A Framework for Water Pollution Monitoring System Using Sensors and Machine Learning Technique

By: Elias Mohammed

A Thesis Submitted to

Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment for the Degree of Masters

of Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

January 2020

Adama, Ethiopia

A Framework for Water Pollution Monitoring System Using Sensors and Machine Learning Technique

By: Elias Mohammed

Name of Advisor: Prof (Dr.). Jayant Sheker

A Thesis Submitted to

Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment for the Degree of Masters

of Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

January 2020

Adama, Ethiopia

DECLARATION

I hereby declare that this MSc thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all source of material used for the thesis has been fully acknowledged.

Name: Elias Mohammed

Signature: ___________________________

This MSc thesis has been submitted for examination with my approval as

thesis Advisor

Name: Prof (Dr). Jayant Sheker

Signature: __________________________________

Date of submission: ____________________________

APPROVAL BOARD OF EXAMINERS

We, the undersigned, member of the Board of Examiners of the final open defense by "*Elias Mohammed*" have read and evaluate his/her thesis entitled "*A Framework for Water Pollution Monitoring System Using Sensors and Machine Learning Technique*" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the degree of Masters in Computer Science and Engineering (CSE).

______________________	____________________	__________________
Supervisor/Advisor	Signature	Date
______________________	____________________	__________________
Chairperson	Signature	Date
______________________	____________________	__________________
Internal Examiner	Signature	Date
______________________	____________________	__________________
External Examiner	Signature	Date

ACKNOWLEDGMENT

First, and foremost, I believe that thank Allah for granting my knowledge, health, and the patience to complete this thesis.

I respect and thank Mesfin Abebe (Ph.D.), for providing me an opportunity to do the research work in Real-time Water Pollution Monitoring Systems and giving me all support and guidance, which made me complete the research duty. I am incredibly thankful to him for providing such helpful support and guidance, although he had a busy schedule managing the coordinate affairs.

I would also like to show my appreciation to Prof. (Dr. Jayant Shaker), my advisor, for excellent feedback and continuous encouragement of constructive feedback for the report and valuable comments in the progress.

I would also thank my family, Mohammed Gena, my father, Caltu Mohammed, my sister, Kedir Mohammed, my brother, and all my family, those who helped me financially and morally in supporting and encouraging to me in this study time.

Finally, I would also like to show the most profound sense of gratitude to my best friends Kemal Temam, Musa Ahmed, and Kemal Jado. Additionally, I also would like to thank my fellow Smart Software SIG (3S) lab mates for their appreciation for finishing this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATION AND ACRONYM

AARC	Assela Agricultural Research Center
AMF	Assela Malt Factory
BOD	Biological Oxygen Demand
BPNN	Back Propagation Neural Network
CA	Clustering Analysis
CCUD	Chine Centered for Urban Development
COD	Chemical Oxygen Demand
DA	Discriminate Analysis
DO	Dissolve Oxygen
EC	Electrical Conductivity
ELM	Extreme Learning Machine
FAO	Food and Agricultural Organization
FDI	Framework for the Development of Environmental Information
FN	False Negative
FP	False Positive
HTML	Hypertext Markup Language
IoT	Internet Of Things
KNN	K-Nearest Neighbors
KPCA	Kernel Principal Component Analysis
LR	Lasso Regression
LSSVM	List Square Support Vector Machine
MLRM	Multi Leaner Regression Analysis
NP	Non-Determinant
NTU	Nephelometric Turbidity Unit
PC	Python Code
PCA	Principal Component Analysis
PH	Potential Hydrogen
PM	Process Monitoring

QC	Quality Control
RFE	Recursive Feature Elimination
SVM	Support Vector Machine
SVM	Support Vector Machine
SWM	Solid West Management
TDS	Total Dissolved Solids
TN	True Negative
TP	True Positive
TSS	Total Suspended Solid
UFS	Universal Feature Selection
UV	UltraViolet
WHO	World Health Organization
WWTP	West Water Treatment Plant

