A Classification and Prediction Based Approach for Real-Time ETP Outlet Monitoring through E-IoT and Remote Sensing using Machine Learning and Deep Learning

by

Md. Mehedi Hossain
15201033
Md. Jahid Hasan Mridha
16301052
Sazid Md. Imran
18201193
SK Ayub Al Wahid
19241023

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Mehedi Hossain

Md. Mehedi Hossain
15201033

Md. Jahid Hasan Mridha

Md. Jahid Hasan Mridha
16301052

Sazid Md. Imran

Sazid Md. Imran
18201193

SK Ayub Al Wahid

SK Ayub AL Wahid
19241023

Approval

The thesis/project titled "A classification and Prediction based Approach for Real-Time ETP Outlet Monitoring through E-IoT and Remote Sensing using Machine Leaning and Deep Learning" submitted by

1. Md. Mehedi Hossain (15201033)
2. Md. Jahid Hasan Mridha (16301052)
3. Sazid Md. Imran (18201193)
4. SK Ayub Al Wahid (19241023)

Of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January 11, 2021.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, Phd
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Prof. Mahbub Majumdar
Chairperson
Dept. of Computer Science & Engineering
Brac University

Mahbubul Alam Majumdar, Phd
Professor and Dean
Department of Computer Science and Engineering
Brac University

Abstract

Water is a vital element in our environment but day by day water pollution is increasing in an alarming rate in our country. In Bangladesh's perspective, industries such as textile and ready-made garments (RMG) contribute to a massive amount of waste or effluent. Effluent treatment plant (ETP) are used to remove as much suspended solids from wastewater as possible before it gets back to the environment. However, according to a report published by the Environment and forests ministry, seven state-run factories don't have any effluent treatment plant (ETP) to treat their waste before disposal. And also even the factories which has ETP do not always keep the ETP up and running because it consumes a lot of electricity. The purpose of our research is to establish a setup which will monitor the real-time quality of water outside the industries and inform us whether the ETP is turned on or not with the help of E-IoT and various classification algorithm. It will also predict the seasonal impact where the ETP might be turned off again and what will be the quality of water with the help of various machine learning and deep learning algorithms such as CNN, KNN and LSTM. We have also tracking the sensor value for monitoring and the ETP outlet with RGB color analysis. We have successfully achieved an accuracy of 99% for KNN, 97.5% for CNN and 94.9% forecasting model accuracy for LSTM.

Keywords: : Effluent Treatment Plants (ETP), E-IoT, Water monitoring, RGB color analysis, Video classification, Water Quality Index (WQI).

Dedication

Dedicated to our loved ones for all their support and inspiration.

Acknowledgement

First and foremost, we take this opportunity to express our profound gratitude and deep regards to our supervisor, Dr. Golam Rabiul Alam, Phd whose contribution in stimulating suggestions and encouragement, helped us to coordinate our research especially in coming up with innovative ideas and moving forward to the next level. Moreover, we thank all our friends inside the Department of Computer Science and Engineering at Brac University and specially the academic and admin staffs for their reluctance support throughout our struggle in completing our thesis on time. We also would like to express a great gratitude towards Brac University for providing an excellent environment to conduct our research. We are also thankful to the Department of Computer Science and Engineering for providing us with all the technical support required by us, on time. Furthermore, we would like to dedicate our thesis to our parents and loved ones, for their unconditional love, support, patience and understanding during this work. It would be close to impossible for any of us to have succeeded in playing our roles during this thesis, without their encouragement and most importantly moral support. Lastly we thank our Creator, the Almighty, we all are grateful to for blessing us and making us capable to completing the Thesis immense thanks to all the friends and family for their roles and moral support which enabled us to complete our Thesis on due time.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

BOD	Biological Oxygen Demand
CNN	Convolutional Neural Network
COD	Chemical Oxygen Demand
DL	Deep Learning
DO	Dissolved Oxygen
$E-IoT$	Embedded Internet of Things
EC	Electrical conductivity
ETP	Effluent Treatment Plant
IoT	Internet of Things
KNN	K-nearest Neighbour
$LSTM$	Long Short Term Memory
ML	Machine Learning
RGB	Red Green Blue
TDS	Total Dissolved Solid
$UART$	Universal Asynchronous Receiver-Transmitter
WQC	Water Quality Class
WQI	Water Quality Index

Chapter 1

Introduction

ETP or Effluent Treatment Plant is a setup which is designed for purifying industrial waste water and reusing it. Another purpose of ETP is releasing safe water to the environment so the environment can stay safe from the harmful effects of the effluent. Almost most of the industries in our country has got ETP and they have been deployed to keep the water of our environment pollution free. Almost 71% of the earth's surface is covered with water, of which 2.5% is considered as freshwater[1]. However this limited amount of freshwater resources are under threat of water pollution, mostly generated by human activities. Diseases like cholera, diarrhea, dysentery, hepatitis A etc are directly linked to unhygienic and contaminated polluted water. Every year millions of people die due to waterborne diseases, mostly in third world countries. Bangladesh, being like rest of the third world countries facing severe water pollution for a long period time. And one of the most chief reason of water pollution is industrial production. To reduce the effect industrial water pollution a lot of solutions have been proposed and deployed in the field. ETP is one of them. As per the Bangladesh Environment Conservation Act, 1995 (Amendment 2010) proper management of water waste is mandatory [2]. All the industries in our country are advised to use ETP and keep it running for the whole working period of the industries. But according to the Environment and Forest ministry, seven-state run factories do not have ETP and also the industries which have ETPs, they don't always keep the ETP up and running because it consumes a lot of electricity. The officials visits the industries to make sure that the ETPs are turned on but that is once in a month or a week which is not sufficient to ensure twenty four hours monitoring of the plants. So the water pollution is still severely at risk by industrial waste and effluent. The physicochemical properties of water can effect biological life form of water and also can decide the quality of water. There are physical properties such as temperature, turbidity, electrical conductivity and chemical properties for instance dissolved oxygen (DO), Biological oxygen demand (BOD), Chemical oxygen demand (COD), all come under the group of physiochemical properties[3]. To ensure the quality of water, these parameteDOrs must be in a range and if they go above or below the standard range, the water is not safe for the environment, water life forms and also can be the reason for many waterborne diseases.

We have established a setup with the help of IoT where a module is made with stm32 embedded with raspberry pi 3 and connected with various sensors which will monitor the water coming out of an ETP outlet and measure the quality of water and decide whether the ETP is turned on or off. The data taken by the sensors will be sent on a database and then it will be transferred to a machine learning algorithm which will analyze the data and represent a forecast of the upcoming dates of occurrence. A camera is also be implemented alongside the sensor setup which will capture the video of the water and with the help of RGB color analysis it will monitor the quality of water. The system will be trained with thousands of images taken from the ETP and it will be able to tell the quality of water from just the picture after a period of time. We have used LSTM for time forecasting and KNN for predicting the upcoming occurrences.

Chapter 2

Literature Review

Funabiki et al. [4] applies an analytic system on Smart Environment Monitoring system (SEMAR) that is an ongoing project which is a real time system based on IoT and Big data for monitoring water conditions. Previously, the conventional visualization was used by SEMAR and it could not handle the real-time data. In this research, they upgrade the visualization of SEMAR to the real-time visualization. In this paper they propose the big data analytic system as SEMAR's extension in real time. The water quality sensors were attached with Physical Devices and Controllers which acted as nodes and they were scattered at several points along the river. The nodes use Raspberry Pi 3, type B. In this research. The data communication protocol uses MQTT. The application layer of SEMAR has 3 processes. 1) The learning process where the building of the classification model used in the system is described, 2) Real-time classification, which shows the schema of real-time analysis and 3) Real-time visualization, which demonstrates the process of real-time visualization on the front-end of a web interface. This paper uses data which was retrieved from Surabaya River. The data are laboratory test data and live sensor data. The data of laboratory test is the data from the daily laboratory test result with 1,347 samples and 20 attributes. The data retrieved by the live sensor is the data from the sensors placed in several places of Surabaya's river. The determination of the class label uses the Pollution Index method. From the Pollution Index category, there are four possible labels, namely; Fulfil Standard (0), Lightly Polluted (1), Polluted (2), and Heavy Polluted (3). Classification algorithms which would be used were Support Vector Machine and Decision Tree. SVM with the linear kernel is implemented in this research. One method contradicting the other is chosen to be used in this study. The decision tree or classification tree is used in learning the classification function which infers the dependent attribute (variable) value given by the independent attribute (input) value. This research uses CART. The real-time classification uses big data analytic technology. The generated classification model is loaded before the classification of new data. The water data is read by Spark from Kafka Broker. The classification result is loaded into Kafka Broker using Kafka Producer. Further, Node JS takes data from the real-time classification result of Kafka Broker. By using Node JS, the flow of real-time data can be managed to the front end. In the two datasets, the labelling results only produced two water classes, lightly polluted (class 1) and polluted (class 2). This shows that the water of river Surabaya is differentiated in class of either 1 or 2. The number of mislabeled data, the accuracy rate, and the MSE (Mean Squared Error) by each algorithm on each dataset that were calculated from the confusion matrix. Both algorithms show good performance with the accuracy rate of more than 90% and the MSE around 0.019075. In future, SEMAR is expected to be used in the air environment, and for real-time clustering in mapping the water conditions in the river.

Anjana et al. [5] discusses the main goal is to save water from unnecessary wastage and also monitor the quality of the water that we get daily from our water supplier. Addition to these measures the system also generates an approximate prediction of the water bill. The data is sent to the web by the help of BeagleBK and ipV6 is assigned to all the necessary sensors and motes. Mainly flow sensor, heat sensor along with pH sensor and ORP sensor is used in the underground tank for sensing the water flow, heat and the quality of the water. Flow sensors are also added to each water supplying outlet like toilet, kitchen etc. The flow sensors spin when the water flows and it measures the volume of the water which is used and thus with the help of this data, the bill is generated. The information is sent to the web by the help of BeagleBK and the information is saved in a server, displayed in the webpage and also in the mobile app. In this way the process of this paper is done and clients can see the quality of the water and also predict the water bill.

Liu, P., et al. [6] discussed rapid economic growth and increasing urbanization, Water pollution is becoming more and more alarming. In recent years, establishing a reliable water quality prediction model is becoming more popular day by day. However, traditional water quality models cannot be comprehended as there are a lot of factors which are physics, chemistry, biology, meteorology and hydraulics. And so, to predict water quality and promote scope of application variety new technology has drawn attention such as fuzzy mathematics, stochastic mathematics, 3S technology and artificial neural networks (ANN) are emerging above them, ANN is becoming a popular predictive method because of the excellent applicability to uncertain and non-linear situations. Traditional data collection is solved using big data collections using Iot. Here, based on remote sensing, Iot, cloud computing, big data and artificial intelligence (AI) RS based wireless sensor networks built to smartly monitor water. Where on the basis of the historical data collected by smart water monitoring a predictive model can be done for water quality by observing change in quality parameters with the corresponded to multi-monitoring data to predict successful water quality . Thus, to forecast this LSTMs deep neural networks has been used. Establishing the sample, the data preprocessing parameter settings and learning automatically by LSTMs has given a feasible approach to monitor. Because of the excellent performance of long short-term memory (LSTM) models in the field of time-series prediction, the use of LSTM in environmental research has increased. LSTM is a special kind that have four interactive layers. As a model which is extended with multiple layers of hidden LSTM and each layer containing numerous memory cells, it learns the deep learning technique faster. Lin algorithm has been chosen to obtain a good estimation on missing values but in terms of non-stationary data time series is very poor. Although this paper only considered the single dimensional data with lack of optimization on non-linear data but in this paper there has comparison LSTMs method with time analysis such as the Arima model and Support Vector Regression (SVR) using MSE value to conduct it a reliable research. Zhang, L., et al [7] studied the real-time identification and display of water quality parameters such as pH, water temperature, conductivity, dissolved oxygen and chlorophyll with the help of STM32 and necessary sensors. The device can collect real-time water sample data and conduct remote transmission in the field under unattended control and eventually show all water quality parameters in the background in real time. Here The machine uses ST's main chip based on ARM Cortex-M3 core as main controls , and the periphery is mainly composed of a relay group that controls the timed water collection, a GPRS module, and a 485-serial port. It also controls the water pump by poor electrical power regulation to perform stratified pumping of water samples. The sensor equipment detected the water quality parameters of different layers and the data collected were transmitted via RS485 to the single-chip microcomputer. The SCM is sent via the serial port to the GPRS module and sent to the backend server by the GPRS module. It is beneficial to promote the realization of automated control and data collection in real time, which prevents the low manual productivity and identification inaccuracy.

Zhang, L., et al [8] shows RS techniques have opened a door to helping people widen their ability to understand the earth. RS image data warehouses expand daily, including images with different spectral and spatial resolutions. In this article, we survey recent developments in DL for the RS field and provide a technical tutorial on the design of optical RS data based DL methods. Although there are also several advanced DL techniques for synthetic aperture radar images and data from point clouds for light detection and range (LiDAR), they share similar basic DL ideas of the model for data analysis. The basic algorithm used here is CNN as it is good for mapping. An Auto encoder is a neural network which symmetric that is applied to learn the features from a data set by an unsupervised way. It minimizes the reconstruction error between the input which is taken by the ending layer and the output which is the reconstruction of the decoding layer. An RBM is generally used as training model which is layer-wise in the construction of a DBN. It is a network which two-layered, presenting a specific type of Markov random field. The DBN was applied to the spatial spectral classification of the RS image and demonstrates superior performance compared to conventional dimensionality-reduction approaches such as main element analysis (PCA) and classifiers such as vector support machines (SVMs).For object recognition and scene classification, it has also been successfully proposed in recent years. Sparse coding DL for RS data from four perspectives: 1) image preprocessing, 2) pixel-based classification, 3) target recognition, and 4) scene understanding is a type of unsupervised method for learning sets of bases which are complete and thus it represents the data efficiently. Inputs of the DL networks can be divided into three categories: the spectral feature, the spatial feature, and the spectral–spatial feature a general framework of target recognition using DL methods. The high-level features which are learned by the deep networks are transferred to the classifiers so that they can be classified. In Scene Understanding objects are assembled into scenes. For analyze, they compared our proposed supervised DL method, i.e., the random convolutional network (RCNet), with the spatial pyramid matching kernel (SPMK) method, the SSC or SIFT+ sparse coding approach described in, and our method of unsupervised learning, i.e., the sparse AR method which is saliency guided which was proposed previously.

M. Mohammadi et al [9] talks about how in present IoT is the next big thing

as it connects all the necessary technological components around us and makes out life efficient and easier. But the implementation of IoT is not always very easy as it involves a big number of real time fast data streams which is not very easy to handle with mediocre technology. In this paper, a class of advanced DL (Deep Learning) approaches have been discussed and their nature and properties are thoroughly shown. DL is a suitable approach for Incorporating IoT and applying various DL architecture with IoT helps us to get the desired details which helps us with analytics, insights and predicting the future. The challenges of using DL are also discussed in this paper. Also a major discussion is held on the reported research in this sector. The smart IoT devices that have been already implemented with this technology is also discussed. How to optimize with real time streaming data and the involvement of Big Data in this sector is shown also. Various frameworks of DL are discussed and other approaches jointly with DL are mentioned. The challenges and future predictions in this field is shown. The lesson that is learned from all the research and experiments is shown in the end.
Until now most of the related researches focused on mainly water quality monitoring with only limited sensor data and some of them applied several machine learning algorithms.In our research we have developed a methodology where we have implemented RGB color analysis for tracking and monitoring either the ETP is turned on or off based on the quality of water and validated the WQI that represents the sensor values. We have also implemented video classification by the help of CNN. These methods have never been used before for monitoring any water refinery outlet.

Chapter 3

Background Analysis

3.1 Feature selection

3.1.1 Temperature

Temperature has a big impact on water bodies. If the water temperature rises, the level of dissolved oxygen becomes lower and if the temperature becomes lower, the opposite happens as there is an inverse relationship between water temperature and dissolved oxygen. Also temperature has its own influence on water chemistry. The rate of chemical reaction increases with the temperature of water. Water with high temperature can also dissolve more minerals from surroundings. Therefore it results in high electrical conductivity. So we can say the temperature is a vital property of water bodies. The temperature of the water coming out of an ETP outlet can be of different temperatures which will have an effect on water DO and EC. That is why it is necessary that we take this parameter in utility.

3.1.2 pH

pH is mainly an indicator of expressing the acidity or alkalinity of a solution. On a logarithmic scale 7 is neutral and the lower is more acidic and higher is more alkaline. Each number portraits a 5 times change in the acidity or basicness of water. Water pH of 4 is ten times more acidic than water pH of 5. pH in water in mainly changed by chemicals so if the pH of a water body is changing we can say that the water is changing chemically. Bleaches, ammonia and these type of chemicals are used in industries which has pH over 10 and also many acidic chemicals are used in productions. So pH is a very important parameter in deciding the quality of ETP outlet water.

3.1.3 Electrical Conductivity

EC or Electrical Conductivity of water is basically the ability of water to pass current or electricity. Dissolved salts and other chemicals which are not organic can conduct electrical current, therefore conductivity increases along with the salinity of water. A higher conductivity indicates that there are more chemicals dissolved in the water. Pure water has no conductivity rather it is a good insulator. So we can decide whether the water is fresh or filled with chemicals by the help of EC. The waste water of industries are full of inorganic elements which conduct electricity so the water coming out of ETP outlet should not have high EC. That is why we have chosen this feature of water.

3.1.4 Turbidity

Turbidity in water is mainly the result of suspended materials like clay, slit, organic and inorganic material which are finely divided, soluble colored compounds, plankton and microscopic organisms. It is mainly an expression which describes the optical properties of water. Light can be absorbed or scattered in water rather than being transmitted in a straight line. If turbidity is high, then the water is filled with a lot of suspended materials which is not good for the quality of the water. So turbidity is an important aspect to look at while deciding the quality of water.

3.1.5 Dissolved Oxygen

Dissolved Oxygen is one of the most important factors of water quality. Dissolve Oxygen is basically the amount of oxygen present in a water body. The whole aquatic life depends on this property of water. If Dissolve Oxygen level of water drop below 5.0 mg/l, aquatic life form cannot survive. Fish, invertebrates, bacteria and plants and other aquatic life form die if the amount of oxygen is low in water. If water oxygen remains 1-2 mg/l for a few hours, it results in large fish kills. The water coming out of and ETP outlet travels into many water resources around like lakes, ponds and rivers. So to ensure the safety of water life forms, the measurement of Dissolved Oxygen is necessary.

