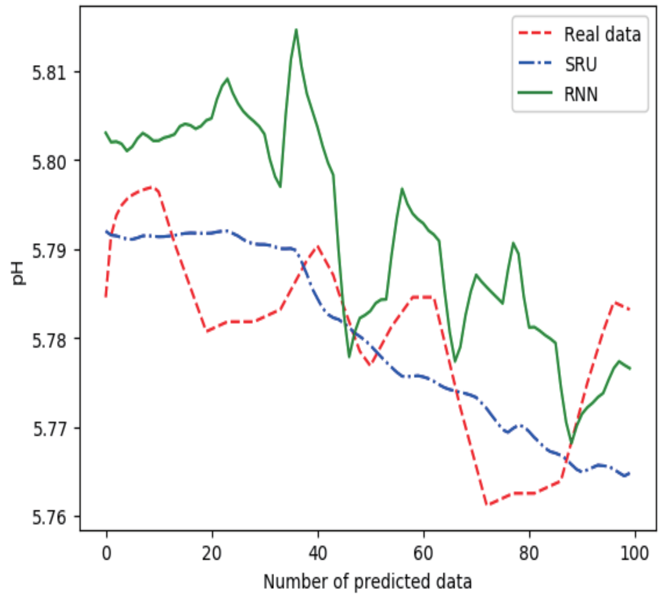

Figure 2.2 Comparation of predicted pH with real data

In modern intensive river aquaculture management, water quality prediction plays an important role. The water quality indicator series are nonlinear and non-stationer. Hence, the accuracy of the commonly used conventional methods, including regression analyses and neural networks, were limited. A prediction model based on Smooth Support Vector Machine (SSVM) is proposed by Wijayanti Nurul Khotimah[2] to predict the aquaculture water quality. SSVM is an algorithm that is used for solving no linear function estimation problems. The data used in this research are data of river in Surabaya collected for two years. The data have twenty variables that indicate water quality such as temperature, turbidity, color, SS, pH, alkalinity, free CO2, DO, Nitrite, Ammonia, Copper, phosphate, sulfide, iron, Hexavalent Chromium, Manganese, Zinc, Lead, COD, and Detergents. The Root Mean Square Error (RMSE) of the experiment is 0.0275. This value shows that SSVM proven to be an effective approach to predict aquaculture water quality.

Many researches to predict water quality has been done. Feifei Li et al. established short-term forecasting model to predict dissolved oxygen using backpropagation (BP) and autoregressive (AR) . Palani et al. developed a neural network model to forecast the amount of dissolved oxygen in seawater [6]. However, neural networks suffer from a few weaknesses, such as over-fitting. In this research, we propose SSVM to predict water quality (especially dissolve oxygen).

Nowadays, Support Vector Machine (SVM) becomes a method that is constantly evolving and increasingly popular in machine learning. This method has several advantages over other methods such as neural network. Those advantages are [10].

a. Assuring global optimum

b. Parameters that must be estimated relatively few

c. The model is stable

d. Relatively simple to use

e. Successfully applied in most real cases.

Many SVM developments proposed to improve its performance and efficiency. Lee and Mangasarian [11] have proposed a new formula of SVM with linear and non-linear kernel for classification using smoothing method. The method is called Smooth Support Vector Machine (SSVM). The basic concept of SSVM is changing primal SVM formula into non smooth optimization problem without constraint. Hence, the objective function of the optimization problem is not differentiable. Therefore, smoothing function was used in purpose to achieve differentiable objective function. This method was solved by Newton Armijo algorithm. By using the Newton-Armijo algorithm can be seen that the main difference between the smoothing approaches with SVM is that SSVM solve problem using linear equations, while SVM using quadratic programming problem.

The result of dissolved oxygen prediction using SSVM is shown in Figure 2.3. The figure shows the comparison between prediction value of this proposed method and actual value (original data). We used 5-fold for the experiment. The RMSE value of the prediction is 0.0275. This value shows that SSVM proven to be an effective approach to predict dissolved oxygen. The SSVM forecasting method to predict water quality can help avoid economic losses causes by water quality problems.

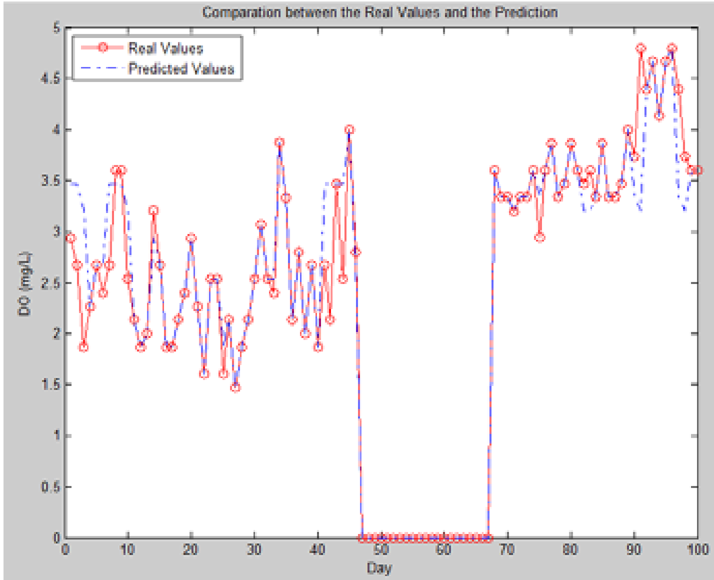

Figure 2.3 The Comparison between Predicted Value and Actual Value of DO

J. Wang et al.[3] investigates the characteristics of dynamic nonlinearity and correlation of water quality parameter information, as well as the gradient disappearance and gradient explosion caused by the training data of traditional RNN network model, etc. The long short-term memory network structure (LSTM) is introduced to optimize the structure of RNN network and the connection weight and threshold of hidden layer. A new water quality parameter prediction model of LSTM-RNN network based on improved RNN network structure is proposed by setting the number of storage units in the hidden layer of the network, the number of structural layers of the network model, and adjusting the time window size of the data training set. Combined with the water quality monitoring data of the River in Shanghai, the model is used to predict and verify the main pollutant index COD (potassium permanganate index) in the River. The simulation results show that compared with the traditional GM (grey model) and RNN network water quality prediction model, the sample approximation accuracy and generalization ability of the training prediction based on LSTM-RNN network model is higher and better than that of the traditional GM (grey model) and RNN network model. Good comprehensive prediction performance of river water quality is presented.

The commonly used methods are time series prediction, artificial neural network prediction method, regression analysis prediction method, grey system prediction method and so on. The time series prediction method belongs to the statistical prediction model with no cause variables. Its characteristic is that the mathematical theory foundation is perfect, the practice is more difficult, and the error is large. The prediction accuracy of the prediction model established by artificial neural network forecasting method can reach a satisfactory degree, but it is not clear from the prediction model that the specific reasons and internal relations of the water quality change trend can be clearly understood from the prediction model. Therefore, the prediction accuracy of the model is affected. Regression analysis and prediction methodthrough a large number of data samples for correlation analysis to obtain the correlation, and establish a regression equation, on the basis of considering the prediction error to determine the future water quality prediction value. Its model is complex, and the distribution of data and samples requires higher requirements. The grey system forecasting method can get regular time series by accumulating the raw data without regular rules once or more, and then establish a forecasting model to predict the water quality in the medium and long term. It is easy to be affected by data instability, and then will produce a large prediction error. Because water quality parameter is a dynamic time series, it is more suitable to use recurrent neural network (RNN). In addition, the prediction process of water quality parameters is gradual, that is to say, the current water quality parameters are correlated with the historical water quality parameters. This requires that RNN can dynamically memorize historical water quality parameter information and retain the historical water quality parameter information while learning new information. So, this paper introduces a long short-term memory network structure (LSTM) [12] to add the hidden layer of RNN.

The LSTM network is a variant type of RNN, which combines learning with model training without additional domain knowledge. The improved LSTM structure is helpful to avoid the problem of gradient disappearance and explosion in typical RNN. This means that LSTM has advantages in capturing long-term dependence and modelling nonlinear dynamics and can be used to deal with long length sequence data. The The generalization ability of LSTM-RNN model is stronger and the prediction accuracy is higher.

Traditional water quality detection schemes are inefficient, which will greatly delay the timely management of fishery water. UAV has expanded our thinking of using it in the fishery field. It focus on the principle of pH, dissolved oxygen and ammonia nitrogen sensors and their application in fishery production practice. Figure 2.4 shows the framework of the fishery water quality assessment and prediction system.