ABSTRACT

Water is the core and most relevant among the five core elements in natural existence and survival of the entire life on the globe. Here we are studying about water pollution issue that has unfortunately become a common problem that occurs in developing countries due to the substantial growth of industrial and urbanization. Polluted water out from industrial can cause a series of problems on the plantation and directly or indirectly Cause health risk on the human being. The core objective of this research study is to design a framework for water pollution classification and prediction through the integration of Sensors and machine learning technique technologies. As a solution to this problem, this research proposed polluted water classification and prediction using machine learning and feature selection techniques to build prediction models. A polluted water dataset gains from online for sample dataset and real-time data collected using the PH sensor, and waterproof temperature sensor through the Arduino Uno communication with Tera term and PuTTY. The sample dataset labeled with the help of domain experts and classified into a binary class that defined as Suitable and Polluted water for reused. This research study employed an experimental approach to determine the best combination of the machine learning algorithms and feature selection technique-based model. SVM, KNN, DT, and RF models trained the whole sample dataset with the feature selection method, RFE, UFS, and LR. The models were evaluated using 10-fold cross-validation to evaluate accuracy, misclassification, and classification performance, measured using well-known metrics; precision, recall, F1-score. Using a sample dataset, the study developed binary class models on selected feature importance. The models based on RF with RFE achieve slightly better performance than the SVM, KNN, and DT models for binary classification. Finally, the prediction of RF with RFE results in an accuracy of approximately 97%. From this result, it is possible to conclude that the RF analysis models should be preferred, since it has a better prediction performance and is less-error prone on classification compared to the SVM, KNN, and DT analysis approaches.

Keywords: *water pollution, PH sensor, Waterproof temperature sensor, machine learning, Tera term, PuTTY.*

CHAPTER ONE

1. Introduction

This chapter introduces the background of the study, Motivation of the study, Statement of the problems, Significance of the study, Research questions, Objective of the study, Scope of the study, limitation of the study, and application of the result, followed by the organization of the research study.

1.1. Background of the Study

Water is the core and most relevant among the five core elements in natural existence and survival of the entire life on the globe. We use water for a different purpose, including for drinking, cooking, personal hygiene, agricultural activities, industrial process, and all in all in everyday activities. In addition, plants and animals also depend on water for their basic survival. In short, all living organisms need a large quantity and good quality of water to continue their process of life and to give fruit full results [1]. Also, Water is one of the common factors for qualities in smart cities, urban development plans, and vital resources for humans, animals, and plants [2].

Polluted water is the world's biggest problem in health risk and environmental effects. Water pollution is releasing a contaminant into the environment that causes harm, illness, disability, and death worldwide. Water pollution and environmental factor can be made by human activity and human force in rare [2]. During the raining time, water on the road flows into rivers and lakes, carrying different toxic chemicals and picks up into the water, pick up dirt into the water, trash, carrying disease-carrying organisms along into the water, and makes polluted water and polluted environments. Many other water resources due to lack of basic protections, they vulnerable to pollution from factory farms and industrial plants. Due to that, a classification model or framework of water pollution protection way required to present the excellent worth of the water environment [3].

Currently, water quality reduced due to different water pollution caused by the disposal of pointed and non-pointed sources of water pollution. Among them, human waste, industrial waste, and automobile waste are the core issues in reducing smart cities, city development plans or urbanization, and cause a health problem in human life. To overcome all problems listed above, water quality evaluation, water pollution monitoring, and prediction using machine learning techniques become an important and hot issue to study water pollution monitoring in smart city development and urbanization plan [4]. As one of the researchers in Kenya, Dhaka city-states that the highest percent of environmental pollution and the current issue is represented below figure 1.1.