3.1.6 Total Dissolved Solids

Total Dissolved Solids or TDS is the minerals, salts metals, anions and cations dissolved in water bodies. Total dissolved solids consists of inorganic salts like magnesium, calcium, sodium, bicarbonates, sulfates, potassium and a little amount of organic matter that are dissolved in water. If the amount of TDS is high, then the water is not drinkable. The water becomes unacceptable if the TDS goes over 1200 mg/liter. That is why we have taken this parameter in measure.

3.2 Water Quality Index(WQI)

[6]Water quality index is a numerical value which can be calculated easily and it is used for describing the overall quality of water bodies. It is a fast and straightforward approach to judge the quality of water by looking at a single value and matching it with the standard scale. Necessary water quality parameters are needed to execute the calculation and selected parameters must have standard limit established by WHO/BIS/ICMR. There are a few ways to measure the WQI. We have used the Weighted Arithmetic Index method[10] . There are mainly 3 steps to perform this method. The steps are described below: Step: 1. Calculate the unit weight (Wn) factors for each parameters by using the formula

$$Wn = \frac{k}{sn} \tag{3.1}$$

Where

$$K = \frac{1}{1/S1 + 1/S2 + 1/S3 + \cdots + 1/Sn} = \frac{1}{\Sigma \frac{1}{Sn}} \tag{3.2}$$

Sn = Standard desirable value of nth parameters

Our summation of all selected parameters unit weight factors,
Wn = 1(unity)

Step: 2. Calculate the Sub-Index (Qn) of the value by using the formula

$$Qn = \frac{[(Vn - V0)]}{[(Sn - V0)]} * 100 \tag{3.3}$$

Where Vn = mean concentration of the nth parameters

Sn = Standard desirable value of nth parameters

Vo= Actual values of parameter in pure water (generally for most of the parameters Vo = 0, except for pH and DO)

$$QpH = \frac{[(VpH - 7)]}{[(8.5 - 7)]} * 100 \tag{3.4}$$

Step: 3. Combining Step 1 and Step 2, WQI is calculated as follows

$$\text{Overall WQI} = \frac{\sum WnQn}{\sum Wn} \tag{3.5}$$

This is how the WQI is calculated and the value is matched with the following scale to identify the overall quality of water.

Water Quality Index Range	Water Quality Class
0-20	Very Good
20-50	Good
50-70	Bad
70-100	Very Bad

3.3 Selection of Sensors And Necessary Software

3.3.1 pH Sensor

pH is a figure which is used to express the acidity or alkalinity of a solution. A logarithmic scale of 7 is neutral, the lower the value the more acidic the solution is and higher values are more alkaline. To measure the pH we have used Gravity Analog pH sensor. This sensor is specifically designed for measure the pH of a solution and reflect the acidity of alkalinity. The installed voltage controller chip underpins the wide voltage supply of 3.3 - 5.5V, which is viable with 5V and 3.3V principle control board. In this sensor, potentiometric method determine pH of a solution by using the pH sensitive electrodes. The hydrogen ISE gives an electrochemical potential that is impacted by the hydrogen particle action of the arrangement.

$$
\begin{aligned}
pH &= -\log[H+] = \log \tfrac{1}{[H+]} \\
pOH &= -\log[OH-] = \log \tfrac{1}{[OH-]}
\end{aligned}
\tag{3.6}
$$

The reference electrode, however, is planned to develop an electrochemical potential that doesn't rely upon the composition of the sample. The distinction between these potentials, the voltage (mV) showed on a pH meter, decides the pH value depending on the Nernst equation.

$$\text{Ecell } = E^{\circ} \text{ cell } + (0.0591/n)\log(Q) \tag{3.7}$$

Turbidity Sensor

Turbidity is basically cloudiness of a fluid made by a large number of suspended particles that are not visible through naked eye. This a key test of water quality. We have used gravity analog turbidity sensor to measure the turbidity of effluent waste water. This sensor supports voltage supply of 5V from control board and operates in 40 mA current max. This sensor, along with microcontroller unit measures the turbidity by using light. There are 4 pins in the sensor which do the main work .Pin 1 interfaces with the thermistor of the sensor and sends the signal produced by it to the microcontroller of the device on which it is set. Pin 2 associates with the photodiode of the sensor. This part goes about as an output device for turbidity estimations. Pin 3 is the expansion of the photo semiconductor, it is the input unit for the sensor. Pin 4 is the normal voltage for the sensor.

$$\text{Turbidity } = (2.3 * A)/L \tag{3.8}$$

Where A is the absorbance and L is the length of the optical path.

3.3.2 TDS Sensor

Total Dissolved Solids (TDS) is the term used for describing the inorganic salts along with some organic matters present in solution of water. The main particles are usually calcium, magnesium, sodium, and potassium cat ions and carbonate, hydrogen carbonate, chloride, sulfate, and nitrate anions. For measuring the TDS, we have used the Analog TDS sensor. This sensor supports 3.3 - 5.5V wide voltage input and analog voltage output of 0 - 2.3V. The sensor has TDS measurement range of 0 - 1000ppm and measurement accuracy of ± 10% F.S. A TDS sensor usually reads the conductivity of a particular solution and gives the measurement in Voltage form. But it can also be converted to NTU (Nephelometric Turbidity Unit).

$$TDS = KE * EC \tag{3.9}$$

Where KE is the correlation factor and EC is the electrical conductivity.

3.3.3 Dissolved Oxygen Sensor

: DO (Dissolved Oxygen) is the amount of oxygen present in water. Just as we need oxygen to live, in the same way dissolved oxygen is needed for the survival of fish, invertebrates, bacteria, and underwater plants. To measure the amount of oxygen present it the factory water, we used Gravity: Analog Dissolved Oxygen sensor. This sensor has a wide range of power supply of 3.3 - 5.5V and output signal of 0 – 3.0V. It is an electro-chemical DO sensor where dissolved oxygen in the sample diffuses through an oxygen permeable membrane into the sensor. After that, the oxygen goes through a chemical reduction reaction, which produces and electric signal that can be read by the dissolve oxygen instrument.

$$Id = \frac{4 * F * Pm(t) * A * po2}{d} \tag{3.10}$$

Where Id is the amount of current produced, Pm (t) is the permeability of membrane as a function of the temperature, A is the surface area of the cathode, po2 is the partial pressure of oxygen and d is the thickness of the membrane.

3.3.4 Temperature Sensor

The temperature of water is a physical property of how hot or cold the water is. Though hot and cold are both arbitrary terms, temperature can be defined as average thermal energy of a substance. The overall quality of water also depends on the temperature of water. To measure the temperature, we have used DS18B20 temperature sensor. This sensor supports wide range of power supply of 3.0 - 5.5V and has a temperature range of -50 - +125℃. This sensor can directly convert temperature signal to the serial digital signals for computer processing.

3.3.5 Camera

For the purpose of live video recording and sending the video to Raspberry pi, we have used C525 HD webcam. The camera has 1 GHz of CPU, 512MB ram and 200MB hard drive space. The webcam can support video recording up to 1280 x 720 pixels. The camera has auto focus option so it adjusts with the changes of lighting. For being Hi-Speed USB 2.0 certified, the camera streams smoothly without any interruption. It is easy to setup with the whole data acquisition module as it is very small in size. Overall this camera is a perfect fit for our setup.

3.3.6 STM32

STM32: We have used STM32duino as our microcontroller. STM32duino which is also commonly referred as STM32 blue pill has all the capabilities of an Arduino module but in cheaper price. Bluepill is basically a 32 bit system with a speed of 72 MHz and bus width of 32 bits. Bluepill is superior to Uno in all aspects except the mechanical capability of Arduino shields. Though the mechanical inconveniences cannot be fixed, it is possible to make it run in the Arduino environment. For this reason we have chosen this microcontroller for our data collection.

3.3.7 Raspberry pi 3B+

As microprocessor we have used Raspberry pi 3B+ as it is the best model of Raspberry pi at the present. It has got processor of Broadcom BCM2837B0, quad-core A53 (ARMv8) 64-bit SoC @1.4GHz. This microcomputer has got 1GB LPDDR2 SDRAM. This device has Wi-Fi connectivity and supports both 2.4GHz and 5GHz IEEE 802.11 b/g/n/ac wireless LAN. For our video input this device was the best choice as it has got 1 x full-sized HDMI port, MIPI DSI display port, MIPI CSI camera port, 4 pole stereo output and composite video port. It has got microSD format for OS and data storage.

3.3.8 Google Colab

We have done most of our coding on Google Colab. Google Colab is a product of Google research which allows anybody to write and run python codes on the browser while being provided with free computer resources including GPUs. It is especially well suited for machine learning and data analysis. Google Colab is associated with IPython3 kernel. It runs 64 bit code on the platform.

3.3.9 Putty

Putty is a serial console, open source terminal emulator and network file transfer application. It likewise can imitate control sequences from xterm, VT220, VT102 or ECMA-48 terminal emulation, and permits all kinds of port forwarding with SSH. The network connection layer of Putty supports IPV6 and SSH protocols. This application can also be used with local serial port connections.

3.3.10 MySQL

MYSQL database is a database service which is fully manageable to deploy cloud native applications. The powerful analytics engine of MYSQL improves the performance by 400 times. It is mainly a relational database system based on Structured Query Language or SQL. We have used MYSQL workbench to store the dataset from our sensors and from the functions of table data export wizard we can convert the data to CSV format with minimal effort.

Chapter 4

Working Model

4.1 Proposed Model

In this paper, an embedded system has been proposed based on real-time IoT for ETP outlet monitoring. Two real-time monitoring systems have been proposed that are shown in fig-4.1 for data acquisition. One system is a sensor-based automation system and another one is remote sensing based on video processing. To automate the sensor data, a microcontroller was selected. It is connected to different kinds of digital and analog sensors. The sensors used here are pH, total dissolved solids, dissolved oxygen, and turbidity sensor which are analog sensors, water temperature sensor is a digital sensor based on one wire system. For this system, we have selected STM32 as the microcontroller. STM32 is based on the ARM Cortex-M 32-bit processor core. It has ADC therefore it can take both digital and analog data at the same time. Moreover, a camera is positioned in the ETP outlet to stream a real-time video. Raspberry pi 3B+ is a 64bit quad-core processor that is a natural choice for fulfilling the work of microprocessors in the system. The main purposes of microprocessors in this system are data processing and establish connections between microcontrollers and databases. SQL based databased has been set up to in-store data and sent it to machine learning and deep learning to analyze, predict, validate, forecast, and visualize to provide some logical information on the monitoring

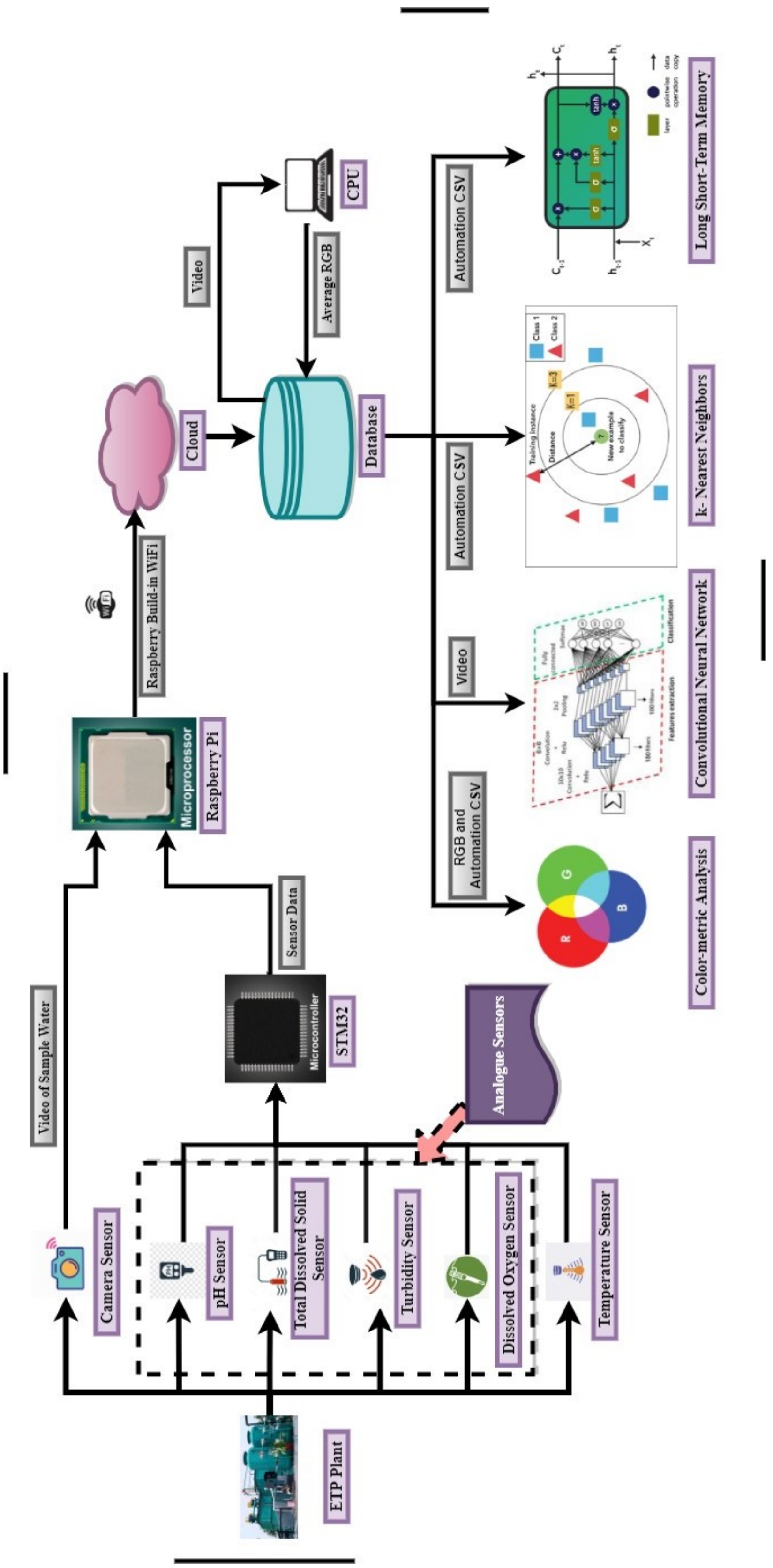


Figure 4.1: Proposed Model

4.2 Workflow

In fig -4.2, the process of the proposed system can be seen. To begin with, we take simultaneously both automation data and Image data on a specific spot of the outlet. So, we will begin the whole embedded system by reading the data of sensors with some delay place in for efficiency. we read the data into STM32 then we convert the data using build-in ADC in STM32. then we send the converted data to Raspberry pi3B+ to further process. However, when this sensor data being read at the same time the camera started taking pictures of the picture. Conversely, it is passive remote sensing, it will be acquiring data when there is light. This video then directly sends to raspberry pi as direct data. The raspberry pi is a microprocessor and thus it will be connected to the STM32 microcontroller and camera with ports. There is S.grab() function to catch the data in raspberry pi. Consequently, the Raspberry Pi sends the data to SQL based cloud with the help of the internet using a built-in Wi-Fi module inside of Raspberry Pi. From the database, it has to be decided by us whether the data is sensor data or not. If the data that is presented to us is video data then using path split the video was sent to two different processes. In one case using RGB analysis, the average RGB value for each frame of the video was found by us. Furthermore, the data of Automation-based sensor data and Remote Sensing image RGB value merger into a CSV. While sending the video directly in another case for classification. This is how the data is acquired. It is then sent to Google colab for preprocessing. In this process, it is analyzed using a k-Near Neighbor along with that it is forecast using LSTM. and CNN to classify the image. Additionally, mapping a color metric analysis. Finally, we validate the model we create for this deep learning. After getting the result the whole procedure can be concluded.

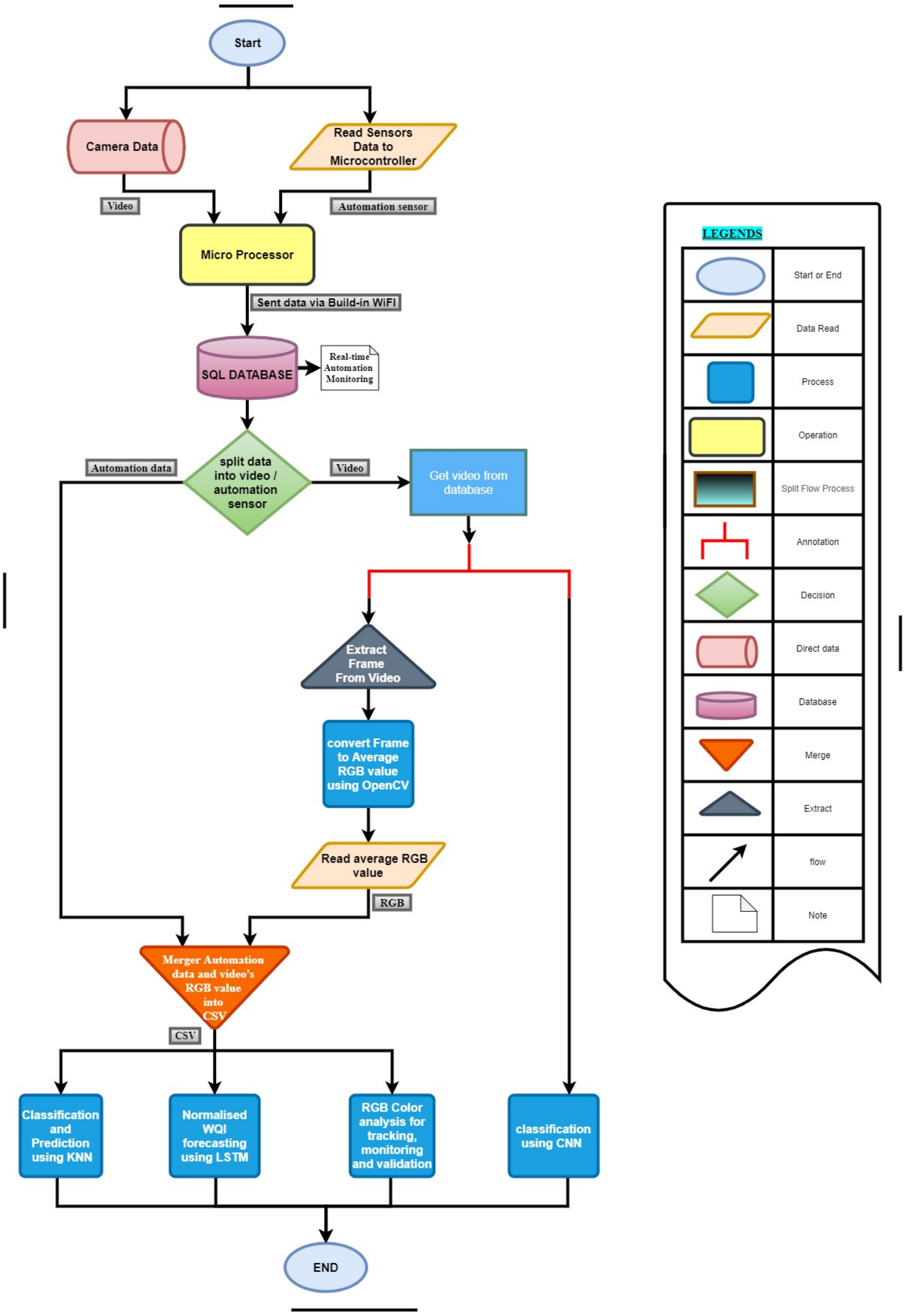

Start
Camera Data
Read Sensors Data to Microcontroller
Video
Automation sensor
Micro Processor
Sent data via Build-in WiFI
SQL DATABASE
Real-time Automation Monitoring
Automation data
split data into video / automation sensor
Video
Get video from database
Extract Frame From Video
convert Frame to Average RGB value using OpenCV
Read average RGB value
RGB
Merger Automation data and video's RGB value into CSV
CSV
Classification and Prediction using KNN
Normalised WQI forecasting using LSTM
RGB Color analysis for tracking, monitoring and validation
classification using CNN
END
LEGENDS
Start or End
Data Read
Process
Operation
Split Flow Process
Annotation
Decision
Direct data
Database
Merge
Extract
flow
Note

Figure 4.2: Workflow

Chapter 5

Methodology

5.1 E-IoT Setup for ETP Outlet Monitoring

5.1.1 Hardware Architecture

We have setup an E-IoT system for real-time ETP outlet monitoring. In this section, we look at the hardware architecture setup in fig-5.1 for monitoring the water quality index (WQI) and predict the water quality class (WQC). For our Microcontroller and Microprocessor, we choose STM32 and Raspberry Pi 3B+ respectively. All of the pin of stm32 is digital. Since there is ADC in stm32, there are 10 analog pins in stm32 from PA0 to PB1 in stm32. For proficiency, the stm32 was placed on the breadboard. A UART named FTDI was used to link our stm32 to Raspberry. At one end it has a pin connection and on the other end it has an USB connector. We connected our STM32's ground and 5V to common GND (-) and common VCC/VDD (+). Thus, Raspberry pi and Stm32 is now serially connected. DS18B20 is used for temperature sensor. It is a digital sensor that has 3 pins where one is VDD which will go to common VCC/VDD.