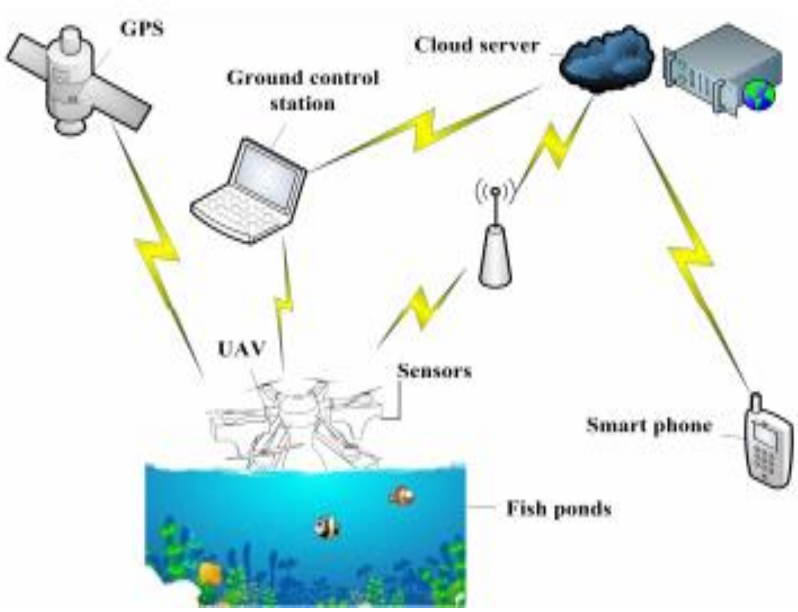

Figure 2.4 Framework diagram of fishery water quality prediction system

Dong Yao et.al.,[17] discussed the Remote control has been used to park the drone on the planned route, use the computer terminal ground station to monitor the UAV's flight status, and the GPS satellite to locate the real-time area. After the drone is parked on the surface, three groups of sensors hidden in the EVA material began to monitor the water quality. During the floating sampling period, the data is transmitted to the database of the cloud server through GPRS module, and the data is processed at the back end. The evaluation and prediction results are finally displayed on the application side and PC side. In this system, each group of sensors uploads the average sampling value to the cloud server, and data are uploaded and updated once every 3 seconds interval. In the 40-minute evaluation operation, 10 minutes are used for sensor pre-processing, the remaining 30 minutes for data acquisition, each group of sensors will update 600 times of experimental data.

All sensors data will be processed by singular data rejection, ADF detection, periodicity test, orthodoxy test, autocorrelation and partial correlation discrimination, ACF and PACF rule test, AIC order determination, white noise test, residual fitting, and finally complete the modeling and prediction. Figure 2.5 and Figure 2.6 shows the predicted data is substantially close to the actual value during the predicted time period of 1500 seconds to 1800 seconds.

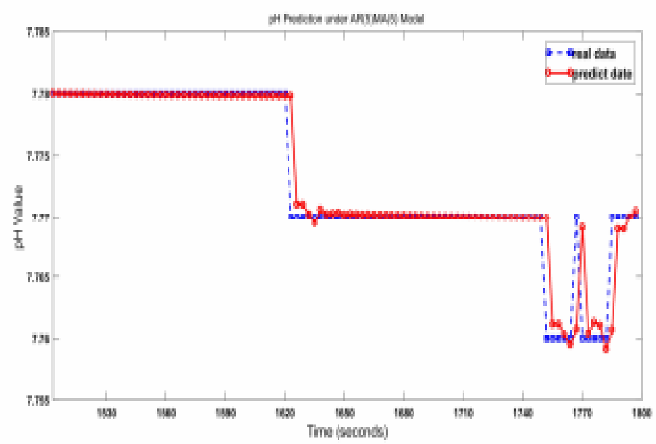

Figure 2.5 The pH Prediction under ARIMA Model

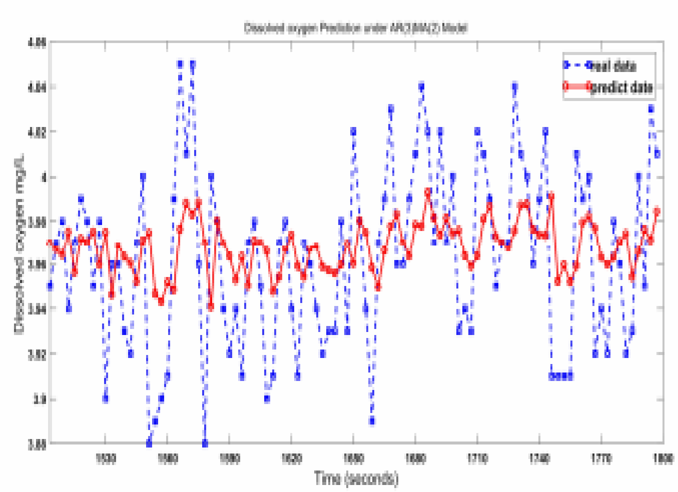

Figure 2.6 Dissolved oxygen (DO) Prediction under ARIMA Model

2.2 Comparison

The following table (Table 2.1) gives advantages and disadvantages noted from the above research papers.

Table 2.1 Comparison

Research Paper	Advantage	Disadvantage
River Water Quality Parameters Prediction Method Based on LSTM-RNN Model	The improved LSTM structure is helpful to avoid the problem of gradient disappearance and explosion in typical RNN. This means that LSTM has advantages in capturing long-term dependence and modelling nonlinear dynamics and can be used to deal with long length sequence data.	LSTMs take longer to train. Extending the number of hidden layers to test the application effect of multi-layer LSTM-RNN network structure or starting with many parameters of LSTM-RNN model, seeking more effective parameter optimization methods and so on can be done.
Automatic and Accurate Prediction of key water quality parameters based on SRU deep learning in mariculture.	SRU performs a higher degree of fitting with the real value, so the SRU water quality parameter prediction model has better prediction result.	The SRU adds a new unit states based on the structure of the RNN to control information in the network, and it slightly redundant in the training time
Aquaculture Water quality prediction using Smooth SVM	SSVM has several advantages over other methods such as neural network. Those advantages are. a. Assuring global optimum b. Parameters that must be estimated relatively few c. The model is stable d. Relatively simple to use e. Successfully applied in most real cases.	The method introduced here employs SSVM approach for predicting water quality, especially dissolved oxygen concentration. It is not predicting for other parameters which is important.
Automated fish farm Aquaculture Monitoring System.	Image processing is used to detect the color of the water.	Better intelligence image processing system has to be used as water color may alter.

2.3 Summary

In this chapter we mainly focused on existing system through literature survey and various IEEE paper analysed and we specified some important points of each paper and related diagrams or graphs are included. In comparison section we have mainly highlighted few important advantages and disadvantages in each paper and comparison between those papers. This chapter also introduces drawbacks of existing system and functionality of proposed system.

CHAPTER 3

REQUIREMENTS SPECIFICATION

3.1 Functional Requirements

The system should be able to provide these functionalities efficiently.

- **Resource Visualization**: The visualizations should be self-explanatory which can be easily understood by the user. There will be line plots and graphs which can be used as an effective measure while devising any new program.
- ML algorithm should be able to predict the output efficiently and accurately.
- On exceeding the critical conditions, alert should be sent to the aqua farmers.
- Predict the water quality parameters in hourly manner. i.e., predict for next hour based on past data.
- Alert should be given through web application.

3.2 Non-Functional Requirements

Non-functional requirements are requirements that specifies criteria that can be used to judge the operation of a system rather than the behavior.

- **Usability:** System has been made user friendly by developing a web application, so it's easy to use.
- **Scalability**: If more parameters required, it can be added easily. Number of visualizations can be increased. Currently the system predicts for hourly manner this interval can be changed accordingly.
- **Reliability:** System should give reliable predicted results.
- **Performance**: Our LSTM model will have improved performance because of the use of datasets with lowest time intervals and has high precession. For checking the accuracy we have shown the performance metrics using RMSE.
- **Documentation**: Coding standards are maintained throughout the project.

- **Maintainability**: This project has easy maintainability of the web application, can be modifiable and integrated with advanced computational and operational technologies.

3.3 Hardware Requirements

- System: Core i5 Processor
- Hard Disk :1 TB.
- Monitor: 15” LED
- RAM: 8GB

3.4 Software Requirements

- Operating system: Windows /UBUNTU.
- Programming Language: Python 3, HTML, CSS, JavaScript, JSON
- Software: Anaconda -Jupyter Notebook
- Web browser -Google chrome, Firefox, IE8+.