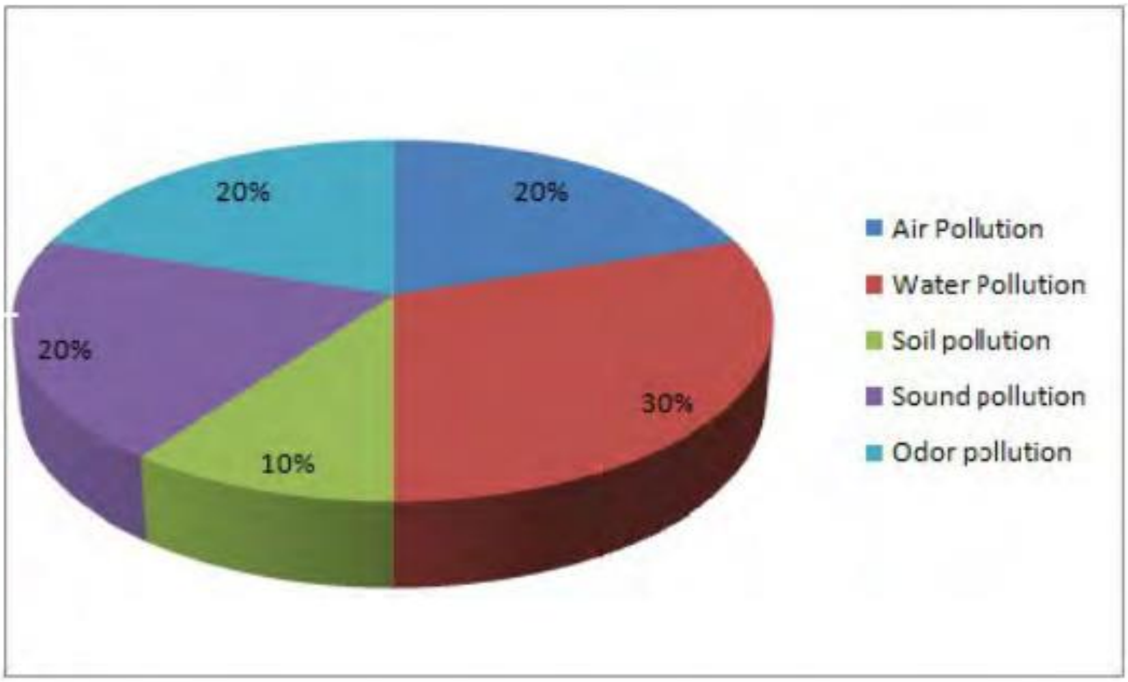

Figure 1.1: Statistical Representation of Environmental Pollution in Dhaka City [13].

Water pollution takes place either on the surface or underground levels. Underground water consists of water that has percolated below the topsoil. Surface water includes streams, rivers, oceans, lakes, runoff, and usually polluted through industrial wastewater that contains suspended solids, dissolve solid, and biological material when discharged to the environment [2]. In this study, water quality control, water pollution monitoring is selected as an application field to demonstrate the proposed approach in achieving high accuracy. It is the most significant to clarify that water is an essential element for life and has an active link between water and health. Besides that, safe water is a key to life and having poor quality water intact the environment and harm the wildlife, as well as kills the peoples due to the cause of water-born disease [5]. The effect of human activities on water quality can be evaluated in terms of the ecosystem and water usage [3].

Wastewater presents problems because of their variable composition and possible high concentration of suspended solids and dissolved solids. Physical, chemical, and biological characteristics can define wastewater quality. Physical parameters include temperature, total suspended solids, total dissolved solids, total solids, electrical conductivity, pH, and turbidity [2]. Polluted water is the leading cause of several diseases (such as gastritis, diarrhea, dysentery, cholera, typhoid, centric fever, and are the most recent decade problem of water pollution pathogenic microorganisms that cause around 2.3 billion people to suffer from the borne disease [6]. worldwide web as a world health organization report. Polluted water does not only affect the lifestyles of the peoples or generation, but it also may affect the lifestyles of future generations due to it cause environmental and climate change. Water pollution generated diseases are spreading quickly in the monsoon season due to the west water industrial system as well as less awareness of people as do not take into consideration these issues seriously.

Furthermore, water pollution monitoring and controlling are determining both the health of human beings and the environment as some paper had stated that around two to four million people die due to water pollution problems and environmental pollution [6]. So, situation engaged condition to monitor and controlling water pollution in every center of industrial. Tradition ways of water pollution controlling have many problems, such as time-consuming, need high labor force, and high cost. To overcome those problems, transform water pollution monitoring to Sensors and machine learning-based is significant.

One of the attempts to make the application more useful and efficient in monitoring pollution of water is using the sensor to get data in real-time for analysis and transform it to the controller or window terminal, which read data in digital format. Many developing countries like Ethiopia are ongoing to a rapid phase of industrialization and urbanization. In the process of the ruler to urbanization as well as farming to industrialization, it comes with the issue of an environmental management problem. Among that water, pollution management is the core issue for citizen's life, and every direction of development trigger the environmental issue.

The progress of urbanization and industrialization lead to increases in municipal sewage and industrial wastewater. Due to the monitoring lag of water quality and low automation level, existing wastewater treatment plants (WWTPs) are difficult to be maintained in stable operation in the case that influent water qualities are frequently varying. The main reasons are the lack of on-line sensors in WWTP to monitor the water qualities, so it is challenging to realize closed-loop control effectively. Therefore, the on-line method or real-time water pollution monitoring is urgently needed.