1) GND which will go to common GND and the digital Data pin will be connected to the c13 pin of stm32. While there is A 4.7k ohm in between the Data pin and VDD.

2) The next Sensor here is a pH sensor which is an analog sensor. It contains T0, D0 ,P0 ,VCC ,2 GND . However, we don't need to T0 and D0. The pin P0 is needed as it is for pH. VCC and 2 GND will go to Common VCC/VDD and common GND respectively. while the P0 pin will be connected to the stm32 A3 pin.

3)For the turbidity sensor, the connector has 1,2 ,3 ,4 , G , A ,D , V pins . The pin 1,2,3, is connected to the sensor's cathode, while we connected the V and G to common VCC/VDD and GND respectively. A and D pins are our output pin which is analog and digital respectively. So, we use A to connect stm32 pin A2.

4) After that, we have connected the DO(Dissolved Oxygen) sensor with an analog sensor to stm32. It has 3 pins. One pin is VCC which we connected to our common VCC/VDD. The second pin is ground which we connect to our common GND. Now the data pin is connected to A1.

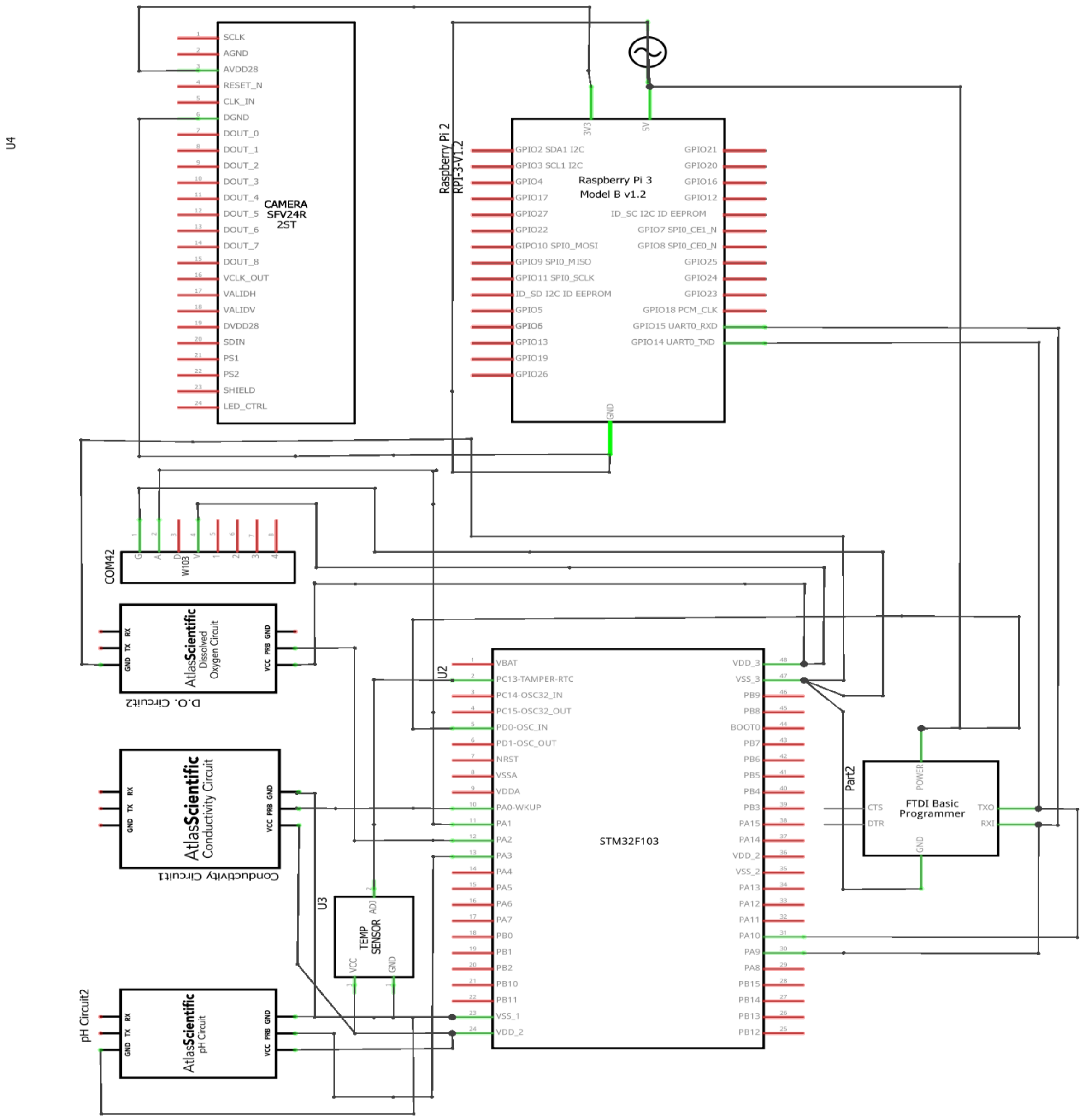


Figure 5.1: Schematic diagram

5) Lastly, we used the TDS sensor which is also an analog sensor . It has VCC, Ground, and Data pin. Ground and VCC connect to common GND and common VCC/VDD. The data pin is connected to pin A0 of stm32

6) We connect stm32 using UART name FTDI which contains 6 pins. We need RX, TX, GND, and VCC pins which will be connected to stm32's A9, A10, common GND, and common VDD/VCC respectively to connect serially. Another end of FTDI has a USB cable import port. So, we use the USB port to connect to raspberry pi's one port to make the stm32 and Raspberry Pi GPIOs connected.Now The Raspberry pi is capable of extracting data from stm32 . As raspberry takes Digital data using stm32's ADC function, the analog data can be converted to digital data for raspberry to process. Finally, the data was sent to the SQL database using a built-in Wi-Fi module.

5.1.2 Data Preprocessing

The sensors mostly take input data in voltage. So, the data needs to be preprocessed before we execute our system. Therefore, to preprocess the data, we called all the necessary libraries.

At first we call the <SoftwareSerial.h> library to send data and <Arduinojson.h> for sending parse data

We take the input data and save it to different variables with the same name as the pins. The data comes as voltage as it needs to be extracted and identified as the key values we need . For that purpose, we define some value and call different methods We define:

VREF 5000, ADC_RES 1024

We used two-point calibration to find the DO data. For that purpose, we also defined.

CAL1_V (1600),CAL1_T (25), CAL2_V (1300),CAL2_T (15)

We also established a matrix for DO to range the DO value.

5.1.3 Data Extract

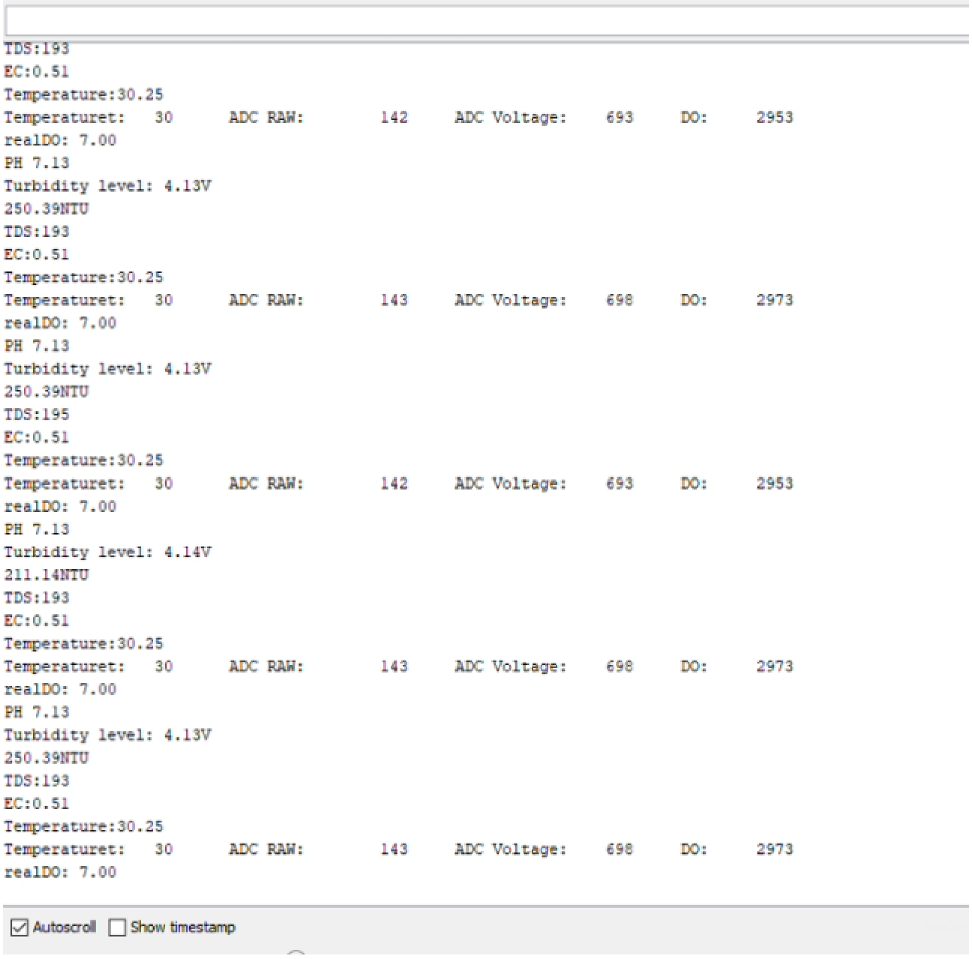
In order to interpret the data, we called some manual methods

realDO(): Here we transform the DO using two calibrate methods and return the value of DO. Here the calibration happens. Calibration is the method of finding a particular value from the perspective of two value

Setup(Void): We begin the system in 115200 frequency

Loop(void): Here we process the voltage of pH and calculate the pH level of the water. As it is where we get the data of turbidity(V) directly, we calculate it as NTU which is the standard for measurement of the transparency of water. Finally, we called readtdsquick() for gathering our other remaining data.

readtdsquick(): Here we got our temperature data by calling the Dallas index. Then we got the electrical conductivity with the help of TDS sensor. Using EC and Temperature we calculate the total dissolved solid using the temperature compensation formula. Lastly, we use the voltage of the DO sensor and temperature to initiate two-point calibration to find the Dissolved Oxygen level. In fig-5.2 we showed the data being collected in real-time.



```
TDS:193
EC:0.51
Temperature:30.25
Temperaturet:   30      ADC RAW:        142     ADC Voltage:    693     DO:     2953
realDO: 7.00
PH 7.13
Turbidity level: 4.13V
250.39NTU
TDS:193
EC:0.51
Temperature:30.25
Temperaturet:   30      ADC RAW:        143     ADC Voltage:    698     DO:     2973
realDO: 7.00
PH 7.13
Turbidity level: 4.13V
250.39NTU
TDS:195
EC:0.51
Temperature:30.25
Temperaturet:   30      ADC RAW:        142     ADC Voltage:    693     DO:     2953
realDO: 7.00
PH 7.13
Turbidity level: 4.14V
211.14NTU
TDS:193
EC:0.51
Temperature:30.25
Temperaturet:   30      ADC RAW:        143     ADC Voltage:    698     DO:     2973
realDO: 7.00
PH 7.13
Turbidity level: 4.13V
250.39NTU
TDS:193
EC:0.51
Temperature:30.25
Temperaturet:   30      ADC RAW:        143     ADC Voltage:    698     DO:     2973
realDO: 7.00
```

Figure 5.2: Automation Data

5.1.4 Algorithm(Automation)

Step 1: Start

Step 2: Read data of Temperature, TDS, DO, Turbidity, pH into one_wire_bus 2, analog(A0), analog(A2), analog(A1), analog(A3) respectively

Step 3: Define DO saturation table

Step 4: begin the system at115200 frequency

Step 5: Calculate pH using following pseudo code :

```
for $i < 10$ do
    buf[i] = Take input for $i < 9$ do
        for $j < 10$ do
            if $buf[i] > buf[j]$ then
                buf[i] = buf[j]
            end
        end
    end
end
for $i < 8$ do
    avgValue = buf[i]
end
float phValue = (float)avgValue * 5.0 * 1.689 / 1024 / 4.75
```

Step 6: Find turbidity value using following formula.

Turbidity(V)= Sensor voltage value* (4.2 / 1024.0)

Step 7: Following the pseudo code find NTU standard of Turbidity.

```
if $Turbidity(V) < 2.5$ then
    ntu = 3000
end
else
    ntu = -1120.4*square(Turbidity(V))+5742.3* Turbidity(V)4353.8
end
```

Step 8: Call the readtdsquick() to find out rest of the sensor data

Step 9: To find water temperature value we directly get it from Dallas temperature by calling the index

Step 10: To find Electrical Conductivity and Total dissolved solid we used these formulas.
rawEc = Sensor voltage analog data ” aref / 1024.0
temperature compensation formula(temperature Coefficient):
finalResult (25^C) = fFinalResult(current) $/(1.0 + 0.02 * (fTP - 25.0))$
Electrical Conductivity(ec) = (2*rawEc / temperatureCoefficient) * ecCalibration[temperature and calibration compensation]

Total Dissolved Solid(tds)= (133.42 *ec3 - 255.86 * ec2 + 857.39 * ec) * 0.5
Step 11: Here, readDO(Sensor data of DO in voltage,Water temperature) to calibrate the Dissolved Oxygen to find v_saturation . In order to get Dissolved Oxygen using following pseudo code:

```
if TWO_POINT_CALIBRATION == 0 then
    uint16_t V_saturation = (uint32_t)CAL1_V + (uint32_t)35 *
      temperature - (uint32_t)CAL1$_T * 35$
    return (uint64_t(VREF) * DO_Table[temperature] * raw) /
      (uint32_t(ADC_RES) * V_saturation)
end
else
    uint16_t V_saturation = (int16_t)((int8_t)temperature - CAL2_T) *
      ((uint16_t)CAL1_V - CAL2_V) / ((uint8_t)CAL1_T -
      CAL2_T)+CAL2_V
    return (uint64_t(VREF) * DO_Table[temperature] * raw) /
      (uint32_t(ADC_RES) * V_saturation)
end
```

5.2 Video Streaming

C525 HD webcam was used for streaming video in real-time. With the help of a USB cable, the camera was connected to Raspberry PI. In raspberry pi the OS calling an SSH command s.grab() fetches the port data into the raspberry pi.

5.3 Sending data to cloud

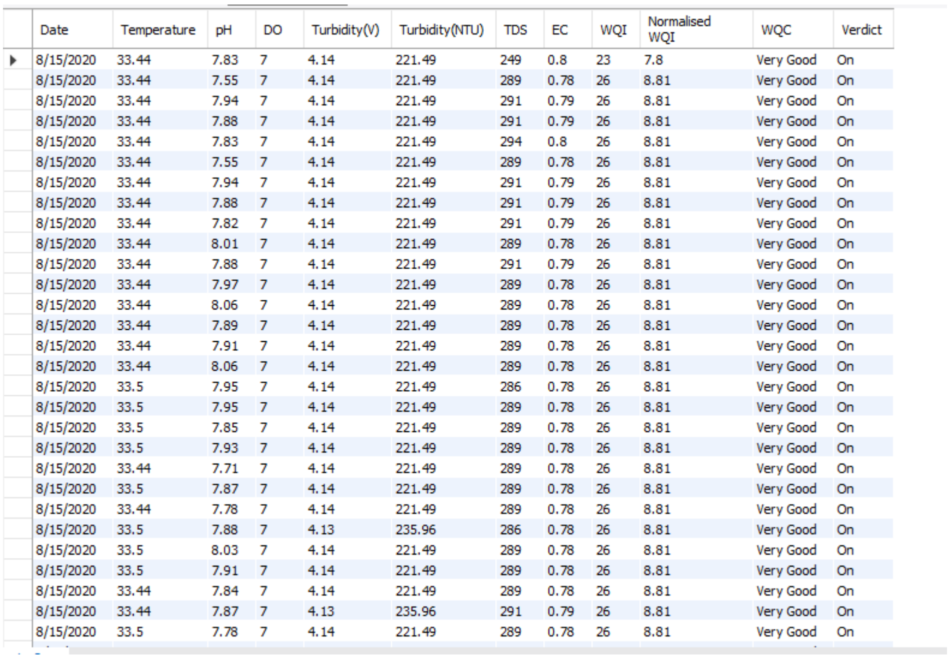
We find the Raspberry pi's IP address using the Advance IP scanner. Now we use PUTTY.exe to connect the host to Raspberry pi. Using SSH we get access to our stm32 data from the port. After the raspberry pi got access to the data it was saved in a CSV and sent to SQL using an API.