3.5 Summary

This chapter gives an insight into the functional an non-functional requirements that the system provides. It also describes the hardware and software requirements that are required for building the system.

CHAPTER 4

SYSTEM ANALYSIS AND DESIGN

4.1 System Architecture

The proposed system architecture (Figure 4.1) shows the complete working of the system starting from training the model using the collected dataset to showing the predicted result and appropriate message on the web application.

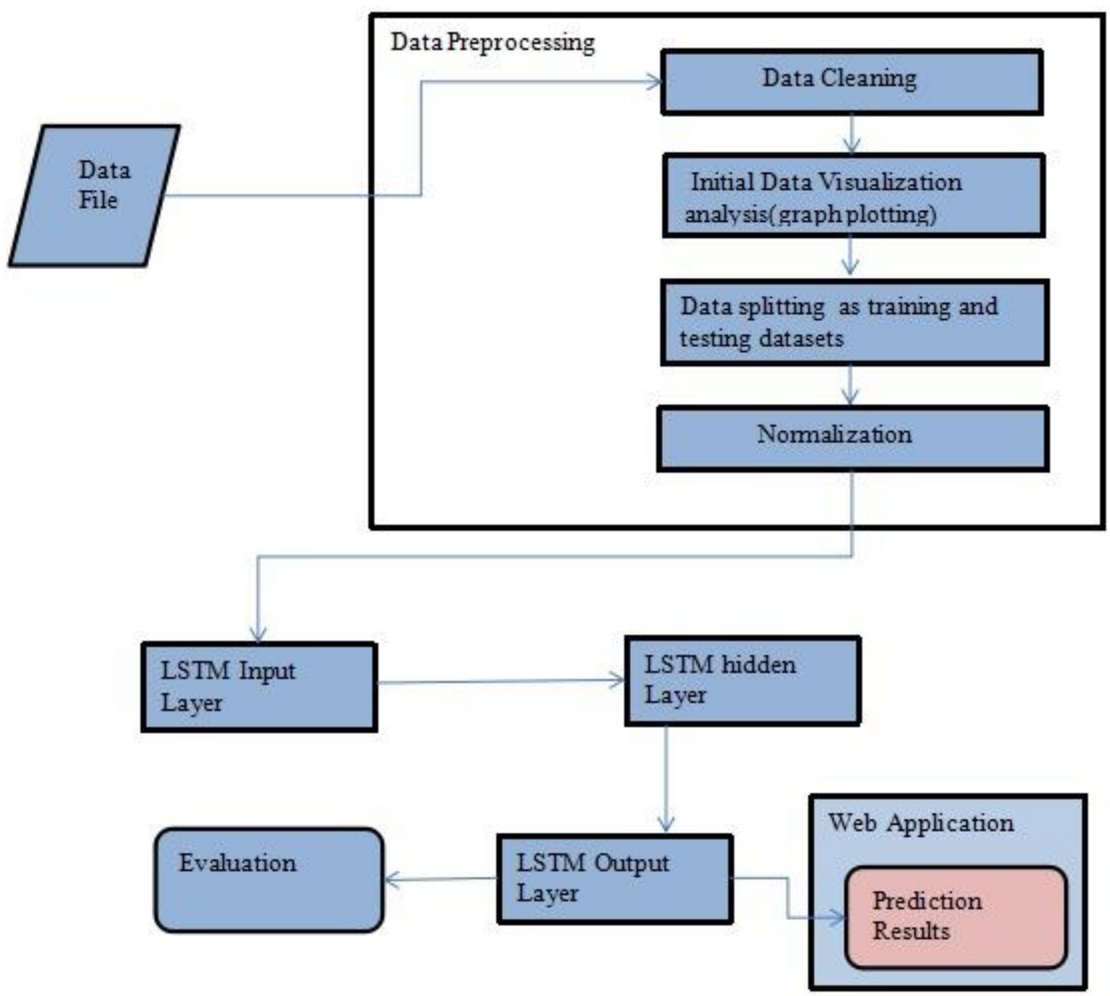

Figure 4.1 Proposed System Architecture

4.2 Flow Chart

The below flow chart (Figure 4.2) shows the step by step execution implemented at the backend and frontend of the system.

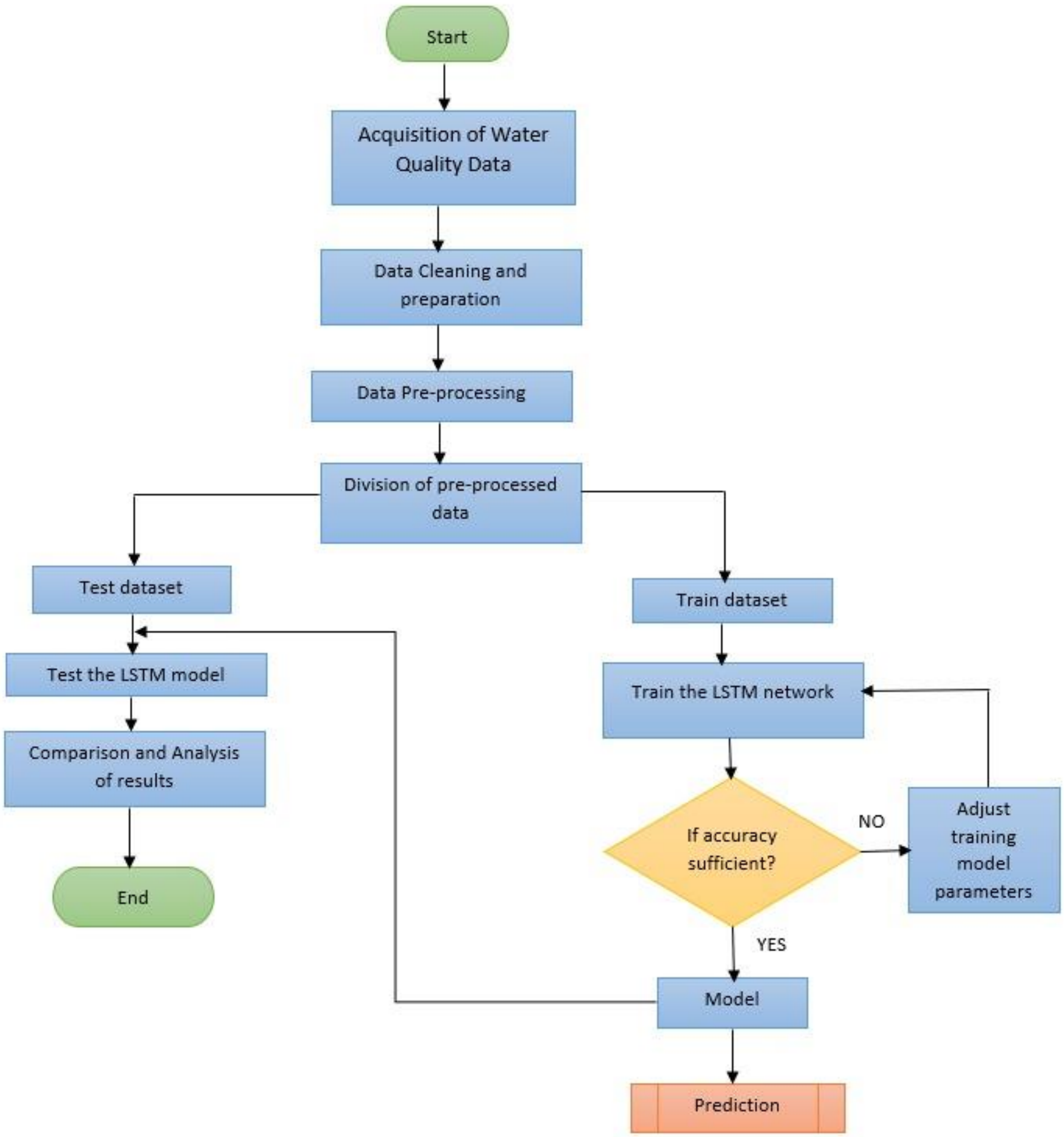

Figure 4.2 Flowchart of the Water Quality Prediction Model

The dataset gathered from USGS's online information archive initially contained 1,50,729 rows. Data cleaning and preparation is performed and then the data preprocessing which involves normalization, scaling is done. The model is created using the preprocessed data. The preprocessed data is divided into train and test data. Fit/ Train the model using training data, matrix of input parameters and array of output, epoch and batch size. Next, the model is evaluated by using the test data and the performance indicators are calculated. The next hour values are predicted, and if they exceed the range of parameters specified, an alert is sent as in Figure 4.3.