In recent years, wastewater from different sources (malt factory), either naturally or through human activities like runoff, industrial waste, residential areas, agricultural farms are released to the environment, especially in a developing country. Similarly, around Assela town, Kulumsa village wastewater from the malt factory is released to the environment. This wastewater contained total dissolved solids and total suspended solids and turbid when it released from the factory. Assela Malt Factory was the first malt factory in our country.

In this factory, the annual consumption of water is 700,000 m3 during the processing of Malt (IPS). This water passes through different processes for different purposes, such as for steeping, transportation, to cooling barley in a germination box when the temperature of the barley increases. At the end of all this, water gets contaminated and released as wastewater [7]. The existing techniques of application work discussed as blew. As manual testing, lab testing consumes more time, cost, human labor, energy, and the parameter can be used to avoid water pollution, and smart cities that used for smart urbanization and reduced problems face through water pollution to the environment. The main aims of this research study are to develop a framework for water pollution monitoring with Sensors and machine learning techniques concerning real-time monitoring and dynamic controlling at the industrial center.

1.2. Motivation of the Study

Water pollution is a major worldwide problem that needs continuous evaluation and modification of water resource center guiding principles at the international level down to individual levels of well. It has been serving water pollution, the leading cause of death and disease worldwide. Some research shows that more than 14,000 people die [8] of water pollution-related illnesses every day. In many developing countries, dirt or contaminated water is being used for many activities and drinking without any proper former treatment. One of the issues for this occurrence is that unawareness of the public, administration, scarcity of water, and lack of water quality monitoring system, which creates a whole series issue. Also, natural phenomena, such as algae tints, volcanoes rainstorms, and earth quack, change the quality and ecological station of water. Currently, the condition with water pollution includes several challenges for the international community.

Especially for developing countries, those challenges are increasing water pollution-related disasters, climate change, desertification, and drought, along with the lack of capacity to ensure integrated management and monitoring water resources for reuse in irrigation and to further exacerbates this situation [8]. So, water pollution poses a serious issue to the environment and need a series of the threat to human living standards, but the source of pollution may vary from place to place. For instance, polluted and non-pointed sources of water pollution cause a severe problem in environmental water pollution [9]. As world water quality control and statistics agency state that different source of the west is raised from glob to make water pollution impacts in 2010 [10]. Among those global water pollution, human west, human health impact (waterborne disease), and the industrial west. Among that global water, pollutions take a high rate of impact. As WHO 2010 states that due to global water pollution, every day, 2 million tone of sewerage, industrial, and agricultural west water discharged into the world water, form 6.8 billion peoples, and cause severe problems for the environment and life [10]. As water is a most important issue for all living organisms, it is essential to monitor and predicts its, and water quality monitoring is one of the first in the rational development, reduce the scarcity of water and management of water resource.

Assela malty factory has a comprehensive approach to viewing water governance. The company has a good link with agricultural sectors for the purpose of managing water scarcity problems. The good approach the company to reduce water scarcity is that have own artificial lake for water storage. The company has around 3 to 5 artificial lakes around the industrial company for water storage. Somehow, they try to understand the scarcity of water problems in society around the company. But the limitation of the company is that, to monitoring the west water released from the industry, they use the manual system. Such as lab test, and much human labor required. And the system elaborates on high costs and much time for water quality treatments. The other core problems with the company are that they have only secondary tertiary treatment of wastewater and no opportunity to reuse wastewater for the industry due to requiring high costs for tertiary wastewater prediction and monitoring. The other important opportunity from this company is many of the stakeholders are participate in work from the society around the company.

However, the main issue that initiates me to do this research study is the lack of the automated system for wastewater monitoring with low costs, a factor of wastewater released from company to the society and the tertiary monitoring wastewater with low cost and efficient time is the core issue to do these researches. As all the study focus on how to integrate IoT device (like Sensors, Arduino Uno, window terminal and soft sensors) and machine learning techniques for wastewater prediction and monitoring systems with low costs and in short period of times

In this paper study, we design a framework of water pollution monitoring and sensing based on trained, tested and evaluated a sample of data set. To keep the water resource within a standard described for irrigation usage and to be able to take necessary action to restore healthy of degraded water body using different machine learning approach and sensors, this system can collect varies parameter form such as temperature, conductivity, turbidity, PH, dissolve oxygen and many other supported limit adding to the significant or fundamental needed parameter to test and predict water pollution.