5.3.1 Algorithm(Send data to cloud)

```
Algorithm 1: algo of Sending data from raspberrypi to Cloud
Input: Automation sensor data from STM32, Video
Output: Automation sensor data,RGB value , Video
Input Automation sensor data from STM32, Video; into Raspberry PI using
  SSH command named S.Grab(). It allowing to extract all the data from
  any port if it connected .
Define the IP address of the Host Server .
Define Raspberry PI's IP address .
Using API connection Script we Connect the SQL to Raspberry PI .
Check,
  Pseudo code for sending data :
  if S.grab()==1 then
    Send the input data to the cloud;
end
```

5.4 Data storage

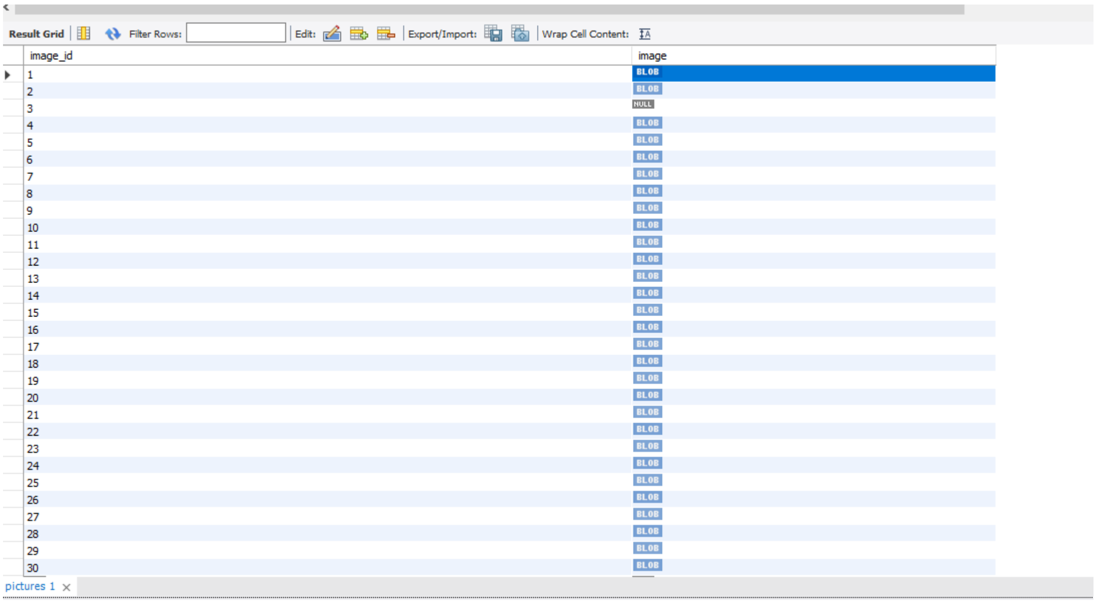
The sensors we have used to measure the quality of water are connected it Raspberry pi 3B+ which has built in Wi-Fi module. The setup sends the data to the cloud and we receive it by the help of MySQL. We have created a database where all the sensor data, images and videos will be stored along with the date and time shown in fig-5.3. A script was made which connects the Raspberry pi with the database and the data stores in real time. The data is at first sent to a web server and then it is transferred to MYSQL. We have connected MySQL with the python codes by the help of MySQL connector. All the sensor data are sent to the KNN code of python. All the images along with the sensor data are sent to the RGB analysis code. And the video is sent to the CNN.



Date	Temperature	pH	DO	Turbidity(V)	Turbidity(NTU)	TDS	EC	WQI	Normalised WQI	WQC	Verdict
8/15/2020	33.44	7.83	7	4.14	221.49	249	0.8	23	7.8	Very Good	On
8/15/2020	33.44	7.55	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.94	7	4.14	221.49	291	0.79	26	8.81	Very Good	On
8/15/2020	33.44	7.88	7	4.14	221.49	291	0.79	26	8.81	Very Good	On
8/15/2020	33.44	7.83	7	4.14	221.49	294	0.8	26	8.81	Very Good	On
8/15/2020	33.44	7.55	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.94	7	4.14	221.49	291	0.79	26	8.81	Very Good	On
8/15/2020	33.44	7.88	7	4.14	221.49	291	0.79	26	8.81	Very Good	On
8/15/2020	33.44	7.82	7	4.14	221.49	291	0.79	26	8.81	Very Good	On
8/15/2020	33.44	8.01	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.88	7	4.14	221.49	291	0.79	26	8.81	Very Good	On
8/15/2020	33.44	7.97	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	8.06	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.89	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.91	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	8.06	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.5	7.95	7	4.14	221.49	286	0.78	26	8.81	Very Good	On
8/15/2020	33.5	7.95	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.5	7.85	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.5	7.93	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.71	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.5	7.87	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.78	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.5	7.88	7	4.13	235.96	286	0.78	26	8.81	Very Good	On
8/15/2020	33.5	8.03	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.5	7.91	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.84	7	4.14	221.49	289	0.78	26	8.81	Very Good	On
8/15/2020	33.44	7.87	7	4.13	235.96	291	0.79	26	8.81	Very Good	On
8/15/2020	33.5	7.78	7	4.14	221.49	289	0.78	26	8.81	Very Good	On

Figure 5.3: Stored sensor data in MySQL database

Images are saved as BLOB file in the database shown in fig-5.4. A binary huge item (BLOB) is concentrated binary information that is compacted into an individual document inside a database. The huge size of the document implies they need unique storage method. BLOBS are binary files, which implies they are typically pictures, sound or other media. Be that as it may, they can likewise be different structures, for example, binary code. A blob is comprised of crude information as a record however somewhat different in makeup. Our images are saved as BLOB format as they are slightly bigger in size. But later on they can be utilized in other programs or codes.

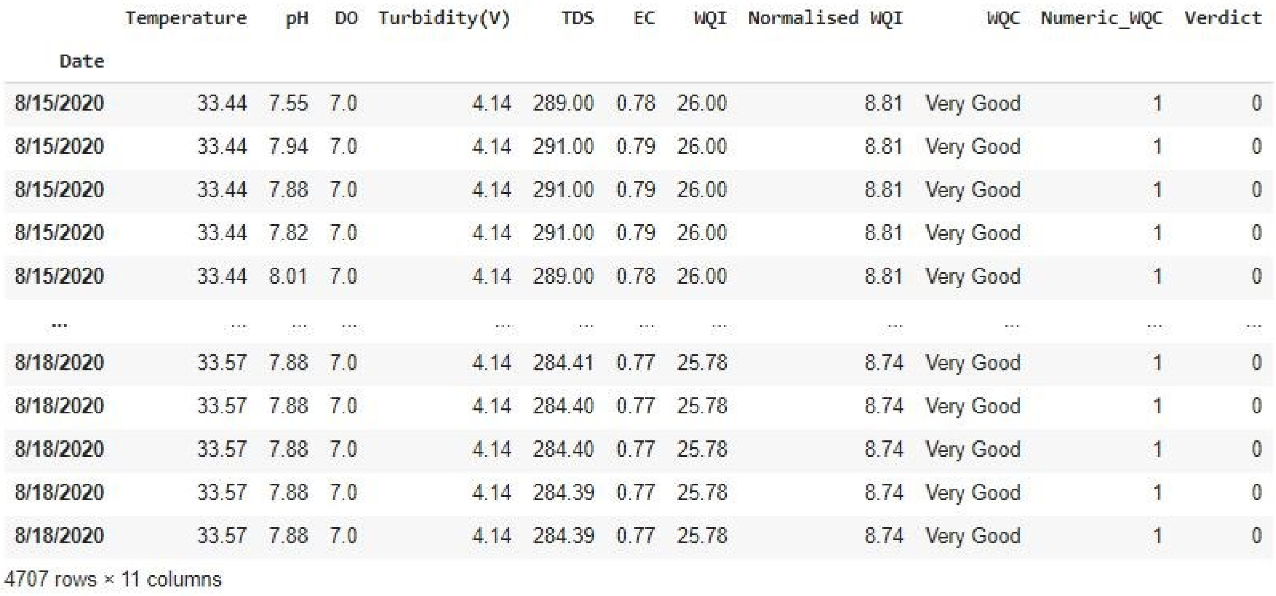


image_id	image
1	BLOB
2	BLOB
3	NULL
4	BLOB
5	BLOB
6	BLOB
7	BLOB
8	BLOB
9	BLOB
10	BLOB
11	BLOB
12	BLOB
13	BLOB
14	BLOB
15	BLOB
16	BLOB
17	BLOB
18	BLOB
19	BLOB
20	BLOB
21	BLOB
22	BLOB
23	BLOB
24	BLOB
25	BLOB
26	BLOB
27	BLOB
28	BLOB
29	BLOB
30	BLOB

Figure 5.4: Stored image data in MySQL database

5.5 K-nearest Neighbors

A tool for classifying objects based on the nearest objects is the KNN algorithm. Machine learning is for obscure or inconspicuous data prediction and data analysis. In Machine learning, a program initially figures out how to play out a part by data set preparing. We transfer both input and output data in Supervised Learning and the result is known as of now. There are two kinds of supervised learning: classification based and regression-based. In this paper, we utilize supervised learning dependent on classification. KNN is a basic calculation that stores all examples accessible and characterizes them as indicated by a size of similarity [11]. Classification issues are directed at defining the features that signify the category to which each entity falls. Both can be used to use this pattern to Understand the current knowledge and foresee how new situations will behave. By analyzing already categorized, data mining builds classification models a pattern will be predicted [12]. Here KNN is implemented for WQC classification and verdict prediction.
Figure-5.5 is showing the dataset that we have used for multiclass classification and prediction. There are 8 columns and 4707 rows whereas the input variables are PH, DO, Turbidity, EC, Temperature, TDS, WQI and the output variables are Verdict, WQC. There are two classes in verdict: ‘On’ and ‘Off’, which are predicted based on



	Temperature	pH	DO	Turbidity(V)	TDS	EC	WQI	Normalised WQI	WQC	Numeric_WQC	Verdict
Date											
8/15/2020	33.44	7.55	7.0	4.14	289.00	0.78	26.00	8.81	Very Good	1	0
8/15/2020	33.44	7.94	7.0	4.14	291.00	0.79	26.00	8.81	Very Good	1	0
8/15/2020	33.44	7.88	7.0	4.14	291.00	0.79	26.00	8.81	Very Good	1	0
8/15/2020	33.44	7.82	7.0	4.14	291.00	0.79	26.00	8.81	Very Good	1	0
8/15/2020	33.44	8.01	7.0	4.14	289.00	0.78	26.00	8.81	Very Good	1	0
...	...	...	...	...	...	...	...	...	...	...	...
8/18/2020	33.57	7.88	7.0	4.14	284.41	0.77	25.78	8.74	Very Good	1	0
8/18/2020	33.57	7.88	7.0	4.14	284.40	0.77	25.78	8.74	Very Good	1	0
8/18/2020	33.57	7.88	7.0	4.14	284.40	0.77	25.78	8.74	Very Good	1	0
8/18/2020	33.57	7.88	7.0	4.14	284.39	0.77	25.78	8.74	Very Good	1	0
8/18/2020	33.57	7.88	7.0	4.14	284.39	0.77	25.78	8.74	Very Good	1	0

4707 rows × 11 columns

Figure 5.5: Dataset for ETP outlet monitoring (used for KNN model)

input attributes and representing the off/on condition of ETP. Further these classes are converted into binary number where 0 represent 'On' and 1 represent 'Off'. On the other hand WQC has four classes: very good, good, bad, very bad. These are representing the water quality of ETP outlet. These four classes are represented as follows: 'Very good' as 1, 'Good' as 2, 'Bad' as 3, 'Very bad' as 4.
Both graphs in Figure-5.6 visualize the count number for classes of WQC and Verdict. For balancing the classes, smote module is used in this model to get an average feature from the nearby neighbors and create a new feature for the particular class.

KNN algorithm has run several times with different K values to select the value of K, and selects the K that decreases the amount of errors we find while retaining the ability of the algorithm to correctly make predictions when data is presented that it has not seen before. Figure-5.7 shows the change of error rate in terms of the value of K.
The efficiency of a KNN classifier is calculated primarily by the choice of K and the distance metric applied. The calculation is influenced by the flexibility of the value of neighborhood size K, since the radius of the local area is determined by the distance to the question of the closest neighbor Kth and various K yields different probabilities in conditional class [12]. The error rate of K=1 for the training sample is always zero. This is because the nearest point to the data point in any training is itself. Therefore, with K=1, the forecast is always correct. Therefore we will raise K and take a wide area around the query into account in order to further smooth the calculation. From the diagram of Error rate vs K-value we can observe that the value of K<8 the error rate just tends to hover around 0.06-0.33. After the value of K>8 the error rate is gradually increasing. From this analysis the value of k is set at 8. The predictions based on the KNN after choosing the value of K. The KNN forecast for classification is the average product of the K nearest neighbors:

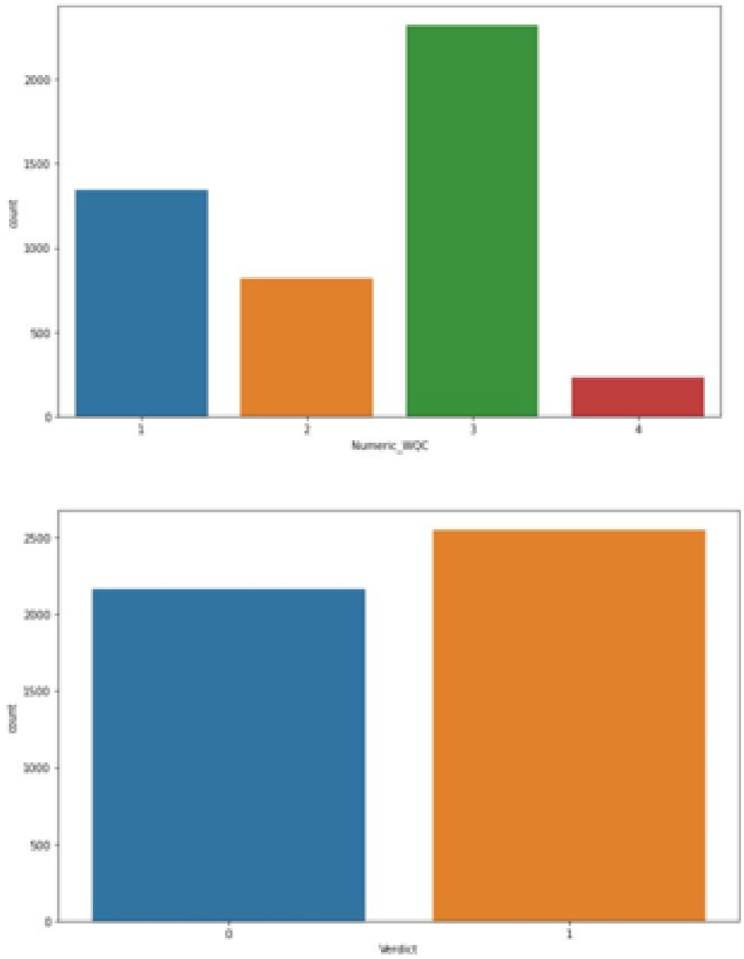


Figure 5.6: Count Number Of Classes of WQC and Verdict

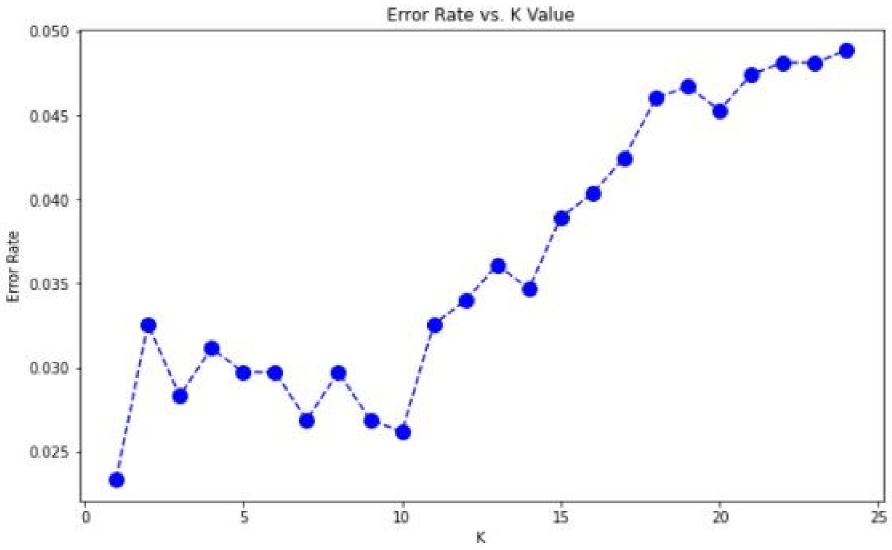


Figure 5.7: Error Rate VS K value

$$\rho = \frac{1}{K}\sum_{i=1}^{k} y_i \tag{5.1}$$

Where y_i is the case of the $\hat{ith}$ sample of the examples and y is the product. KNN's predictions are based on the intuitive premise that near-distance objects are theoretically close, it makes good sense to when making forecasts, differentiate between K's nearest neighbors. A set of W weights, one for each closest neighbor, identified by each neighbor's relative closeness with respect to the query point.

$$W\left(x, p_i\right) = \frac{\exp\left(-D\left(x, p_i\right)\right)}{\sum_{i=1}^{k} \exp\left(-D\left(x, p_i\right)\right)} \tag{5.2}$$

Where $(x_2 p_i)$ is the distance between the query point x and the i^{th} case p_i of the example sample. For classification problems, the limit for each class variable is taken from the above equation. It is obvious that when K > 1, one standard deviation for predictions in regression tasks can be defined normally by using,

$$\text{err bar } = \mp\sqrt{\frac{1}{K-1}\sum_{i=1}^{k}\left(y-y_i\right)^2} \tag{5.3}$$

5.6 RGB COLOR ANALYSIS

RGB color model is using for real time ETP outlet water monitoring and tracking. From industries like food industries, Pharmaceutical industries, tanneries, textile and dye industries, automobile industries are more likely produce a huge amount of waste water that has a major contribution of coloring agent. For instance, the dyeing and washing process in textile industries, which is around 50 percent of the dye must be released into the ETP, is a significant contribution to color on textile waste water. Hence other industries are also producing wastewater containing coloring agent that is removed by decolorizing agent in ETP plant. Therefore Color analysis is essential for monitoring ETP along with WQI that provides the validity of sensor data. In this section of this paper is proposed a model that is monitoring and tracking the real-time average RGB value that is extracted from frame of streaming video. This RGB value is plotted against WQI which we have got from e-IoT sensors value to visualize the presence of coloring agent in terms of WQI value.