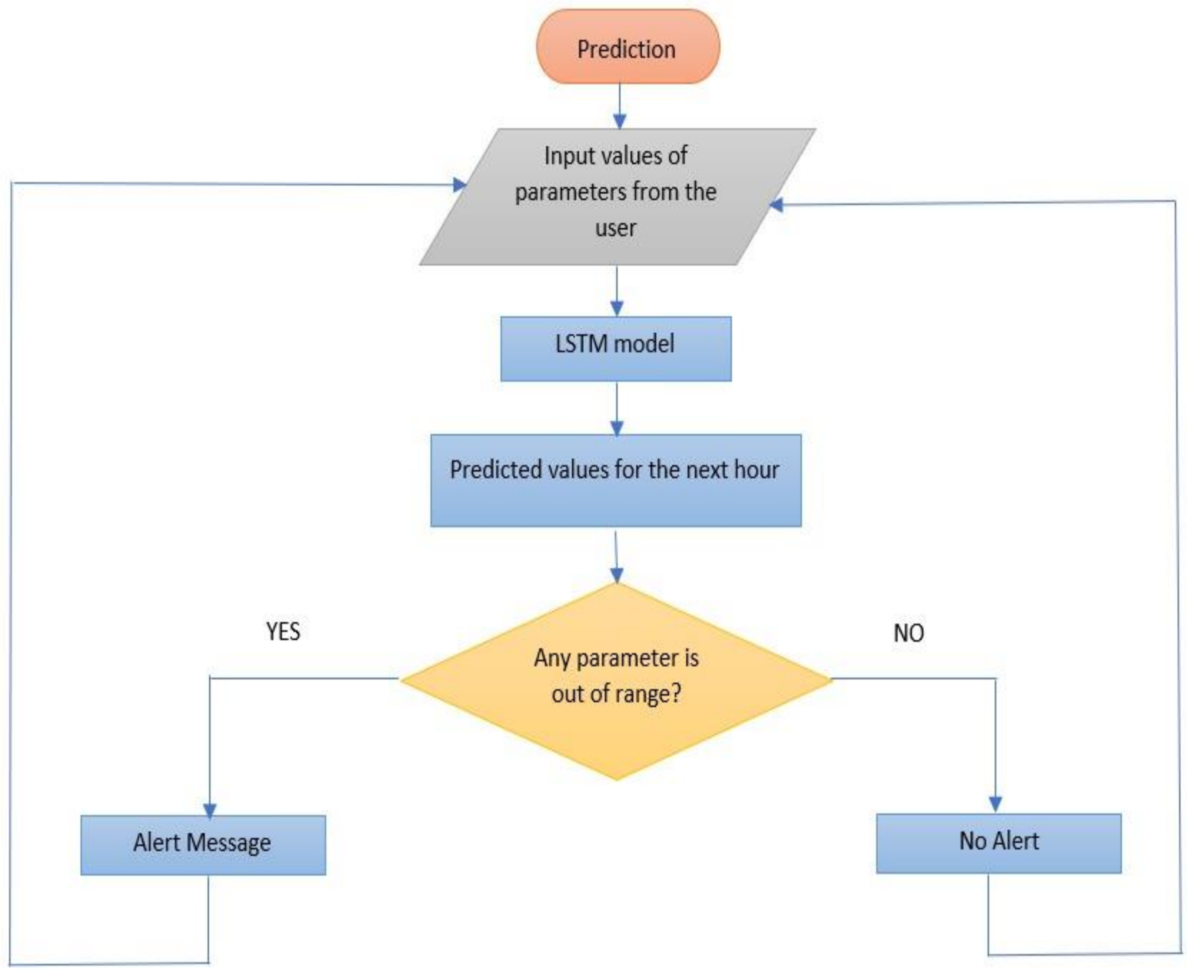

Figure 4.3 Flowchart of Prediction Subroutine

4.3 State Diagram

Figure 4.3 represents the transition between various states of the prediction system. It gives an idea about the various states and the events involved from data collection to generating an alert when any parameter is not in range.

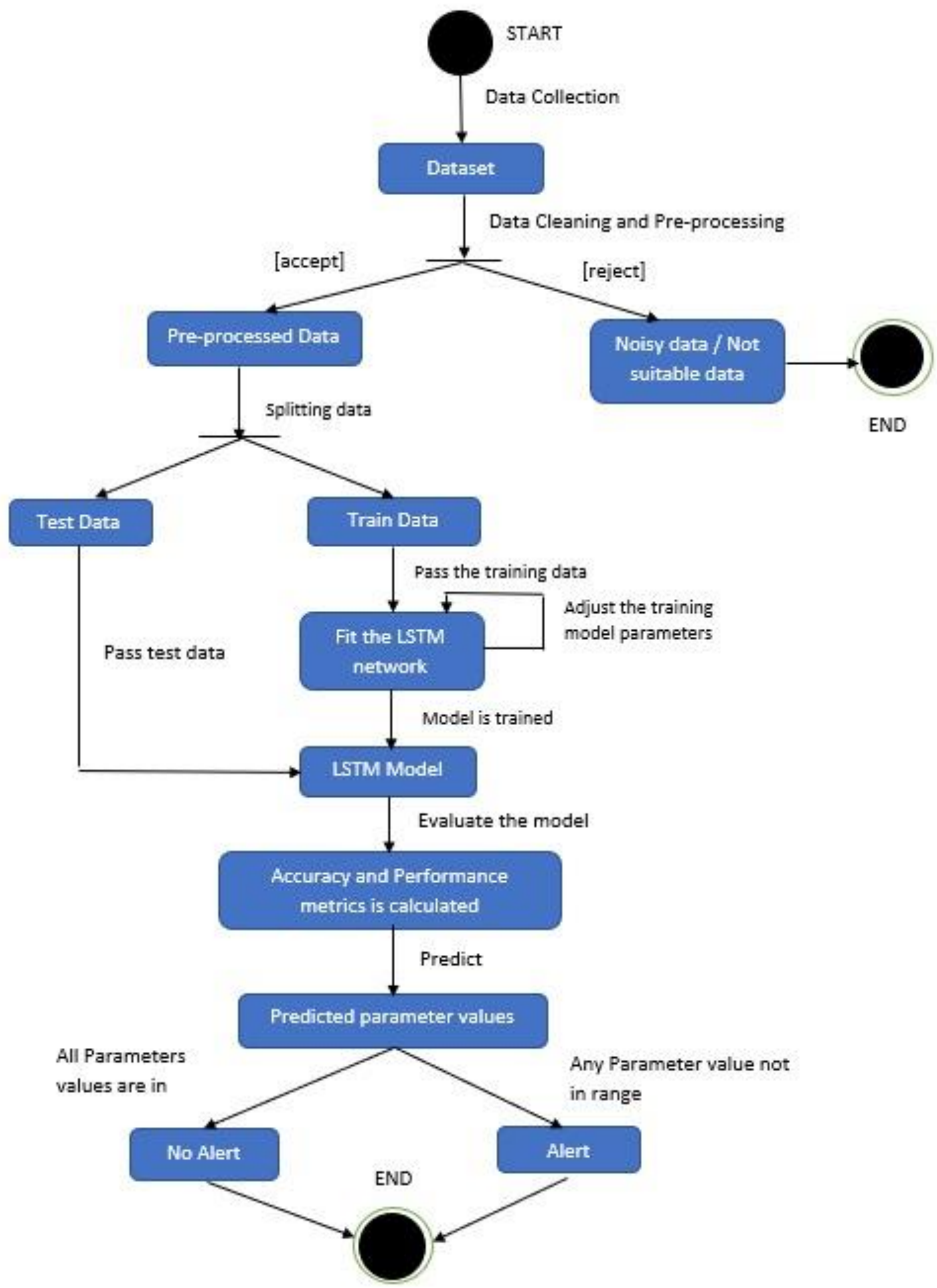

Figure 4.4 State Diagram of Water Quality Prediction System

4.4 Use Case Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved.Here in Figure 4.3 Supervisor is involved in the use cases View prediction result, View Visualization, View Alert and Managing the application. Whereas Aquafarmer is involved only in viewing the prediction result, visualization and alert.