1.3. Statement of the Problem

Water pollution problems are available mostly in urban and industrial districts that cause a health risk. Because manually ways of water monitoring systems are faced with many problems, such as no real-time water pollution monitoring, need more chemical for polluted water treatment and human labor for monitoring system, no dynamic control of water pollution monitoring system, no detail concepts on data quality, and validation issue, no automated system with low cost, those problems or gaps that initiated to study this research: - Real-time water pollution monitor and supervision for water resources at industrial center, real-time water pollution monitor and evaluation systems supporting water quality evaluation and analysis on multiple data or extensive data, and dynamic water quality monitor and analysis at industrial center and for smart cities [4]**.**

This work believes that by incorporating the above list of problems and giving solutions to them in the way of integrating real-time water pollution monitoring and dynamic water quality monitoring and analysis with the method of machine learning approach and Sensors technology. Large scale urbanization production accompanying intensive chemical mixed or pollutant water from the industrial center that causes surface and groundwater contamination. Until recently, commutant and industrial wastewater represent the highest water pollution load, and this trend becomes a severe problem to urban transformation or smart cities.

Basically, around five papers which are more related for this research study state the gaps of wastewater prediction and monitoring system using IoT device and machine learning technique, such as some models can be found to excel which complex, nonlinear, and uncertain for wastewater monitoring system, limitation on integration of application platform with software side for polluted water prediction and monitoring system, directing calculation to stretch out the pollution monitoring system to a broad region and some lies on the integration of sensors and other device.

Based on the above-listed problem, the study incorporating to minimize or resolve waterborne disease, cost, climate change (environmental pollutions and waters scarcity), and human labor force by modernizing or automating water pollution monitoring and sensing using machine learning technique and Sensors. Because Sensors and machine learning technique-based water pollution monitoring, controlling and sensing gives real-time monitoring and real-time information to the end-user.

1.4. Research Questions

The most existing approach for water pollution monitoring is manual and thus cannot measure water pollution with cost and time-effective manner. Only some technique is dynamic with a high cost in some limited areas. Therefore, the research intended to get an answer to the following research questions.

***RQ1**:* How Sensors and machine learning algorithms integrated to monitor real-time water pollution monitoring?
RQ2: What are the basic mechanisms and core IoT devices used to develop a useful water pollution monitoring system framework?
RQ3: What are the most and effective machine learning techniques used?
RQ4: What is the performance of the proposed system and how it can be improved?

1.5. Significance of the Study

There are many reasons or importance to study, these core areas for living things as well as life as all. Because water pollution is one of the five environmental pollution issues. Mostly water pollution has a direct link with the human being and living thing life. However, the pollution of water is the core area that needs to study for living things. On another hand, water pollutions are the core and source of other environmental pollution issues (like air, land, and soil pollution). So, it is significant to study water pollution monitoring and prediction system. For example, the following list of the issue are the core problems solved when water pollution monitored and controlled in measurable situation west water monitoring and prediction. The most relevant points used to study water pollution monitoring are: -

- ❖ To kept water quality and water resources with monitoring and treating water pollution for reuse in irrigation.
- ❖ To design a framework used in a real-time method that reduced time and human power for water pollution monitoring and prediction as well as suggest them to provide guideline policy for west water treatment.
- ❖ To enhance the mechanisms used to reduce waterborne disease and the cost required for west water treatment or water pollution monitoring.
- ❖ To reduce the human labor force needed to monitor water pollution and enhance the performance of water pollution prediction.
- ❖ Finally, to enhance Sensors and machine learning technique based water pollution monitoring, to gain measurable monitoring and prediction as well as measurable information for end-users [11].

1.6. Objective of the Study

1.6.1. General Objective

The objective of the research study is to design a framework for water pollution monitoring system using sensors and machine learning Techniques.

1.6.2. Specific Objectives

To achieve the above general objectives and to answer the research questions, the following specific objectives will be considered:

- ❖ Conduct a literature review on the water pollution monitoring and prediction framework.
- ❖ Identify the mechanisms used in identification for water polluted areas as well as their use and factors.
- ❖ Determine the required devices, tools and approaches for data collection.

- ❖ Collect dataset and labeling dataset with domain experts.
- ❖ Select the related features used for polluted water identification.
- ❖ Select the appropriate machine learning techniques to apply for polluted water classification and prediction.
- ❖ Implement the required and significant machine learning model with exploratory data analysis.
- ❖ Determine the best analysis technique for polluted water prediction.
- ❖ Build a framework that saves time, cost, and human labor for access control solutions based on Sensors, and machine learning techniques.
- ❖ Evaluate the model and framework performance of water polluted monitoring.