Monitoring and tracking color objects in computer vision is an important and fundamental subject. Color is a significant element of knowledge in the classification of image processing. The three primary colors of red, green, and blue represent the RGB space; the other colors consist of the three primary colors. The RGB model, as seen in the figure 5.8, is represented by the Cartesian coordinate system. 1. The three axes represent R, G, B, and each point in the three-dimensional space implies the three components of the value of brightness. The value of luminosity is between 0 and 255[13]

Natural gray shades are generated in the RGB color model by adding equal amounts of all three color values: red, green and blue. If those three components are assigned to 255 (255,255,255), the complete presence of the color illuminates white, while the

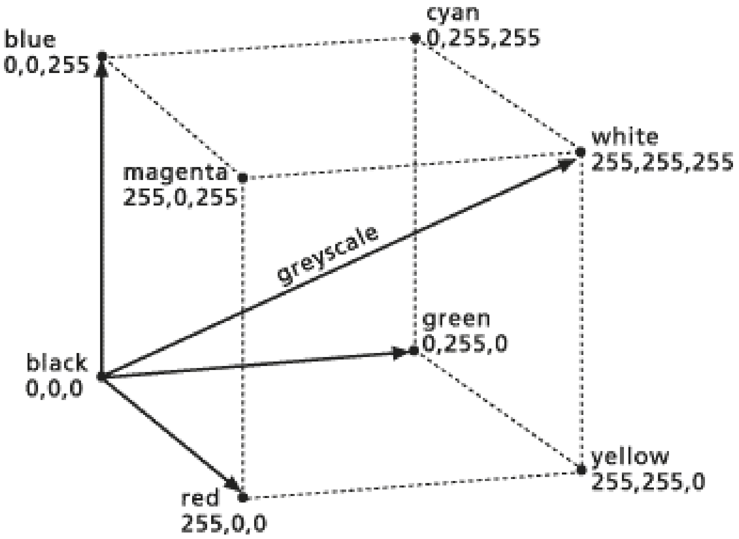


Figure 5.8: RGB color model

absence of the color illuminates black if all three color values are set to zero.

We are using digital camera for streaming video that is connected with raspberry pi. This paper uses the CVCAM technology of OpenCV to simultaneously notice the collection of visual sensor, encoding, and streaming video collection, and realize the processing, storing, streaming file. For all kinds of image and video processing, such as facial recognition and identification, license plate interpretation, picture editing, advanced robotic viewing, optical character recognition, OpenCV is used. We extract one frame per 60 frame from the video because of WQI value that is calculated from e-IoT sensor data is received per 2 minute interval and the video is streaming in 30fps. The algorithm for extracting average RGB value is shown in algorithm 1.

Algorithm 2: RGB color analysis

```
Input: Streaming video
Result: Average RGB value
Capture= cv2.VideoCapture(0);
i = 0;
Frame_skip = 60;
while cap.isOpened() do
    ret, frame = cap.read();
end
else
    BREAK
end
if $i > frame_{s}kip - 1$ then
    Computing the RGB value Computing the average RGB value Store
      average RGB value
end
i += 1
```

Average RGB value is computed from each frame and merged with sensor value.

Matplotlib is used for data visualization. It is a two-dimensional plot library of pythons for visualizing details and generating dynamic graphs. It becomes simple to visualize data from large and complex data.

5.7 Convolutional Neural Network

A subset of deep learning neural networks is the convolutional neural network (CNN). A major advance in image recognition is expressed by CNNs. They are most widely used for digital imagery processing and are also used in image recognition behind the scenes. Video classification is the strategy for taking input and producing a class or the likelihood that the information is a sure class. In this paper video classification is implemented to predict the verdict. There are three types of classification: Binary, multilabel and multiclass classification. We have used binary classification to predict the condition of ETP that is running or not. Figure 5.9 shows a basic flowchart of CNN algorithm that is implemented in this paper.

In order to construct our CNN model, we obtained a sample video captured during the ETP is off and on. The frames extracted from the video during ETP is running, are stored in the 'On' folder and the frames extracted from the video during ETP is off, are stored in the 'On' folder. There are 2289 images in ‘On’ folder and 2820 images in ‘Off’ folder to train, test and validate the model.Figure-5.10 and Figure-5.11 shows some of the sample images from 'On' folder and 'Off' folder.

The size of all images are set to 180×180×3. We are using list format to handle the data. Here all input data is in x-axes and label is in y-axes. All images of ‘On’ folder are assigned as ‘0’ and images of ‘Off’ folder are assigned as ‘1’. Further we have normalize all data because the training may not converge without scaling. Normalization is a rescaling of the initial range data. In order for all values to be between 0 and 1. CNNs have a layer of input, a layer of output and hidden layers. The hidden layers normally consist of convolutional layers, layers of relu, layers of pooling, and fully connected layers. With the receptive field size of 3 x 3 and stride 1, we have three convolution layers. There are 32 channels in the first and second convolutional layers, and 64 channels in the third convolutional layer. For each convolutional layer, 'relu' activation function is used. In this model the max pooling size is 2 x 2. As it is a binary classification the output layer shell has only one output depending on probability value. If the probability value is close to zero then the prediction will be '0' that is labeled as On and If the probability value is close to one then the prediction will be '1' that is labeled as Off. The last layer is the sigmoid layer for the class probability calculation.Figure-5.12 shows the model summary of our CNN model.

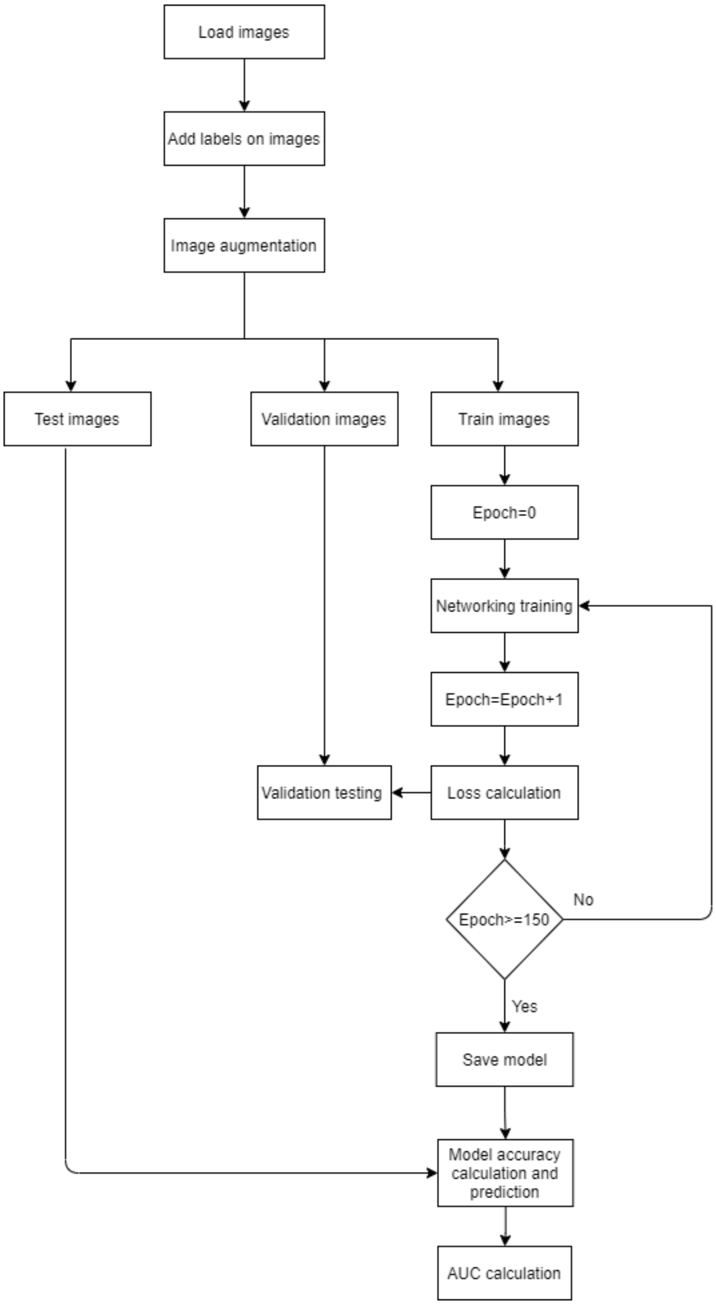


Figure 5.9: Flowchart of CNN algorithm

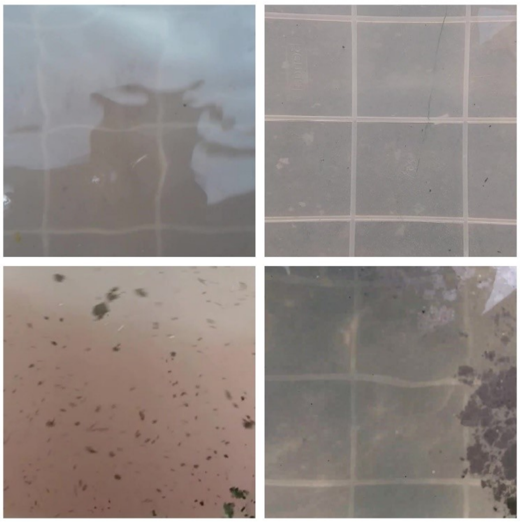


Figure 5.10: Sample images of 'on' folder

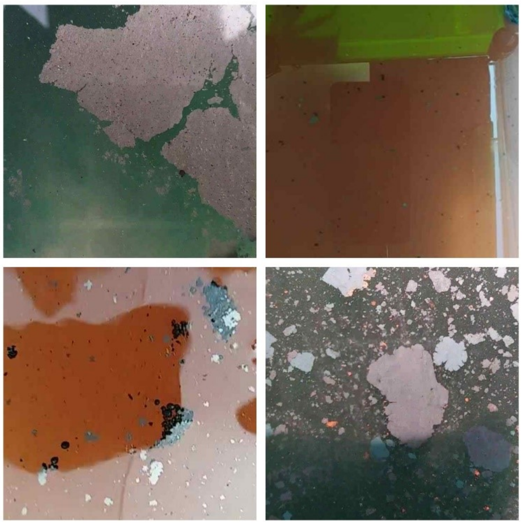
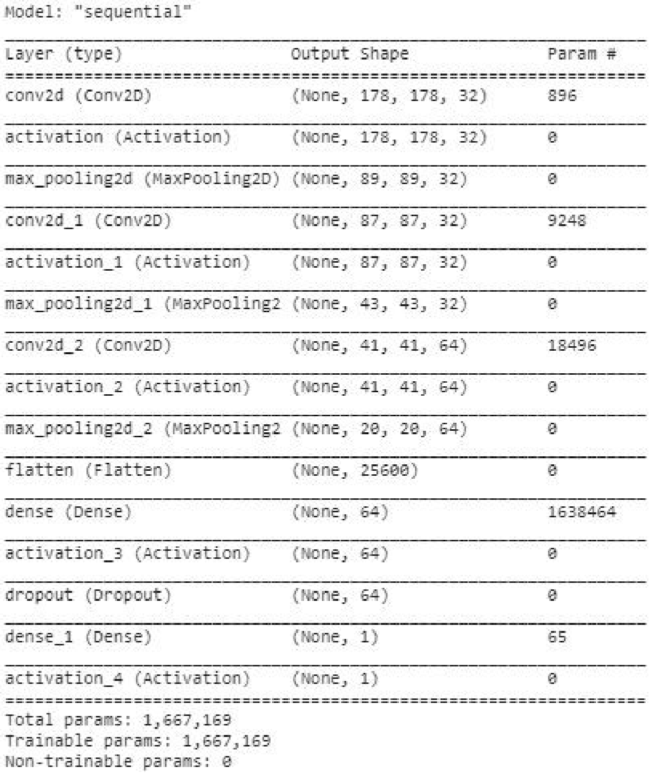