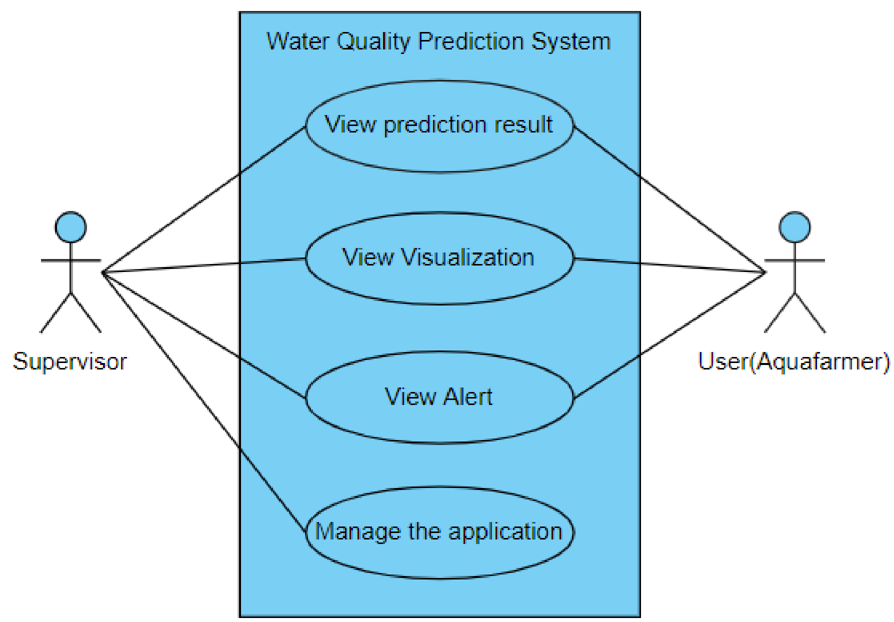

Figure 4.5 Use Case Diagram of Water Quality Prediction System

4.5 Sequence Diagram

Sequence Diagram (Figure 4.4) is interaction diagram that detail how operations are carried out.it shows how they interact over time and they are organized according to object (horizontally) and time(vertically). From the data set data is extracted and processed, respectively. Then ML model is trained, and input data is given by the user and future prediction is made. User can view the prediction results and visualization through the Web application.

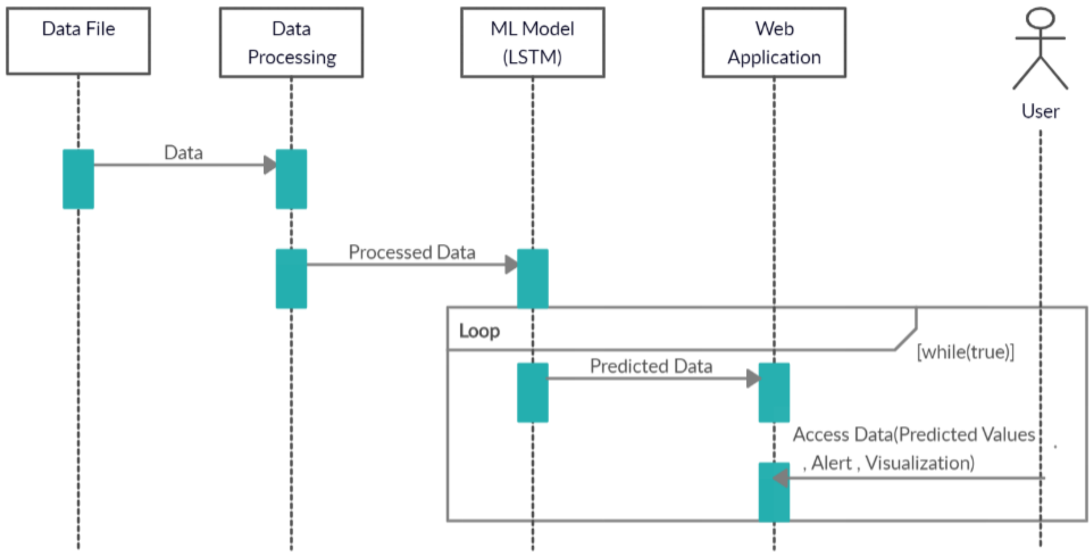

Figure 4.6 Sequence Diagram of Water Quality Prediction System

4.6 Summary

This chapter discusses the working of the system through proposed system architecture. The flow diagram shows the working of ML algorithm. The usecase diagram shows interaction between actors and the system. The sequence diagram is shown.

CHAPTER 5

IMPLEMENTATION

5.1 Long Short Term Memory

Recurrent neural networks (RNN) as in Figure 5.1 are networks with loops in them, enabling the information to persevere. When the gap between the related information and the place it is required is small, RNNs can learn to utilize the past information. Unfortunately, as the gap increases, RNNs become unfit to learn to associate the information.

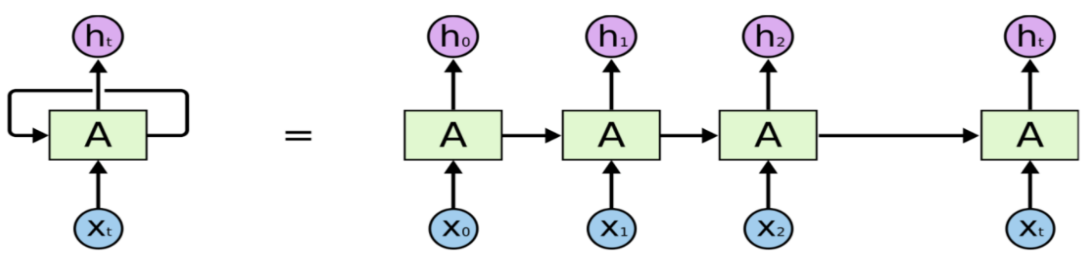

Figure 5.1 Recurrent Neural Network

LSTMs are an extraordinary sort of RNNs, equipped for adapting long term conditions. Recollecting information for long periods purposes their default behaviour. LSTMs also have a chain like structure, yet the repeating module has an alternate structure, not at all like RNNs. Rather than having a single neural network, there are four layers, cooperating in a unique manner. The way to LSTM is the cell state. The cell state is somewhat similar to a conveyor belt. It runs straight down the whole chain, with some minor linear connections. It is extremely simple for information to the cell state, carefully controlled by structures called gates.Figure 5.2 shows the basic LSTM memory block architechure.