1.7. Limitations and Scope

1.7.1. Scope of the Study

As some paper state that, they are many problems concerning real-time water pollution monitoring and sensing on some parameters. Also, the method of monitoring and sensing is the most malty factor. For instance, effects on performance, reliability, accuracy of the framework for water pollution monitoring and sometimes, even compatibility of the IoT device. So, based on the above-listed factors, the study specifies studying the ways of machine learning techniques (with proper performance machine learning technique) integration to Sensors based approach. Even if a critical challenge between the integration of machine learning techniques and most of the researchers focus only on water quality monitoring rather than west water reuse consideration.

By considering all these gaps, the proposed solution focused on the technique of machine learning algorithms and Sensors based approach integration for developing a framework used to measure water pollution sensing and monitoring respect to several considerations, such as performance, reliability, and accuracy. Generally, the area study of bounded with developing a framework for malty factor water pollution monitoring using a machine learning technique and Sensors based approach for the Assela malty factory.

1.7.2. Limitations of the Study

The research study has some limitations because of the following issues: among that limited on-line monitoring system, lack of sufficient open datasets for model developments, lack data storage area in the industrial company, availability of resources used to study, such as wifi support Arduino Uno board, turbidity sensor, DO sensor are the core materials and doesn't apply unsupervised and deep learning algorithms model is the main limitation of the study. The challenging activities at the scope of this research study are device require and insufficient domain expert on the data labeling and the more challenged or uncovered activities in this research study put as future work.

1.8. Beneficiary of the Results

The primary beneficiaries of this study are any users that use the water and environment for the day to day activities. On the other hand, any malt factor benefits by this work results as input for enhancing better water pollution prediction or monitoring framework for the Assela malt factory on their areas because they are trying to develop best ways water pollution monitoring in the content of modernization system and it helps to protect from more challenged polluted during the time to use in irrigation area for Kulumsa agricultural centers. Additionally, the researcher can replicate the proposed research for any other malt factory center used in Ethiopia and can serve as a baseline for related work-issued or use the real data to improve the research. Finally, the researcher also benefits from the study as input.

1.9. Organization of the Thesis

The structure of this organized thesis is as:

- Chapter one describes the introduction of the research study, which includes the overall research area, motivation of the study, statement of the problem, objectives of the study, significance of the study, and followed procedures to achieve the goals.
- Chapter two describes the literature review and related works of this research study.
- Chapter three describes the research methodology used to achieve the goal of the research.
- Chapter four describes the proposed solution. In this part, we design the framework and a machine learning model. How the solution works also explained in detail.
- In Chapter Five, the experimental tools and implementation were described. In this chapter, we try to see in detail how to set the simulation and what kind of hardware and software were used to implement the proposed solution and how to implement the proposed solution in the water pollution monitoring environment.

- ❖ In Chapter six, results and evaluation were discussed. In this part, we analyzed the result obtained from the simulation performed and evaluating our algorithm with the existing approach.
- ❖ Chapter seven describes the conclusion, recommendation, and further challenges or future work to study in this specified area.

CHAPTER TWO

2. Literature Review and Related Work

A literature review is reading or finding the most related research articles or journals on a framework of water pollution monitoring systems using Sensors and Machine learning techniques for more understanding, analyze, evaluate, compare the existing framework and gain further study to the research area. The primary objective of this literature review study is to summarize and read the concept, information, and workflow or framework of water pollution monitoring systems.

Water quality monitoring becomes more interesting among all research areas in this twenty-first century. Numerous works are ongoing in this topic focusing on various aspects of west water treatment or real-time water pollution monitoring systems. The central theme of the study was to develop an efficient, cost-effective, real-time water quality monitoring or real-time west water monitoring system which could be integrated machine learning techniques and sensors. A brief digest of previous works in this field of study given below. Wastewater means, domestic sewage and industrial waste discharged into drainage ditches and sewers and the entire contents emptied into the nearest watercourse [12].

2.1. Detail Supportive Concepts of the Study

Water quality degradation or water pollution is one of the most critical problems currently faced in many developing countries. With the increasing trend of urbanization and industrialization, the quality of water is continuously declining or become polluted. For this purpose, we propose a Sensors based and machine learning techniques for monitoring water quality system or water pollution capable of measuring the quality of water in near real-time. The proposed water measurement based on the World Health Organization and the Food and Agricultural Organization defined water quality metrics, or variables used to water pollution identification. Machine learning algorithms applied for classification of water quality, and the experimental results indicate that **deep neural network** outperforms all other algorithms with an accuracy of **93%** [11]**.** Insufficient water pollution monitoring systems in the place of industrial intended to cause various diseases, which increases the mortality rate [13].