Figure 5.11: Sample images of 'Off' folder



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 178, 178, 32)      896
_________________________________________________________________
activation (Activation)      (None, 178, 178, 32)      0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 89, 89, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 87, 87, 32)        9248
_________________________________________________________________
activation_1 (Activation)    (None, 87, 87, 32)        0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 43, 43, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 41, 41, 64)        18496
_________________________________________________________________
activation_2 (Activation)    (None, 41, 41, 64)        0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 20, 20, 64)        0
_________________________________________________________________
flatten (Flatten)            (None, 25600)             0
_________________________________________________________________
dense (Dense)                (None, 64)                1638464
_________________________________________________________________
activation_3 (Activation)    (None, 64)                0
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 65
_________________________________________________________________
activation_4 (Activation)    (None, 1)                 0
=================================================================
Total params: 1,667,169
Trainable params: 1,667,169
Non-trainable params: 0
```

Figure 5.12: Model summary of CNN

After building our CNN classifier model we have classified real-time streaming video by passing it through CNN model. A basic flowchart of algorithm for video classification is given as following fig-5.13:

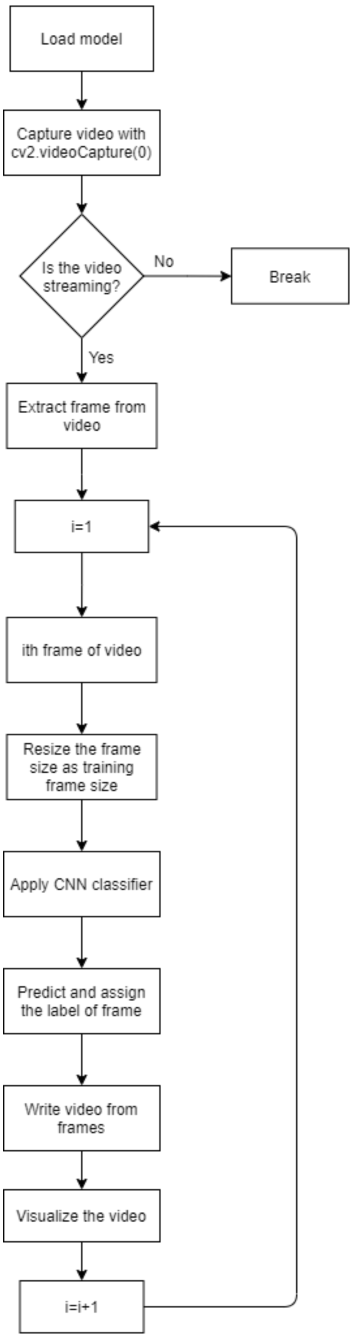


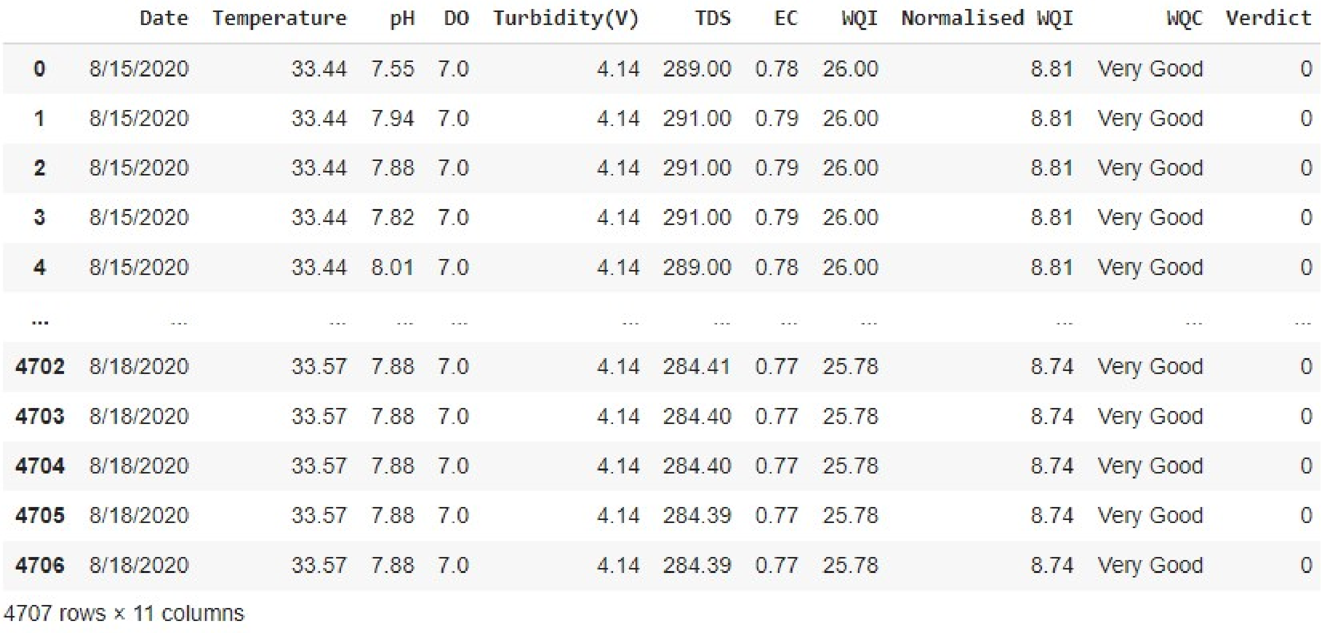
Figure 5.13: Flowchart for video classification algorithm

5.8 Long Short-Term Memory

LSTM is good at remembering previous data. As accuracy can be affected by previous inputs. Thus, it is a natural choice in order to analyze future WQI. And Seasonal outputs.

5.8.1 Data Load

We have pre-processed our data by calling the necessary libraries. From pandas, we input our data shown in fig-5.14. We clean and replace any unnecessary and error data.



	Date	Temperature	pH	DO	Turbidity(V)	TDS	EC	WQI	Normalised WQI	WQC	Verdict
0	8/15/2020	33.44	7.55	7.0	4.14	289.00	0.78	26.00	8.81	Very Good	0
1	8/15/2020	33.44	7.94	7.0	4.14	291.00	0.79	26.00	8.81	Very Good	0
2	8/15/2020	33.44	7.88	7.0	4.14	291.00	0.79	26.00	8.81	Very Good	0
3	8/15/2020	33.44	7.82	7.0	4.14	291.00	0.79	26.00	8.81	Very Good	0
4	8/15/2020	33.44	8.01	7.0	4.14	289.00	0.78	26.00	8.81	Very Good	0
...	...	...	...	...	...	...	...	...	...	...	...
4702	8/18/2020	33.57	7.88	7.0	4.14	284.41	0.77	25.78	8.74	Very Good	0
4703	8/18/2020	33.57	7.88	7.0	4.14	284.40	0.77	25.78	8.74	Very Good	0
4704	8/18/2020	33.57	7.88	7.0	4.14	284.40	0.77	25.78	8.74	Very Good	0
4705	8/18/2020	33.57	7.88	7.0	4.14	284.39	0.77	25.78	8.74	Very Good	0
4706	8/18/2020	33.57	7.88	7.0	4.14	284.39	0.77	25.78	8.74	Very Good	0

4707 rows × 11 columns

Figure 5.14: data set

5.8.2 Feature Selection

Based on the data we need to train. For that purpose, we have selected the feature to train and extract the data for visualization There were features selected based on the date.Im fig-5.15 Featured selected: ['Temperature', 'pH', 'DO', 'Turbidity(V)', 'TDS', 'EC', 'WQI', 'Normalized WQI'] in All timestamps == 4707.

5.8.3 Feature Scaling

To train all data need to be the same parameter.In order to do that we need feature Scaling in fig-5.16. For that we select feature scaling StandardScaler performs the task of Standardization. Usually, a dataset contains variables that are different in scale. The StandardScaler assumes the data is normally distributed within each feature and will scale them such that the distribution is now centred around 0, with a standard deviation of unit variance. It represents the Z-value. The formula is

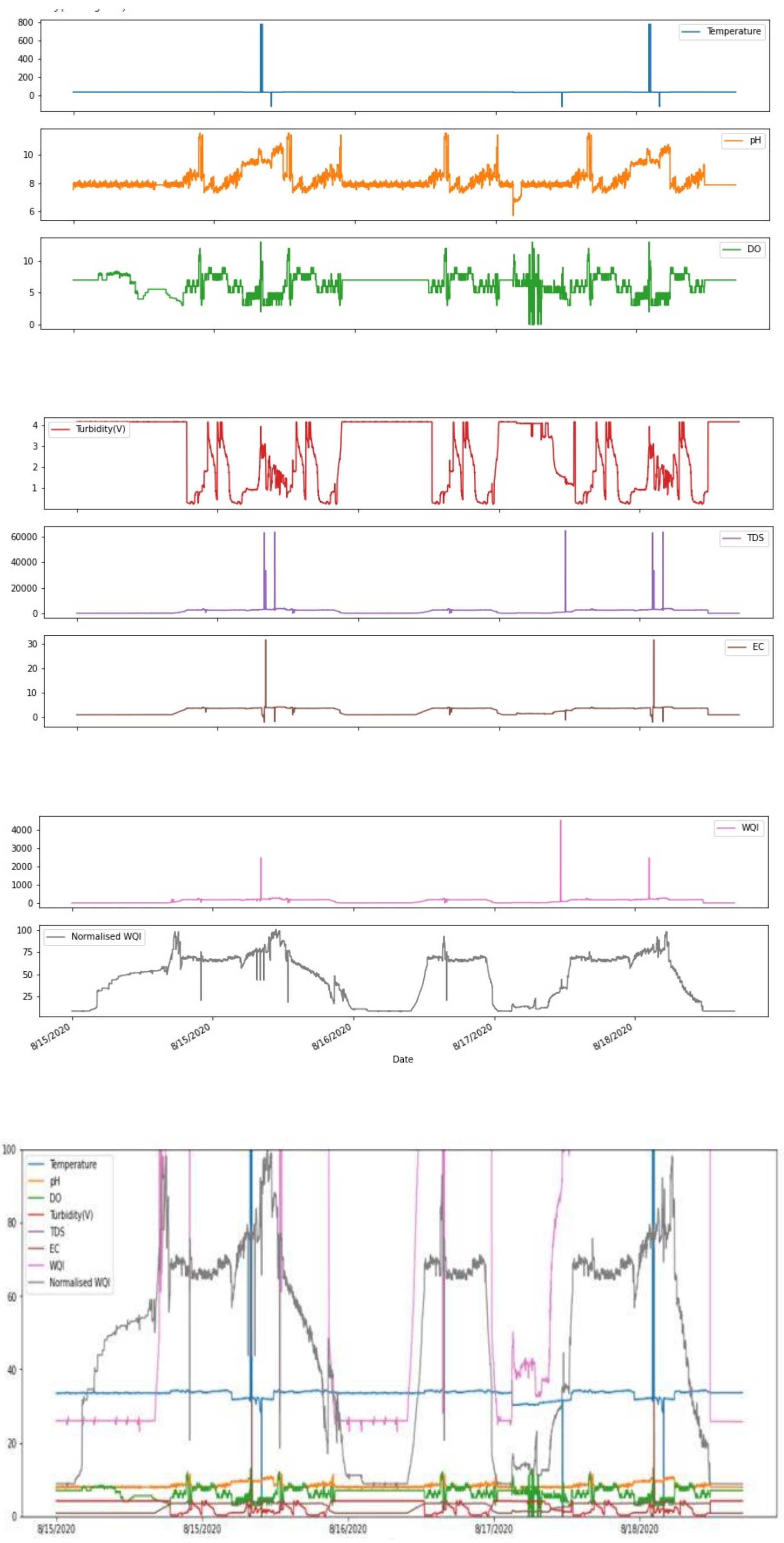


Figure 5.15: Data Visualization

given below
Standardization

$$z = \frac{x - \mu}{\sigma} \tag{5.4}$$

With mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i) \tag{5.5}$$

and Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{5.6}$$

Where,
is the mean of the population.   is the standard deviation of the population.
The absolute value of z represents the distance between that raw score x in a particular time and the population means in units of the standard deviation. z is negative when the raw score is below the mean, positive when above.

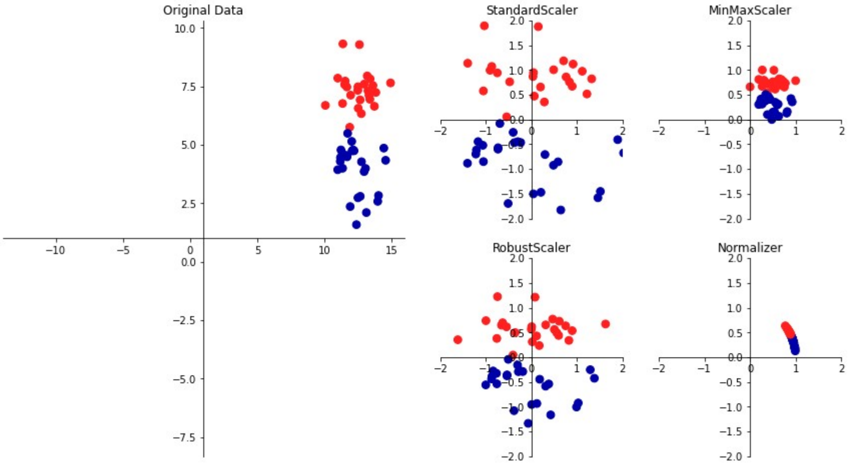


Figure 5.16: Feature Scaling

5.8.4 Modeling

LSTM Generally has 2 layers. input layer and output layer Our model is sequential. We have two input layers. One has units 64 while the other is 25 units. In first layer we keep the return_ sequence true with input being input_shape=(n_past, dataset_train.shape[1]-1). In the 2nd layer, we changed the return_sequence to false to get the same data as the 1st one. In between layers, we gave dropout layer=0.5 so that the data is not excluded randomly. Now for the Dense layer, we use 3 layers. In units=25 we used sigmoid. as gives the output range from -1 to +1 and a sigmoid curve. As our data has been scaled into around zero it will provide us with data out with will not be Underfit.

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

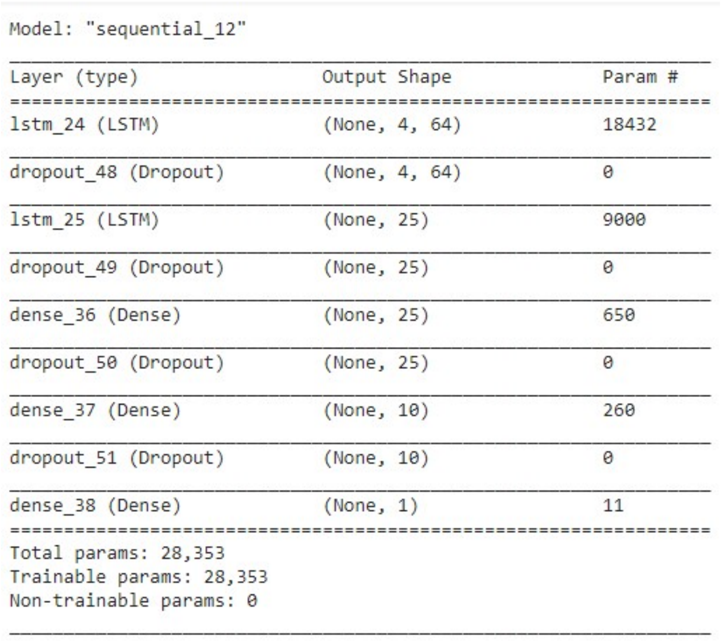
Here S(x) = sigmoid function e = Euler's number Now, in units = 10 we used an activation function called relu. rectifier linear activation. it only keeps the positive data. such that there is no negative data. for that possibility of overfitting is gone.

$$\phi(\mathbf{v}) = \max\left(0, a + \mathbf{v}'\mathbf{b}\right)$$

However, we also have our last output layer with a linear activation function. Linear Activation Function It takes the inputs, multiplied by the weights for each neuron, and creates an output signal proportional to the input. Thus, providing us with data similar to the input.

$$\phi(\mathbf{v}) = a + \mathbf{v}'\mathbf{b}$$

Dropout layer in output layer is 0.2 . for model compiling we used optimizer = Adam(learning rate = 0.01), loss = 'mean squared error', metrics=['accuracy'].This is to find loss and accuracy. As In fig-5.17 it can see out model summery for LSTM.



```
Model: "sequential_12"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
lstm_24 (LSTM)               (None, 4, 64)             18432     
_________________________________________________________________
dropout_48 (Dropout)         (None, 4, 64)             0         
_________________________________________________________________
lstm_25 (LSTM)               (None, 25)                9000      
_________________________________________________________________
dropout_49 (Dropout)         (None, 25)                0         
_________________________________________________________________
dense_36 (Dense)             (None, 25)                650       
_________________________________________________________________
dropout_50 (Dropout)         (None, 25)                0         
_________________________________________________________________
dense_37 (Dense)             (None, 10)                260       
_________________________________________________________________
dropout_51 (Dropout)         (None, 10)                0         
_________________________________________________________________
dense_38 (Dense)             (None, 1)                 11        
=================================================================
Total params: 28,353
Trainable params: 28,353
Non-trainable params: 0
_________________________________________________________________
```

Figure 5.17: Model Summery

5.8.5 Future WQI Forecasting

From the data we got we use the train data to find the pattern from past data. While analyzing given the past data we use the pattern to indicted what will the expected Normalized WQI for the system will be. According to that, we can assume the probability of the ETP being on or off.

5.8.6 Root Mean Square Error(RMSE)

RMSE is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from mean the regression line data points are. it indicates how concentrated the data is around the line of best fit. RMSE closer to 1 means the fit is bad while closer to 0 means good enough fit. If The RMSE value 0.5 then it reflects the poor ability of the model to accurately predict the analyzed data.

5.8.7 R_Squared

The R_square represents the variance of the data that can be explained by the model, let us assume the R^2=0.80 that means 0.20 of the variances cannot be explained by the model, in the logical case when R^2=1 the model completely fit and explained all variance. Explained variance which also called explained variation is used to measure the discrepancy between a model and actual data.

5.8.8 Walk forward Validation

Walk-forward is largely used for validating models that involve time series analysis[14][15][16][17] Walk-forward testing carries the thought of "out-of-sample" testing to another level. It implies to require a portion of your information to optimize a framework, and another portion of information to approve. Thus, here you optimize a window of information say past 100 bars, and after that test it on the following 20 bars. At that point roll the entire thing forward 20 bars and rehash the method. This gives you an expansive out of test period and permits you to see how steady the framework is over time.

5.8.9 Seasonality

The experience is done in winter. However, there is a seasonal impact on our analysis .the dry season had significantly higher contamination loads, which were decreased during the monsoon season[5]. Thus, our decision is based on WQI can be varied due to Season change.

5.8.10 Algorithm(Forecasting)

The working algorithm for LSTM is given below:

Algorithm 3: LSTM for WQI forecasting