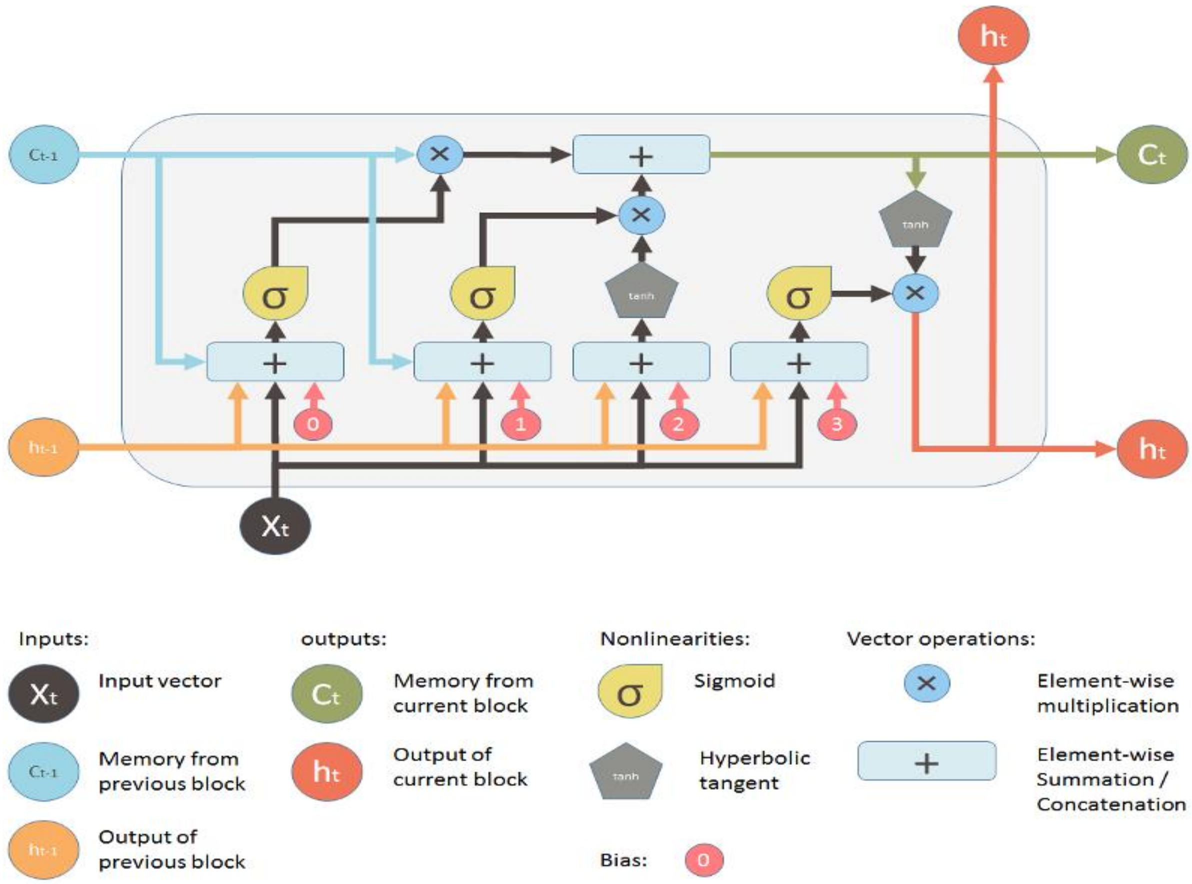

Figure 5.2 Basic LSTM Memory Block

LSTM networks are appropriate for classifying, processing and making predictions based on time series data, since there can be lags of obscure duration between important events in a time series. They were created to manage the exploding gradient and vanishing gradient problems that can be experienced when training traditional RNNs. The activation function of the LSTM gates is frequently the logistic function. The weight of these connections, which need to be learned during training, decide how the gates operate.

A RNN utilizing LSTM can be trained in a supervised fashion, on a set of training sequences, using an optimization algorithm, gradient descent, joined with back propagation through time to calculate the gradients needed during the optimization process, in order to change weights.

Gates as in Figure 5.3 are an approach to alternatively let data through. They are made out of a sigmoid neural net layer and a point wise multiplication operation.

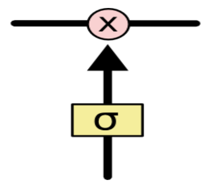

Figure 5.3 Block diagram of a gate

The sigmoid layer yields numbers somewhere in the range of 0 and 1, depicting the amount of every component ought to be let through. A value of 0 signifies "let nothing through", while a value of 1 signifies "let everything through". An LSTM has three of these gates, to secure and control the cell state as shown in Figure 5.4.

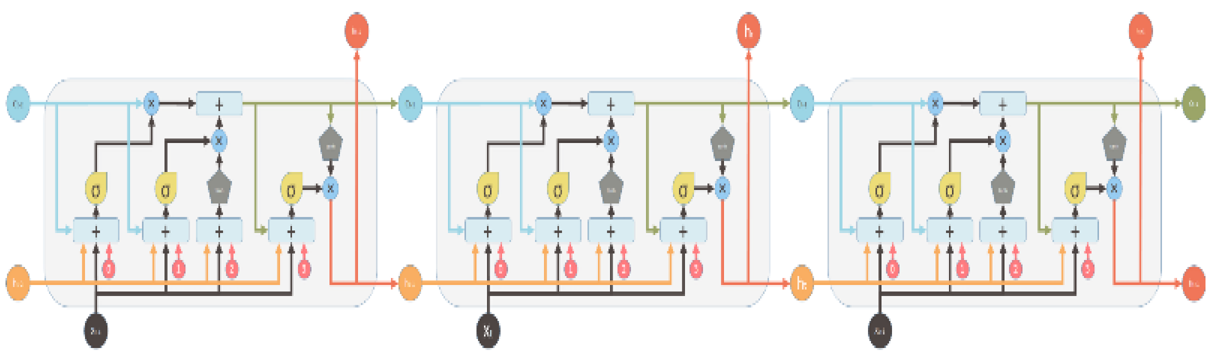

Figure 5.4 LSTM network with memory blocks

The initial phase in our LSTM is to choose what data we are going to discard from the cell state. This choice is made by the sigmoid layer, called the "forget gate" layer. It looks at h_{t-1} and x_t and yields a number somewhere in the range of 0 and 1 for every cell state C_{t-1}. 1 signifies "totally keep this" and 0 implies "totally dispose of this".

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f \qquad (5.1)$$

In eq. (5.1), f_t is the value of the forget gate at t^{th} time. W_f is the weight between the forget gate and the input layer. h_{t-1} is the output of the previous memory block. x_t is the input vector. b_f is the bias vector.

Following stage is to choose what new data we are going to store in the cell state. This has two sections. Initial, a sigmoid layer, called the input gate layer, chooses which values should be updated. Next, a tanh layer makes a vector of new candidate values, C't, which could be added to the state.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{5.2}$$

$$C'_t = \tanh(W_c.[h_{t-1}, x_t], b_c) \tag{5.3}$$

In eq. (5.2), i_t is the value of the input gate. W_i is the weight between the input gate and the input layer. b_i is the bias vector. In eq. (5.3), W_c is the weight between the input gate and the hidden layer.

After that, we update the old cell state, C_{t-1}, into the new cell state, C_t. We multiply the old state by f_t, overlooking the things we chose to overlook before. Then we add $i_t * C'_t$. This is the new candidate values, scaled by the amount we chose to update each state value.

$$C_t = f_t * C_{t-1} + i_t * C'_t \tag{5.4}$$

In eq. (5.4), C_t is the memory from the current block. C_{t-1} is the memory of the previous block.

At last, we have to choose what we are going to yield. This output will be based on our cell state however will be a filtered form. To begin with, we run a sigmoid layer which chooses what parts of the cell state, we are going to yield. At that point, we put the cell state through tanh (to push the values in the range of - 1 and 1) and multiply it by the result of the sigmoid gate, with the goal that we just output the parts we chose to.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o \tag{5.5}$$

$$h_t = o_t * \tanh(C_t) \tag{5.6}$$

In eq. (5.5), o_t is the value of the output gate. W_o is the weight between the hidden layer and the output gate. b_o is the bias vector. In eq. (5.6), h_t is the output of the current block.

Thus, this single unit settles on choice by thinking about the present information, past output and past memory. What's more, it produces new output and adjusts its memory.

5.2 Modules and its Implementation

5.2.1 Data Cleaning and Preparation

The data that we have collected from the USGS NWIS website, has been converted from a tab-separated values (.tsv) file format to a comma-separated values (.csv) file format. The data will be stored in a dataframe and date column has been set as an index. The rows with NaN values will be dropped from the dataframe. The data has been resampled on a daily basis and the daily average has been calculated. Then, the resampled data will be saved in another CSV file.

The above process will be done for all the collected data from 31 water monitoring stations of Georgia. All the CSV files will be grouped together by the date column and the mean is calculated. The output dataframe will be saved in a CSV file.

The resultant dataset will be used for our model training.

5.2.2 Data Pre-processing

The required dataset has been taken and only three columns, i.e. date, dissolved oxygen, and pH, have been used for the computational purposes. Since the values in the date column will be in the string format, we will convert the values in the date-time format and set it as the index. The dataset has been split as training and testing datasets. Training set contains 80%, whereas testing set 20% of the original dataset. MinMax normalization has been used to transform the values of both training and testing datasets in the range of -1 and 1.

5.2.3 Creating the model

To create our LSTM model, first we create a Sequential model, which is a linear stack of layers. We add a visible layer with 6 input, a hidden layer with 16 LSTM neurons, and an output layer that makes a four values prediction. ReLU activation function has been used for the LSTM neurons. ReLu activation function is a piecewise linear function that will yield the input directly, else it will yield 0. This will help in accomplishing better execution, helps in overcoming the vanishing gradient problem, enabling models to adapt quicker and perform better.

5.2.4 Compiling the model

Before training the model, the model has to be compiled using the loss function and the optimizer. Mean Squared Error has been used as the loss function and Adam as the optimizer. Adam is an acronym for “Adaptive Moment Estimation”. It is an optimization technique that can be utilized instead of the classical stochastic gradient process to update network weights in a repetitive manner based on our training data.

5.2.5 Training the model

The model has been trained with 100 epochs. A batch size of 1 has been used while training. One of the callbacks, EarlyStopping has been used, so that the model will terminate itself when the monitored quantity has finished improving.