2.2. Physical Wastewater Property

The main physical characteristics which water quality and wastewater examined are turbidity, Temperature, Electrical conductivity, pH and total solids, total suspended solids, and total Dissolved solids.

Water Quality: -to measure the quality of wastewater for irrigation, we based on three core water pollution identification characteristics, such as physical, chemical, and biologicals. Furthermore, it can be evaluated by specific parameters. The sensors are one of the most IoT based device technology today used to measure real-time water quality and west water monitoring indicators. The most common parameter used to indicate water quality or whether water is contaminated or not is the following with in-depth discussion.

Water Temperature [^{0}c]: - waterproof temperature is one of the physical parameters heavily influenced by air temperature and the depth of the water; the measurement was performed. Temperature affects biological life in the water as well as the number of chemical reactions, solubility, electrical, and toxicity of the water. Higher the temperature, low the dissolved oxygen in the water, and cause many effects on the reuse of wastewater and living things in the water because the temperature of wastewater is warmer than the average temperature of water supply. As the water temperature increases, the disinfectant demand and by-product formation, nitrification, microbial activity, algal growth, taste, and odor episodes, copper solubility increases, dissolved oxygen solubility decreases [2].

PH (Potential of Hydrogen)*:* - It is one fundamental water parameter and in natural groundwater has a typical pH of 7, which is known as neutral; however, many other components can influence pHs of the water, such as soil compound, organic matter, chemical contaminations, and precipitation. A regular pH reading for surface water ranges between 6.5 to 8.0 and not influenced. Acid rock drainage or acid soils can lower the pH of the water to 2, while an algal bloom can increase the pH value to 9.5. These values are extremely low or high and way of the normal pH of 7, profoundly affecting biological life in the water.

Typically, pure water (neutral) has pH equals 7. Rainwater is usually slightly acidic (pH 6.5), while seawater is usually slightly underlying (pH 8.5). Low pH (0–5 units) can be caused by acidic industrial effluents and fallout from the burning of high sulfur fuels. High pH (9–14 units) can be caused by caustic industrial effluents and return irrigation flows from saline agricultural areas [7].

The pH value of the wastewater is usually above 7.0. According to FAO [14], the normal range of pH is 6 – 9, and it is suitable for irrigation. PH values outside these averages are a gives a warning that water is abnormal in quality.

The PH values of wastewater standard for irrigation or reuse are different from country to country. According to the suggestion of the following countries, such as the US EPA, Portugal, Italy, and South Korea the PH value standard for reuse for agriculture is 6.0 – 9.0, 6.5 – 8.4, 6.0 – 9.5, and 6.0 – 9.0 respectively. The low PH values in west water affect the mobility of heavy metals in soil and crops, due to crops and plants observe waters in contaminated water bodies. Generally, according to the above-listed countries more of the existing agriculture affects Wes waters [15].

Conductivity [µS/cm]: - This parameter influenced by soil compound and water temperature - the warmer the water, the higher is the conductivity. Water alone is not conducive. Any change in conductivity allows an understanding of the presence of a compound solved in the water. For example, industrial wastewaters have a conductivity of 10000µS/cm. A reading close to this value would indicate the presence of industrial wastewater in the industry. Electrical conductivity is a measure of how wastewater conducts electric current. The electrical conductivity of wastewater related to the total concentration of ions in the wastewater. According to FAO [14], the guideline for wastewater quality used for irrigation purposes has electrical conductivity less than 3.0 dS/m or 3000 **µS/cm.** Generally, if the EC of irrigation water is below 700µS/cm, plant growth does not affect while the values of electrical conductivity above 3000 µS/cm can cause many problems on plants and peoples directly or indirectly (Ayers and Westcot,[14]).

Oxidation-Reduction Potential [V]: the water potential oxidizing or reducing potential measures. Any issue of change in this value can indicate water contamination, especially by industrial wastewater.