```
Input: Sensor data along with WQI
Result: WQI forecasting value
df ← dataset
Set input units, LSTM units, output units and optimizer to define LSTM
  Network (L)
Feature Scaling for the dataset
for n epochs and batch size do
    rain the Network (L)
end
Run Predictions using L
Calculate the loss function
Using time series forecasting find future WQI
Find accuracy of the result
Calculate R_square to explain variance
Walk-forward model validation finding RMSE
```

Chapter 6

Working Environment

In an industry the wastewater goes into the ETP and three different types of certain materials from the wastewater. Three types of wastewater treatment process are ETP, STP, and CETP[18]. Effluent Treatment Plants or (ETPs) are utilized by driving companies within the pharmaceutical and chemical industry to filter water and expel any harmful and non-poisonous materials or chemicals from it. These plants are utilized by all companies for environment protection. An ETP may be a plant where the treatment of mechanical effluents and squander waters is done. The ETP plants are utilized broadly in mechanical division, for illustration, pharmaceutical industry, to evacuate the effluents from the bulk drugs. During the fabricating prepare of drugs, shifted effluents and contaminants are created. The gushing treatment plants are utilized within the evacuation of tall sum of organics, flotsam and jetsam, earth, coarseness, contamination, harmful, non-poisonous materials, polymers etc. from drugs and other cured stuff. The ETP plants utilize vanishing and drying strategies, and other assistant procedures such as centrifuging, filtration, cremation for chemical handling and effluent treatment. ETP has several tanks with multiple layers in between to clean the water. and one outlet. But There is main two tanks. One tank in-store the water polluted and One keeps the freshwater. The last tank sends the water to the outlet. In our environment shown in fig-6.1, we kept one box as the processed water and another box as the output. We setup our

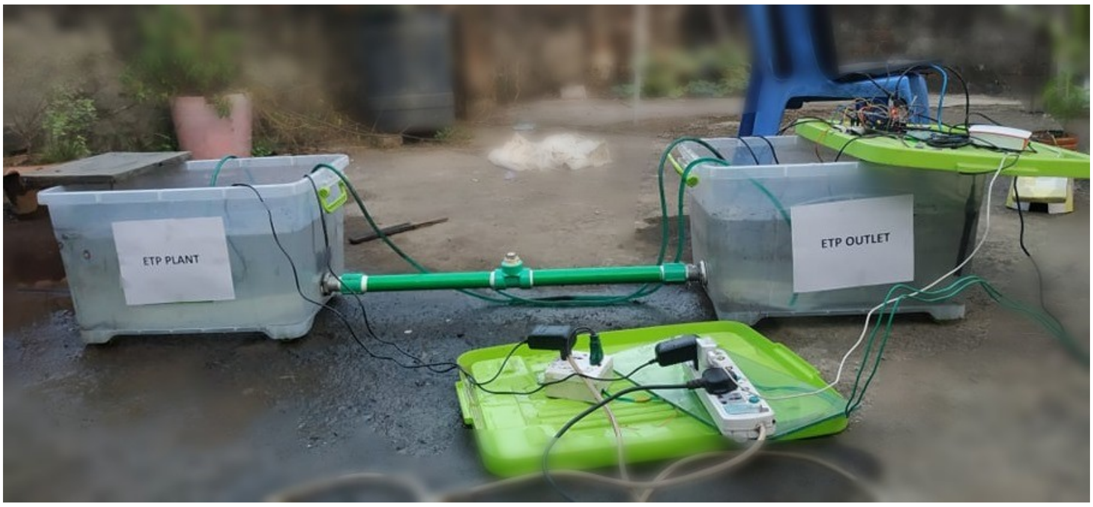


Figure 6.1: Working environment: Demo ETP plant

sensors and camera in the output box.

Our sensor setup is going to be placed in the outlet of the ETP plant which will monitor the water coming out of ETP plant. So we have setup an environment which works as a demo of the whole ETP plant and its outlet. We have taken two water tubs which can hold 40 liters of water each. One tub works as the ETP plant and the other water tub as the ETP outlet. Two tubs are connected with each other by a PVC pipe by which the water flow can be controlled. We initiate the setup by filling both of the tubs with fresh water. The sensor board along with the camera is set over the ETP outlet and the sensors are submerged in water. Then we put different types of colors along with salt and bleaching powder in the ETP plant tub to contaminate the water and slowly start to send the contaminated water to the ETP outlet tub by the help of the PVC pipe. The water of ETP outlet slowly starts to get polluted by the ETP plant water and our sensor starts to picks up the water quality change from the outlet water and sends the data to our database. At this point our setup can detect whether the ETP plant in turned on or off by the quality of water of the outlet. Then slowly we start to supply fresh water to our ETP outlet by the help of a motor and we also throw out the previous contaminated water by the help of another motor. This helps us to capture the gradual transition of water quality in our output from the sensors. We repeat the process several times to capture the transition and time period of the ETP being turned on or off.

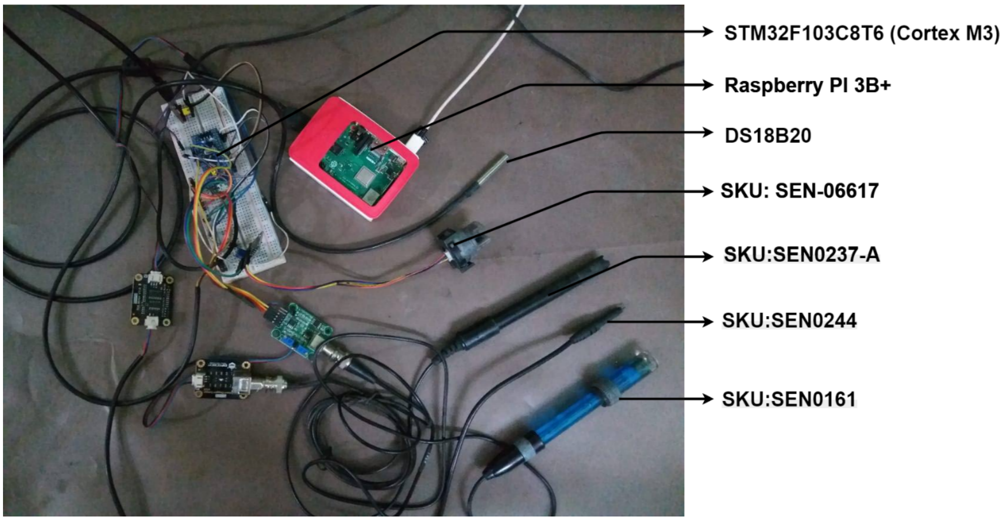


Figure 6.2: Hardware Setup

Here, as it can be seen in fig-6.2 we have created an embedded system for measurement of data sample data. We used STM32 as the core of this process. It was used as micro controller to process the data of the sensors. here we used water temperature sensor , turbidity sensor , TDS sensor, DO sensor , pH sensor to measure temperature , transparency in our water , conductivity , dissolved chemicals or solid , oxygen level and if the water is acidotic or alkalinity . we used a raspberry pi as microprocessor to extract the data from stm32 and used its built—in Wi-Fi module to connect to SQL using internet in real time in order to check seasonality and monitor the data . There we used FTDI UART to connect the stm32 and raspberry pi using USB cable. The system was powered by AC connector which is connected in Raspberry pi.

Chapter 7

Output and Result Analysis

7.1 RGB color analysis

In this paper RGB color analysis is used for data validation, monitoring and tracking. In figure-7.1, figure-7.2 and figure-7.3 is the visual representation of RGB color analysis. The value of three components of RGB model-red, green and blue are between 0 and 255 whereas less the value indicates dark color and the color is getting lighter as the value is getting larger. Here average RGB value is plotted against date.

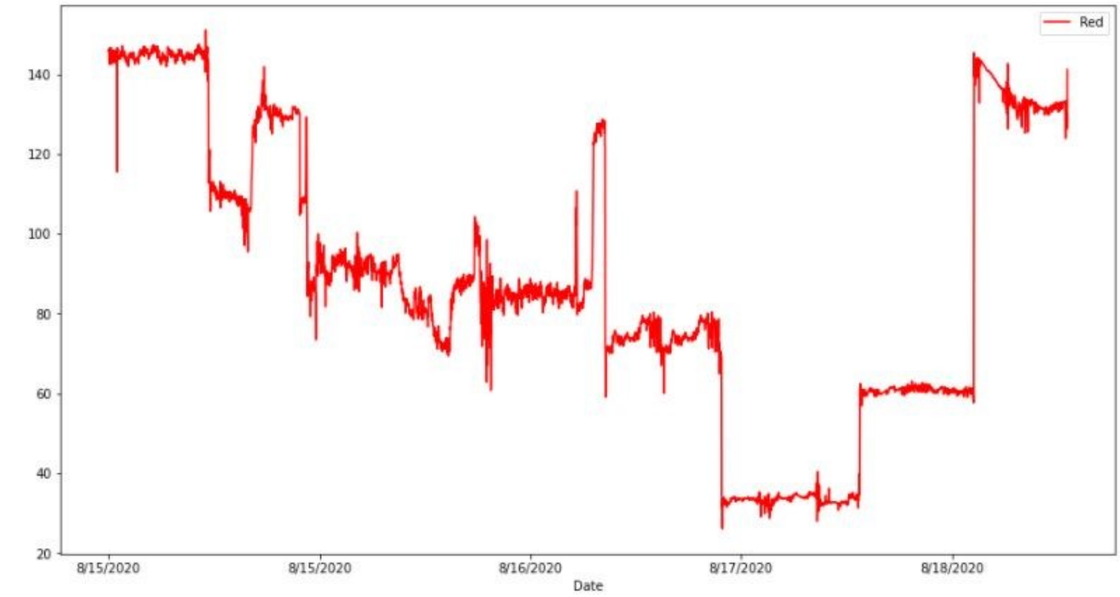


Figure 7.1: Red

Figure-7.4 represent the visual RGB color analysis along with WQI for monitoring the verdict, tracking the transition and validating the sensor value. Here RGB value and WQI value is plotting altogether whereas the range of WQI is 0 to 100. The horizontal line y=50 represents the markup line of verdict in terms of WQI, below 50 represents the ETP is on and above till 100 represents the ETP is off. This visualization shows the RGB color condition of water in terms of WQI where lower

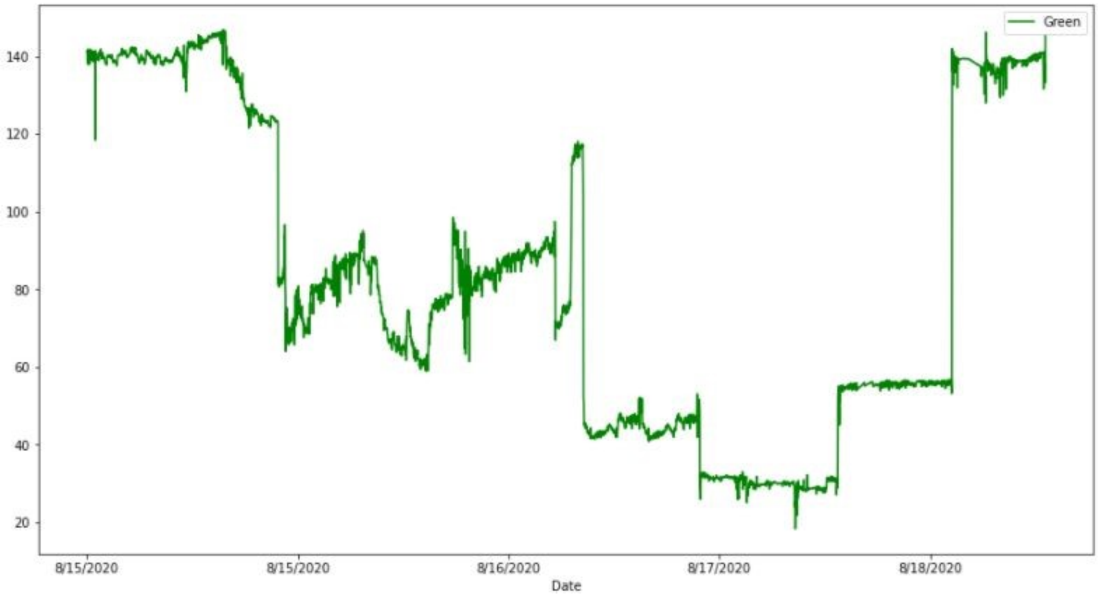


Figure 7.2: Green

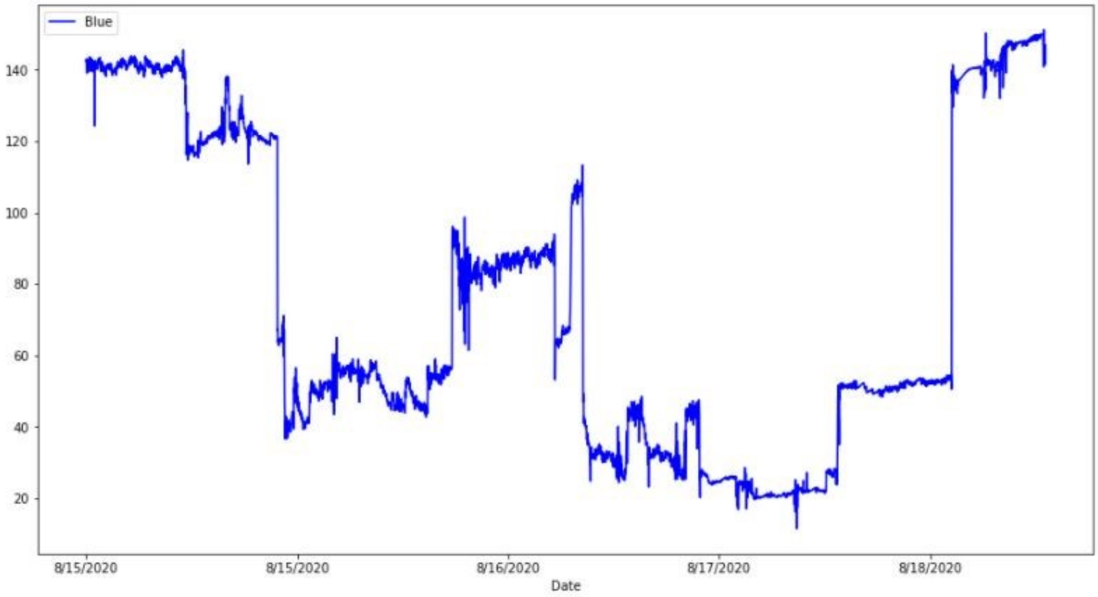


Figure 7.3: Blue

the WQI value greater the RGB value indicates the ETP plant is decolorizing the coloring agent that is present in waste water along with refine the water and make it usable.

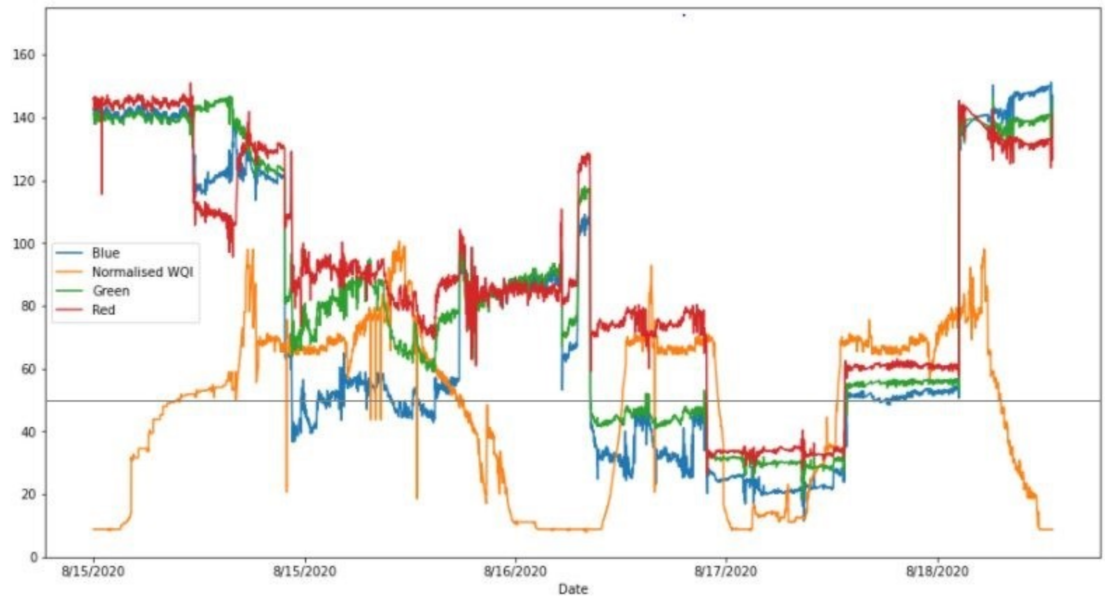


Figure 7.4: RGB color analysis against WQI

7.2 Video classification result using CNN

In this paper we have used CNN binary classification model for video classification to predict the On/Off condition of ETP. There are 2289 images in 'On' folder and 2820 images in 'Off' folder. 150 epochs are trained and tested in our model. After successfully train and test through our model we have obtained 97.3% accuracy that represents a good fit of our model. Figure-7.6 and Figure-7.5 shows the Training and validation loss graph and Training and validation accuracy graph respectively.

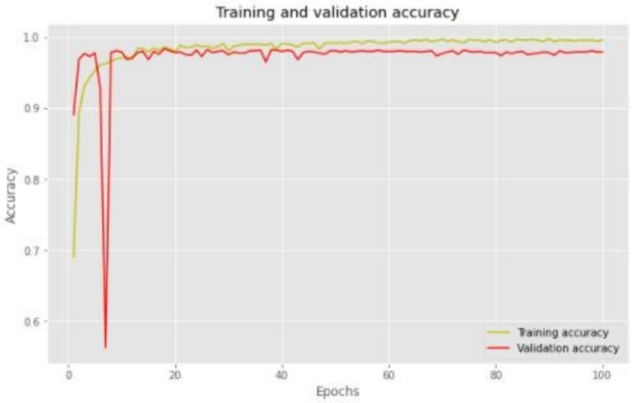


Figure 7.5: Training and validation accuracy graph

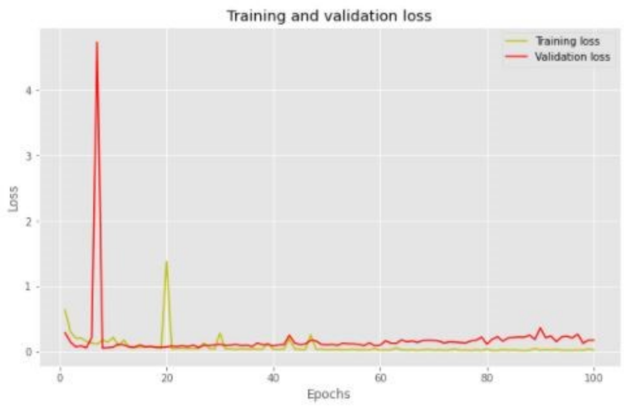
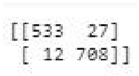


Figure 7.6: Training and validation loss graph

In Figure-7.6 we have seen that training loss and validation loss decreases to a point of stability with a minimal gap between the two final loss values. Figure-7.7 shows the Receiver Operating Characteristic (ROC) curve that represents the capability of your model to distinguish between classes. Accuracy of ROC curve we get is 99%.The ROC curve is a performance measurement at different threshold settings for classification problems. Out of all positive observations (TP/(TP + FN)), the true positive rate is the proportion of observations that were correctly predicted to be positive. Similarly, out of all negative observations (FP/(TN + FP)), the false positive rate is the proportion of observations which are incorrectly expected to be positive. The optimal threshold point we have used in our model is 0.577. Out of all positive observations (TP/(TP + FN)), the true positive rate is the proportion of observations that were correctly predicted to be positive. Similarly, out of all negative observations (FP/(TN + FP)), the false positive rate is the proportion of observations which are incorrectly expected to be positive. Hence we get the confusion as following-



```
[[533  27]
 [ 12 708]]
```

Here the ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

After successfully create our CNN classification model we can classify real-time streaming video of ETP outlet by passing it to our model. This classification model can predict the On/Off condition of ETP. Figure-7.8 shows some successful sample result of video classification. Here ‘0’ represents the ETP is on and ‘1’ represents the ETP is off. The sample output of video classification is added here: https://youtu.be/3VuL2lpB4Xc

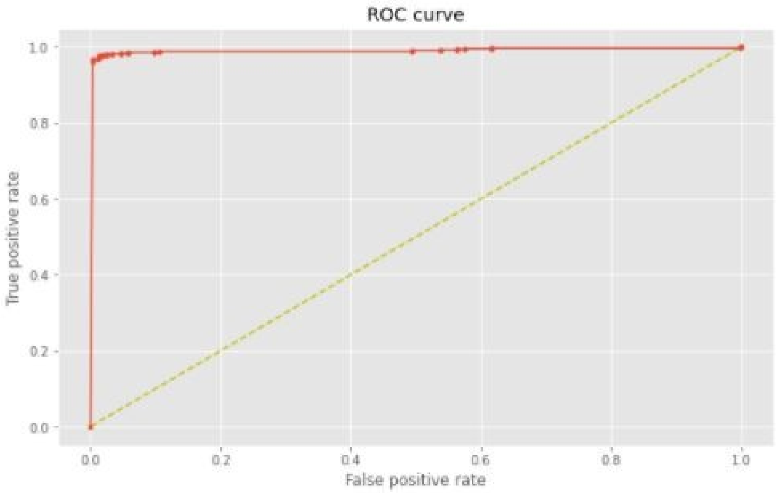


Figure 7.7: ROC curve

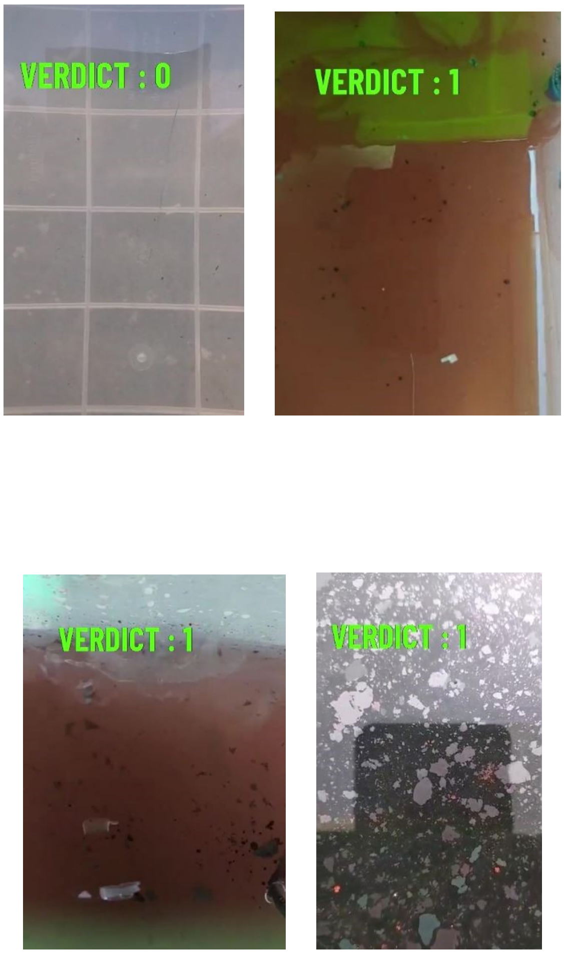
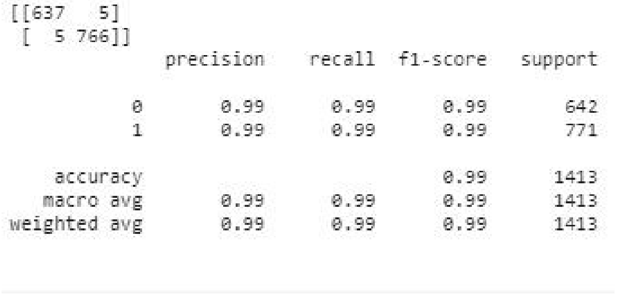


Figure 7.8: Sample prediction of video classification

7.3 Classification and prediction result using KNN

One of the simplest algorithms for classification is the KNN algorithm. It can offer highly competitive outcomes, even with such simplicity. In this paper we have implanted KNN classifier to both classify the Water Quality Class (WQC) and predict the verdict. Though we already have used four chemical and two physical sensor to calculate Water Quality Index (WQI) and WQC, we use this ML approach to avoid any miss prediction because of sensor error. After evaluation we have found the accuracy for prediction is 99% an accuracy for WQC classification is 98%. Thus, it will be simple to use the KNN algorithm for classification and get optimum accuracy by taking into account all the parameters that are discussed. Figure-7.9 and Figure-7.10 shows the confusion matrix and accuracy result of prediction and classification respectively.

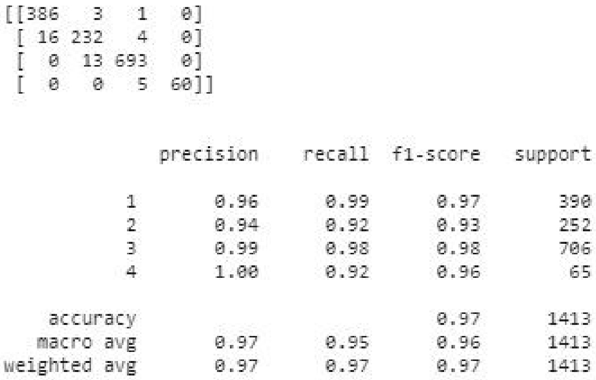