5.2.6 Model Evaluation

R^2 score, MSE and RMSE between the actual values and the predicted values have been calculated. The values of the metrics has been compared with the traditional ANN model and the ARIMA model. Based on the values, LSTM has shown a better performance than that of others. Hence, LSTM will be chosen for prediction. To achieve more accuracy, the LSTM model can be re-trained.

5.2.7 Prediction

Our LSTM model has been used to predict the values. Testing set will be used for the prediction purpose.

5.3 Tools Used

We have used various tools and modules of python in our model to create our proposed model. We have used tools such as Jupyter Notebook to work with our project.

Jupyter Notebook is an open-source application that enables you to make and share reports that contain live code, equations, visualizations, and narrative text. Its uses are data cleaning and modification, numerical simulation, statistical modelling, data visualizations, machine learning and so on.

We have used various libraries of Python to create our water quality prediction model and produce some visualizations.

- **Pandas**:- Pandas is an open-source, BSD-authorized library giving high-performance,easy-to-utilize data structures. It is a software library, utilized for data manipulation and analysis. It gives numerous highlights, for example,
  - DataFrame object for data manipulation and coordinated indexing.
  - Tools for reading and writing data between in-memory data structures and diverse document formats.
  - Data arrangement and incorporated treatment of missing data.
  - Label-based slicing, extravagant indexing and subsetting of datasets.
  - Data structure column addition and deletion.
  - Group by engine permitting split-apply-consolidate activities on datasets.
  - Dataset combining and joining
  - Various leveled axis indexing to work with high dimensional information in a lower dimensional information structure.

- Time-series functionality:- Date range generation and frequency conversion, moving window measurements, moving window direct regressions, date moving and lagging.
- Gives data filtration.

- **NumPy:-** NumPy is the library, including support for expansive, multi-dimensional arrays and matrices, alongside an accumulation of high-level numerical functions to work on these arrays. Its highlights are:-
  - Python options to MATLAB.
  - n-dimensional arrays.
  - Fourier transforms and shapes control.
  - Linear variable based math and arbitrary number generation.
- **Scikit-learn**:- Scikit-learn is a free ML library for the Python programming language. It highlights different classification, regressions and clustering algorithms including Support Vector Machines, Random Forests, Gradient Boosting, K-means, and DBSCAN. It is intended to interoperate with the Python numerical and logical libraries, NumPy and SciPy. Its highlights are:-
  - diminishing the quantity of arbitrary factors to consider.
  - utilized for differentiating, validating and picking parameters and models.
  - utilized for feature extraction and standardization.
- **Keras**:- Keras is an open-source neural network library. It is equipped for running on TensorFlow. It is intended to empower quick experimentation with deep neural networks, it centers around being easy to use, particular and extensible. It takes into account simple and quick prototyping. It bolsters both convolutional networks and recurrent networks, just as combinations of the two. It runs consistently on CPU and GPU. It designs the model for training.

Flask is a micro and lightweight Python web_framework built with a small core and easy-to-extend philosophy. It provides the users with libraries, modules and tools to help build web applications.

5.4 Summary

This chapter explains about lstm working and its architecture and its importance. Process of developing machine learning model and briefing out each steps, and tools used to develop the model.

CHAPTER 6

TESTING

6.1 Testing Objectives

The use of testing is to identify unwanted bugs and errors. Testing is a process of trying to figure or discover every conceivable faults, defects or weakness in a working project. It provides different methods to check the functionality of the components, sub-assemblies, assemblies, integration mechanism and security systems etc. of a project. Testing a machine learning model, either LSTM or ANN is done by with the dual coding and another by giving the input values to the and checking the predicted value.

6.2 Dual Code

The idea behind dual coding testing is to test the quality of the program by testing two different implementations of the same program (one being the main program) for a given set of inputs and comparing their outputs for correctness as shown in Table 7.1.

Table 6.1 Performance of the two models

Model	R^2	MSE
LSTM (Main Model)	0.765	0.0402
ANN	0.724	0.3904

6.3 Different Data Values

We are examining the predicted value by giving various input values and checking if the alert is raised when the parameters are crossing the threshold value as shown in Table 7.1 and Table 7.3. This helps us to check the difference between the input value and predicted value.

Table 6.2 Set 1 Input Values

Input values at time 'H' hour				**Predicted values at time 'H' hour for time 'H+1' hour**			
Temperature	DO	pH	Turbidity	Temperature	DO	pH	Turbidity
29.08	3.10	7.3	22.6	29.164	3.47	7.509	22.1298
29.13	3.2	7.2	22.34	29.052	3.24	7.2959	22.5032
29.18	3.4	6.8	23.2	29.054	3.52	6.834	23.04
29.25	3.49	6.67	23.5	29.125	3.54	6.66	23.760
29.26	3.32	6.4	23.30	29.250	3.23	6.38	24.70
29.28	3.28	6.32	24.29	29.419	3.33	6.39	24.61
29.32	3.19	6.35	24.38	29.459(Alert)	3.14	6.41	25.21

Table 6.3 Set 2 Input Values

Input values at time 'H' hour				**Predicted values at time 'H' hour for time 'H+1' hour**			
Temperature	DO	pH	Turbidity	Temperature	DO	pH	Turbidity
28.8	3.01	7.5	11.6	29.15	3.27	7.64	11.60
29.19	3.19	7.72	11.76	29.27	3.48	7.93	11.627
29.29	3.46	7.99	11.7	29.22	3.72	8.12(Alert)	11.662

6.4 Summary

This chapter discuss about the importance of testing and varies methods that are used to test the model built. This helps us to understand the performance of the system and make the necessary changes accordingly.

CHAPTER 7

RESULTS AND DISCUSSION

7.1 Discussion

The home page (Figure 7.1) of the application were the parameter values to be given to the system can be entered.

Figure 7.1 Home Page of the Application

Once the values of parameters are entered, the predicted values for the next hour are shown to the user using the backend algorithm for prediction and appropriate message is displayed. When all the predicted parameters values are in range, the following is displayed as in Figure 7.2.

Figure 7.2 All parameters in range

When the visualization button is clicked, the graphs for the various parameters are shown as in Figure 7.3, which can help the user understand the changes in parameters over time.

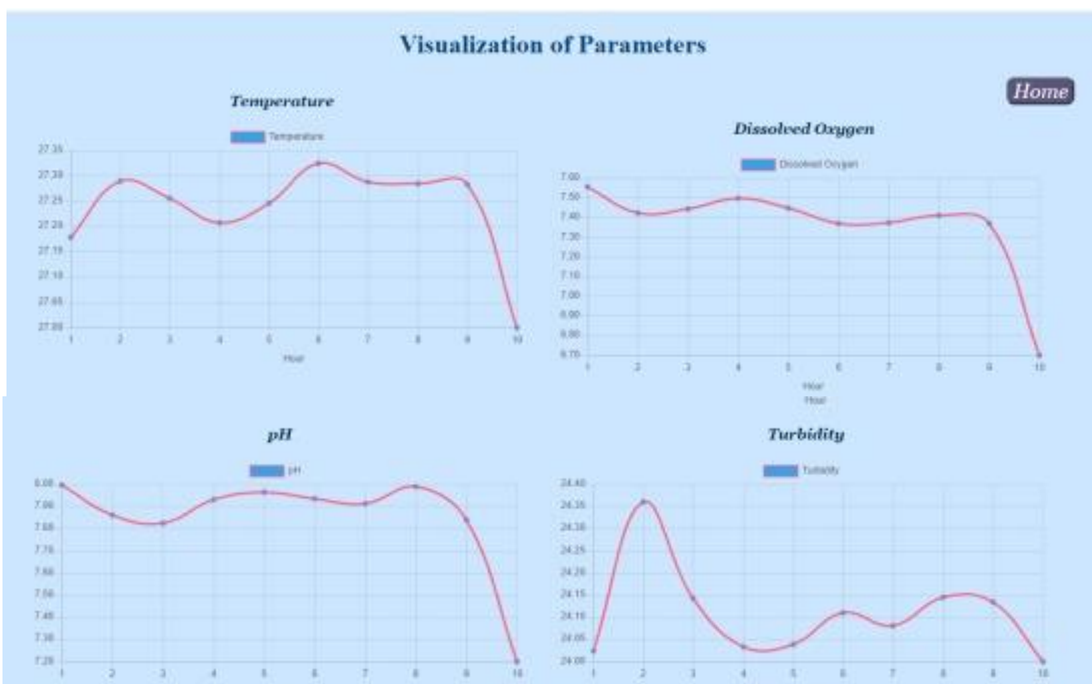

Figure 7.3 Visualization of parameters

Figure 7.4 shows alert message when the predicted value for next hour of dissolved oxygen is out of range.