```
[[637   5]
 [  5 766]]
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       642
           1       0.99      0.99      0.99       771

    accuracy                           0.99      1413
   macro avg       0.99      0.99      0.99      1413
weighted avg       0.99      0.99      0.99      1413
```

Figure 7.9: Result analysis for prediction



```
[[386   3   1   0]
 [ 16 232   4   0]
 [  0  13 693   0]
 [  0   0   5  60]]

              precision    recall  f1-score   support

           1       0.96      0.99      0.97       390
           2       0.94      0.92      0.93       252
           3       0.99      0.98      0.98       706
           4       1.00      0.92      0.96        65

    accuracy                           0.97      1413
   macro avg       0.97      0.95      0.96      1413
weighted avg       0.97      0.97      0.97      1413
```

Figure 7.10: Result Analysis for Multiclass Classification

7.4 WQI Forecasting result using LSTM

In this paper, with batch size 64 and 100 epoch we fit the model.The Validation split is 0.2 and we call the early Stopped method to stop when the value is found and thus saving time, we use the model check to count loss function. We also called the method which is used to reduce the learning rate when a metric has stopped

improving. For 100 epoch we got loss: 0.1601 - accuracy: 0.6595 - val loss: 0.1237 - val accuracy: 0.6809 in fig-7.11, from in fig we have seen that training loss and validation loss decreases to a point of stability with a minimal gap between the two final loss values. The same can be witnessed for training and validation accuracy in 7.12 We know
1) Underfitting This is the only case where loss >> validation loss, but only slightly, if the loss is far higher than validation loss, in another word model cannot fit data correctly.

2) Overfitting loss << validation loss. This means that the model is fitting very nicely to the training data but not at all the validation data, in other words, it's not generalizing correctly to unseen future data

3) Perfect fitting Loss == validation loss
If both values end up being roughly the same and also if the values are converging. We can also say if the accuracy is very good. And validation accuracy not good then is a worse fit.
However, in fig-7.13 our Training Loss and Validation Loss is almost equal respectively 0.16 and 0.13. Whereas we can also see accuracy and validation accuracy are both good enough in 66.65% and 68.75% respectively. This is why we can say our model is fit enough.

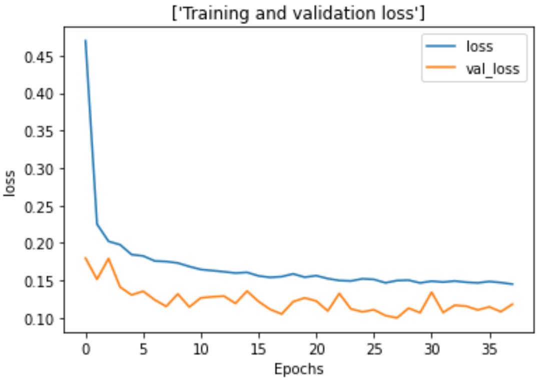


Figure 7.11: Loss V/S Epoch

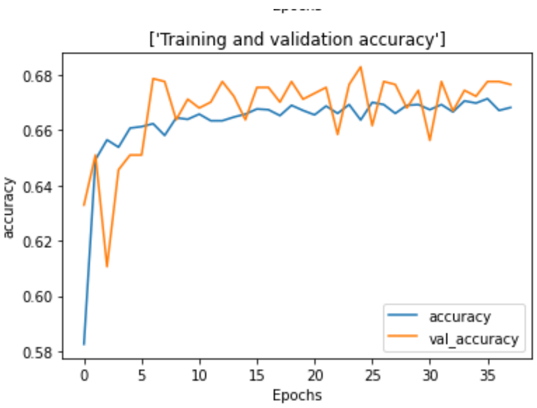
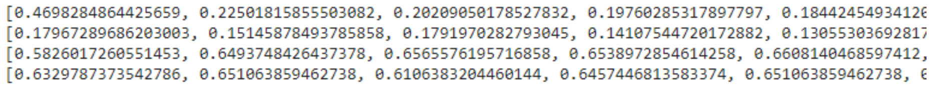


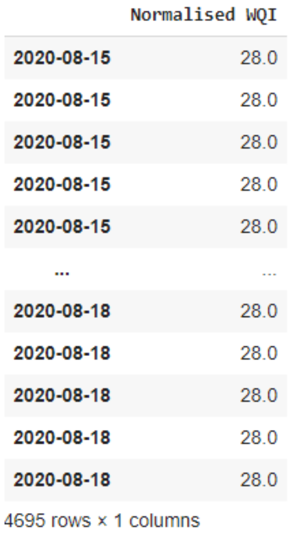
Figure 7.12: Accuracy V/S Epoch



```
[0.4698284864425659, 0.22501815855503082, 0.20209050178527832, 0.19760285317897797, 0.18442454934126
[0.17967289686203003, 0.15145878493785858, 0.1791970282793045, 0.14107544720172882, 0.13055303692817
[0.5826017260551453, 0.6493748426437378, 0.6565576195716858, 0.6538972854614258, 0.6608140468597412,
[0.6329787373542786, 0.651063859462738, 0.6106383204460144, 0.6457446813583374, 0.651063859462738, 6
```

Figure 7.13: loss , val_loss . accuracy , val_accuracy(top to down)

After fitting our data we can use that data to predict and forecast future data according to date. In Figure-7.14, we can see the past data we used to train our model and found the pattern.

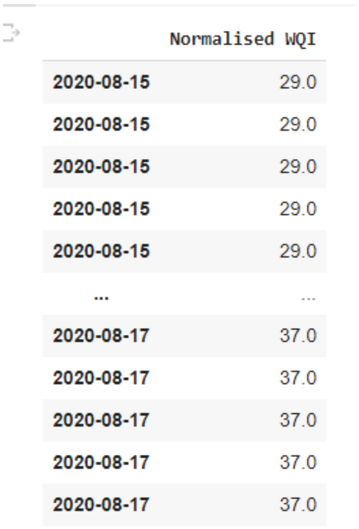


	Normalised WQI
2020-08-15	28.0
2020-08-15	28.0
2020-08-15	28.0
2020-08-15	28.0
2020-08-15	28.0
...	...
2020-08-18	28.0
2020-08-18	28.0
2020-08-18	28.0
2020-08-18	28.0
2020-08-18	28.0

4695 rows × 1 columns

Figure 7.14: Normalized WQI trained In the Past

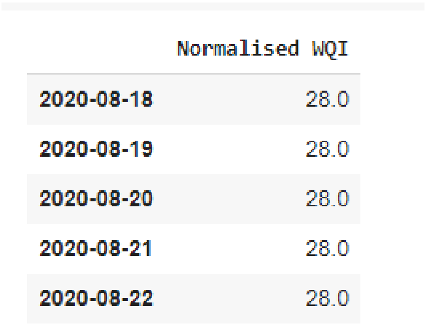
For better understanding here is the 1st 2500 past data shown in Figure-7.15



	Normalised WQI
2020-08-15	29.0
2020-08-15	29.0
2020-08-15	29.0
2020-08-15	29.0
2020-08-15	29.0
...	...
2020-08-17	37.0
2020-08-17	37.0
2020-08-17	37.0
2020-08-17	37.0
2020-08-17	37.0

Figure 7.15: First 2500 Normalized WQI

In Figure-7.16 we have got our future expected Normalized WQI. This is a fit enough model for giving a decision based on the forecast data. The approximate value is 28.



	Normalised WQI
2020-08-18	28.0
2020-08-19	28.0
2020-08-20	28.0
2020-08-21	28.0
2020-08-22	28.0

Figure 7.16: Forecasted Normalised WQI

In Figure-7.17 we can see the actual Normalised WQI , And the trained Normalised WQI and predicted Normalised WQI respect to Dates.

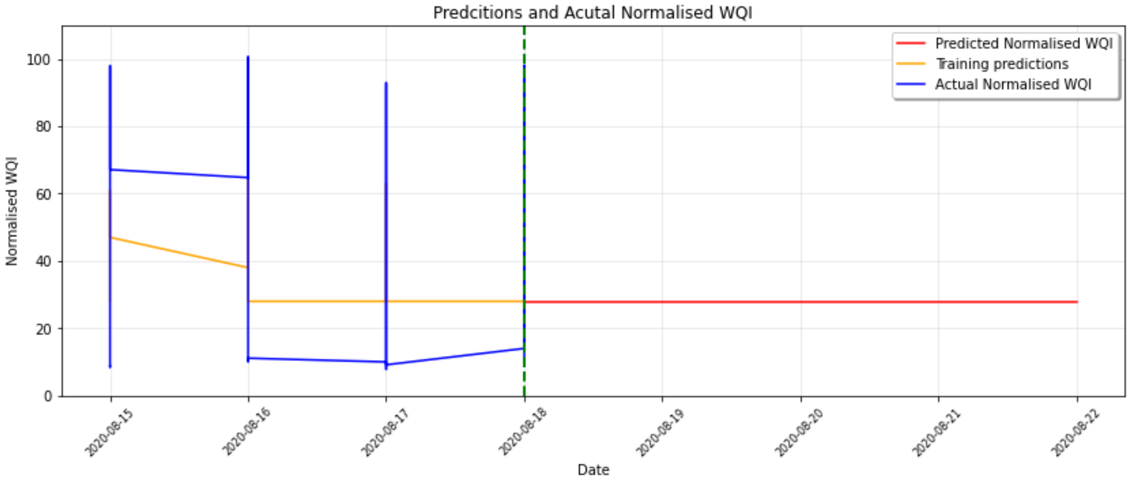
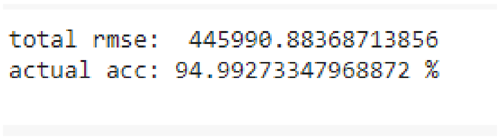


Figure 7.17: WQI Forecasting

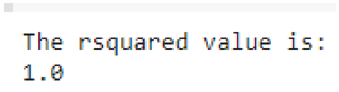
In Figure-7.18 we can see the accuracy of our prediction is 94.99% . we can conclude our future data is 95% accurate



```
total rmse:  445990.88368713856
actual acc: 94.99273347968872 %
```

Figure 7.18: Accuracy of the Result

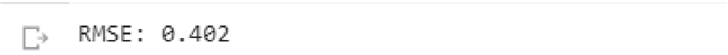
Figure-7.19 shows the R_Square value which is 1.0.So all the variance are explained. Therefore, the data is best fitted. So we can figure out the probability of future dates the ETP likely to be on.



```
The rsquared value is:
1.0
```

Figure 7.19: R_Square Value

In walk forward validation we can check the model validation out of sample data in which we got RMSE = 0.402 shown in Figure-7.20 . As we know if the RMSE is below 0.5 , the model is good . So , we can deduce our model is good enough to predict and forecast if the ETP will be on or off in future



```
RMSE: 0.402
```

Figure 7.20: RMSE from Walk Forward Validation

From our RGB analysis data, we can see if the Normalized WQI is 0 to 50 then the ETP is ON otherwise off. However, our model can varied season by season. The upper limit will change season by season. In this [19] we saw the dry season have higher pollution than the monsoon we can say winter is in the middle. So, due to weather, we assumed the normalized WQI might be 60 maximums to be on during summer. In winter it can below 40 . for 1 mm rain it can be 47, 1mm 42, 3mm below 40. Thus, it can say avg 40 when is we can decide on maximum or off of ETP.

Chapter 8

Limitations

Our research took place during the time of COVID-19. The research was supposed to be conducted by going to a particular industry and setting the sensor module over the water stream of ETP outlet. But that was not possible due to the lockdown. So a demo environment was setup to execute the research. During the data acquisition period, an endless stream of water was not possible to sustain so sometimes the water level went down from the sensor levels. Therefore, at times the data that was taken was null and some values were not feasible. A lot of garbage data was generated by the sensors for not being submerged in water for short a period of time. Also the sensors which was used are applicable for laboratory testing but they are not made for industrial applications. For this a lot of technical difficulties occurred. So if the whole setup is to be made, advanced sensors will be required.

The LSTM algorithm deployed in the research needs training data generated for a long period of time. This was not possible because our demo could work only under our supervision so we collected the data only for 4 days which is not enough for the prediction analytic. If the data is collected for a longer period of time then the result will be more accurate and helpful. The images taken by the camera has shadow of the object which was used to cover the sunlight.

Chapter 9

Conclusion

Bangladesh is a country where industries can be anywhere and most of them are beside river. As there is resource of fresh water is limited and it is threatened by waste especially chemical wastage. Thus, we need to protect it for our own benefit. Hence, Effluent Treatment Plant was established to process the contaminate water to usable water. However, it is hard and almost impossible to ensure the regulation of ETP followed by industries. In this paper we tackle the main problem of the regulation of ETP plant follows or not. In order to overcome such problem, we have come with a monitoring system for the outlet of the plant using real time E-IoT based automation using stm32 and raspberry pi 3B+ along with real time video classification using CNN. With those results we have predicted and forecast the possibility of their future deeds. The ETP monitoring system we have built takes the monitoring aspect to the next level. Our system not only provides us with the information whether the ETP is turned on or off by processing the images and sensor data but also tells us the quality of water going out of the ETP outlet. The system can be built in any industry at a very low cost. The ETP can be monitored remotely from far and the system will provide enough information about the service of the ETP plant. Overall this research will greatly benefit for nutrition and agriculture field by monitoring the industrial wastage.

Our research has created a setup which can be changed and improved in many ways to deploy in different fields in the future. In this arrangement, we have measured the quality of water to monitor the ETP. In future, chemicals can be monitored using appropriate sensors to find out which chemicals are being used in the industry. Moreover, we have used a digital camera to stream the video and get the RGB frames. A Hyper Spectrum camera can be used in the future to get better and more accurate results. Also the whole RGB color analysis can be done using satellite camera on wider fields like rivers, lakes etc. As well the result output of the whole analysis can be displayed using a website and a mobile application to get an overview of the water quality remotely.

Bibliography

[1] P. H. Gleick, "Water in crisis," *Pacific Institute for Studies in Dev., Environment & Security. Stockholm Env. Institute, Oxford Univ. Press. 473p*, vol. 9, 1993.

[2] N. T. Chowdhury, "Water management in bangladesh: An analytical review," *Water policy*, vol. 12, no. 1, pp. 32–51, 2010.

[3] R. Soja and Ł. Wiejaczka, "The impact of a reservoir on the physicochemical properties of water in a mountain river," *Water and Environment Journal*, vol. 28, no. 4, pp. 473–482, 2014.

[4] R. Arridha, S. Sukaridhoto, D. Pramadihanto, and N. Funabiki, "Classification extension based on iot-big data analytic for smart environment monitoring and analytic in real-time system," *International Journal of Space-Based and Situated Computing*, vol. 7, no. 2, pp. 82–93, 2017.

[5] S. Anjana, M. Sahana, S. Ankith, K. Natarajan, K. Shobha, and A. Paventhan, "An iot based 6lowpan enabled experiment for water management," in *2015 IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS)*, IEEE, 2015, pp. 1–6.

[6] P. Liu, J. Wang, A. K. Sangaiah, Y. Xie, and X. Yin, "Analysis and prediction of water quality using lstm deep neural networks in iot environment," *Sustainability*, vol. 11, no. 7, p. 2058, 2019.

[7] L. Zhang and X. Chen, "Research of water quality monitoring and control system based on stm32," in *2018 International Conference on Mechanical, Electrical, Electronic Engineering & Science (MEEES 2018)*, Atlantis Press, 2018.

[8] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[9] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for iot big data and streaming analytics: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2923–2960, 2018.

[10] S. Dendukuri, S. Asadi, and M. Raju, "Estimation of water quality index by weighted arithmetic water quality index method: A model study," *International Journal of Civil Engineering and Technology*, vol. 8, pp. 1215–1222, Jan. 2017.

[11] K. Thirunavukkarasu, A. S. Singh, P. Rai, and S. Gupta, “Classification of iris dataset using classification based knn algorithm in supervised learning,” in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, IEEE, 2018, pp. 1–4.

[12] S. B. Imandoust and M. Bolandraftar, “Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background,” *International Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605–610, 2013.

[13] N. N. Xiong, Y. Shen, K. Yang, C. Lee, and C. Wu, “Color sensors and their applications based on real-time color image segmentation for cyber physical systems,” *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, pp. 1–16, 2018.

[14] K. Żbikowski, “Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy,” *Expert Systems with Applications*, vol. 42, no. 4, pp. 1797–1805, 2015.

[15] N. Kohzadi, M. S. Boyd, B. Kermanshahi, and I. Kaastra, “A comparison of artificial neural network and time series models for forecasting commodity prices,” *Neurocomputing*, vol. 10, no. 2, pp. 169–181, 1996.

[16] A. Dantas and J. Seixas, “An adaptive neural system for financial time series tracking,” in *Adaptive and Natural Computing Algorithms*, Springer, 2005, pp. 421–424.

[17] L.-J. Cao and F. E. H. Tay, “Support vector machine with adaptive parameters in financial time series forecasting,” *IEEE Transactions on neural networks*, vol. 14, no. 6, pp. 1506–1518, 2003.

[18] P. Mondal, *Types of wastewater treatment porcess: Etp, stp and cetp*, 2016.

[19] M. Islam, M. K. Uddin, S. M. Tareq, M. Shammi, A. K. I. Kamal, T. Sugano, M. Kurasaki, T. Saito, S. Tanaka, H. Kuramitz, *et al.*, “Alteration of water pollution level with the seasonal changes in mean daily discharge in three main rivers around dhaka city, bangladesh,” *Environments*, vol. 2, no. 3, pp. 280–294, 2015.