Figure 7.4 Alert Message when Dissolved Oxygen is Out of Range

Figure 7.5 shows alert message when the predicted value for next hour of Temperature is out of range.

Figure 7.5 Alert Message when Temperature is Out of Range

Two or more parameters can be out of range at the same time, in that case alert will be raised indicating both the parameters. Figure 7.6 shows an alert when temperature and pH both the parameters are out of range.

Figure 7.6 Alert Message when Temperature and pH both are Out of Range

Figure 7.7 shows an alert when Dissolved Oxygen and Turbidity both the parameters are out of range.

Figure 7.7 Alert Message when DO and Turbidity both are Out of Range

LSTM model performs better than other models and the accuracy is good for prediction of water parameters.

7.2 Summary

This chapter clearly shows the working of our web application through the screen shots and by including some brief discussion to that. For better understanding of variations of water parameter values screen shot of graphs is also included.

CHAPTER 8

CONCLUSION AND FUTURE SCOPE

8.1 Conclusion

Water quality is an important factor for aquaculture. This project proposes LSTM model to predict the value of pH, Temperature, Turbidity, Dissolved Oxygen which are the main indicators of water quality in determining the aquatic species habitations.

The model is prepared by the recorded information of water quality parameters which are provided by the USGS website which monitors the stations and maintains the database and instantaneous values can be generated by USGS REST service URL tool. The datasets contain the water quality parameters information like temperature, DO, pH and turbidity from the year 2014 to 2019(till 18th Feb) for Georgia State with 60 minutes of intervals. To remove the noisiness and inconsistency, the data cleaning, and data scaling (normalisation) is carried out along with the graphical presentations and exploratory analysis. To improve the prescient precision of the model, a few re-enactments and parameter determination are completed.

The model shows that the RMSE for the LSTM is 0.03 and the prediction results are stable.Our model is fit for multivariate parameters and the predictions are done in a hourly manner. Web application is built to make to it user friendly. Visualization of water parameters gives good insight for understanding the parameter changes for past hours.

8.2 Future Scope

Developing mobile app is more convenient for alerting the aquafarmers when critical conditions exceeds for water parameter values and app is accesible through phones.

Giving suggestion for aquafarmers based on seasons how to maintain water quality for the surveillance of fishes.

Water quality monitoring for aquaculture using IoT and Predicting the water quality using ML as shown in Figure 8.1.

1. Water parameter values will be monitored using sensors.
2. Solar panel can be used as an alternative power source during power cut reduces the system failure and makes it more reliable.
3. Since lot of data is sensed it has to be stored and then use it for ML prediction,so enoromous amount of data can be stored in cloud(thingspeak) and then it can be retrieved for ML prediction.

Altering the water quality automatically using IoT when predictions raises the alert.

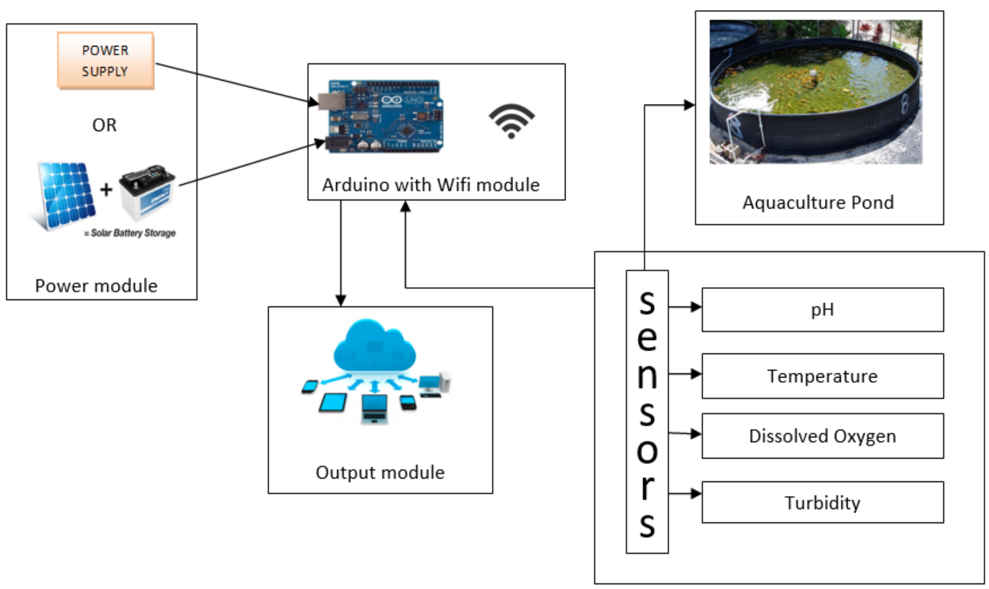

Figure 8.1 Architecture of the future system

REFERENCES

[1] Juntao Liu, Chuang Yu, Zhuhua Hu,” Automatic and Accurate Prediction of Key Water Quality Parameters Based on SRU Deep Learning in Mariculture” 2018 IEEE International Conference on Advanced Manufacturing (IEEE ICAM 2018).

[2] Wijayanti Nurul Khotimah,” Aquaculture Water quality prediction using Smooth SVM”, IPTEK, Journal of Proceeding Series, Vol. 1, 2014 (eISSN: 2354-6026).

[3] J. Wang, F. Xiang, F. Qiu, H. Wang, H. Liu. Research progress of “River Water Quality Parameters Prediction Method Based on LSTM-RNN Model” Guild of Environmental Sciences, Vol.37, No.4, 63-67, 2018.

[4] Tom M. Mitchell, Machine Learning, India Edition 2013, McGraw Hill Education.

[5] Feifei, L., Daoliang, L., & Daokun, M. (2010)., Dissolved oxygenprediction in apostichopus japonicas aquaculture ponds by BP neural network andAR model. Sensor Letters 8, 95.

[6] Sundarambal, P., Shie-Yui, L., & Pavel, T. (2009). Development of a neural network for dissolved oxygen in seawater. Indian Journal of Marine Science, 38 (2), 151-159.

[7] FAO. 2018. The State of World Fisheries and Aquaculture 2018 - Meeting the sustainable development goals. Rome. Licence: CC BY-NC-SA 3.0 IGO

[8] Svobodová Z., Lloyd R., Máchová J., Vykusová, B. Water quality and fish health. EIFAC Technical Paper. No. 54. Rome, FAO. 1993

[9] Chapter 6 - General Relationship Between Water Quality and Aquaculture Performance in Ponds. Claude E.Boyd. Fish Diseases. 2017

[10] Abe, S. (2010). Support Vector Machines for Pattern Classification 2nd edition. New York: Springer- Verlag.

[11] Lee, Y. J., & Mangasarian. (2001). A smooth support vector machine. Journal of Computational Optimization and applications, 20: 5 – 22.

[12] H. B. Wang, Z. P. Yang, Q. Yu, T. J. Hong, X. Lin. Online reliability time series prediction via convolutional neural network and long short-term memory for service-oriented systems. Knowledge-Based Systems, Vol.159, 132-147, 2018.

[13] Sajal Saha, Rakibul Hasan Rajib and Sumaiya Kabir, “Automated Fish Farm Aquaculture Monitoring System”, 2018 2nd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET)27-28 October 2018, Chittagong, Bangladesh.

[14] Po-Yuan Yang and Jyh-Horng Chou, “Prediction for dissolved oxygen in water of fish farm by using General Regression Neural Network”, Proceeding of the IEEE International Conference on Information and Automation Wuyi Mountain, China, August 2018.

[15] Simeonov, Einax, Stanimirova, & Kraft. (2002). Environmetric modeling and interpretation of river water monitoring data. Anal. Bional. Chem, 374 898-905.

[16] Abe, S. (2010). Support Vector Machines for Pattern Classification 2nd edition. New York: Springer- Verlag.

[17] Dong Yao, Lei Cheng, QiuXuan Wu, Gong Zhang, Bei Wu, YuQing He, “Assessment and Prediction of Fishery water quality using electrochemical sensor array carried by UAV**”,** Engineering Research Center of Metallurgical Automation and Measurement Technology of Ministry of Education, Wuhan University of Science and Technology, Wuhan, China