Dissolved Oxygen (DO) [mg/l]: - is measured in percentage and indicates the amount of oxygen in the water relative to the total amount it could potentially hold at that temperature. It heavily depends on the temperature (colder water can retain more dissolved oxygen than warm water), as well as the altitude and speed of the water movement. Its change and low value can also serve as an indicator of organic pollution and consequent reduction of organic species in the water. However, if the water becomes stagnated and excessive growth of aquatic plants occurs, a high DO value, indicating supersaturation (more than 100%), can happen. DO may lead to diseases in the fish population, for example.

Turbidity [NT U units]: - Measures the suspended and dissolved matter in the water as well as the water clarity. It changes with ground soil and precipitation; however, a high turbidity level can indicate a high concentration of metals in the water. Turbidity is the outstanding feature used to identify the existence of suspended particles in wastewater and at which the water quality loses its transparences [16]. Turbidity is used to measure water pollution clarity. Polluted water with high turbidity value is called polluted water, while low turbidity value is suitable for reuse in irrigation. The occurrence of high turbidity value in wastewater is due to cause by sediments suspended in water bodies that results from the scattering as well as absorption of light rays, due to the high turbidity values in the water decrease the amount of sunlight pass through water, it causes the problems of photosynthetic rate on plants and decrease the level of dissolved oxygen [16].

As historically industrial west water monitoring or water quality optimization indication is focused on the above-listed parameters. The occurrence of water with a turbidity of less than 5NTU (Nephelometric Turbidity Units) is acceptable [17].

Except for South Korea, many countries, such as US EPA, Spain, and Greece, used or recommend different standards of turbidity values for irrigation as well restricted individually. For instance, the US EPA and Greece restricted turbidity values in Wastewater to reuse for irrigation, when turbidity value is equal to or lesson than 2NTU. However, Spain suggests that the turbidity values lesson than 10NTU in irrigation for vegetables [18]. Also, the ministry of environment protection suggests that the turbidity value for food crops and possessed food crops uses 2NTU and 5NTU respectively. Most countries do not have their turbidity value standard for west water reuse for irrigation, but they used suspended solid instead. Generally, Spain and Greece are the only two countries that have their standard turbidity values for irrigation.

Total Suspended Solid [mg/l]: - All contaminants of water, other than dissolved gases, contribute to the solids load. Total suspended solids are a measure of the suspended solids in wastewater, effluent, or water bodies, and it is the dry weight of particles trapped by the filter. It is water quality parameters used, for example, to assess the quality of wastewater in a treatment plant. All the suspended particles found in west water is called suspended solids. Mostly, suspended solid exists in sanitary and industrial west waters [19].

As the level of TSS increases, the water body starts to reduce the capacity to support different diversity of aquatic life. The problem of solid particles is that they observe heat from sunlight that increases water temperature and decreases the amount of dissolved oxygen in the water that results in more cooling of waters. To say water is suitable, moderate, and polluted based on TSS, the values are lesson than 20mg/l, between 40 – 80 mg/l, and greater than 150mg.l respectively [2].

Total dissolved solids [mg/l]: - Total dissolved solids (TDS) include inorganic salts such as calcium, magnesium, potassium, sodium, bicarbonate, chlorides, sulfates and a small amount of organic matter dissolved in water. The sources of Total dissolved solids are natural sources, sewage, urban runoff, and industrial wastewater. According to FAO [14], the guideline for the wastewater quality for irrigation is suitable when the value is higher than 450 mg/l and less than 2000mg/l, which considered severe value.

2.2.1. Chemical Parameter Used to Measure Water Pollution Monitoring

They are many variable or parameters used to identify when water contaminated with the different chemical at the industrial center, such as Nitrate (total N), phosphate (total P), faucal coliform, PH, dissolved oxygen, potential hydrogen, biochemical oxygen demand, and chemical oxygen demand.

2.2.2. Physical Parameter Used to Measure Water Pollution Monitoring

As internally chemical parameters used to identify whether water is polluted or suitable for reuse, physical parameters or variables used to indicate water pollution based on different weather conditions and situations, such as suspended solids, temperature, conductivity, turbidity, et.c [13].

2.3. Traditional Wastewater Monitoring Process Diagram

A traditional Anaerobic/Anoxic/Oxic (A/A/O) process is applied to monitoring and predicting unwanted elements in WWTP, which exhibits excellent performance for nutrient removal. However, the performance parameter of any WWTP must be modified according to the actual condition of the A/A/O process.

This study uses a combination of the KPCA and ELM to extract the principal components from past observation data for forecasting inlet COD and BOD concentration, which helps to adjust the performance parameters, such as the balance of carbon source, aeration rates, and reflux ratio.

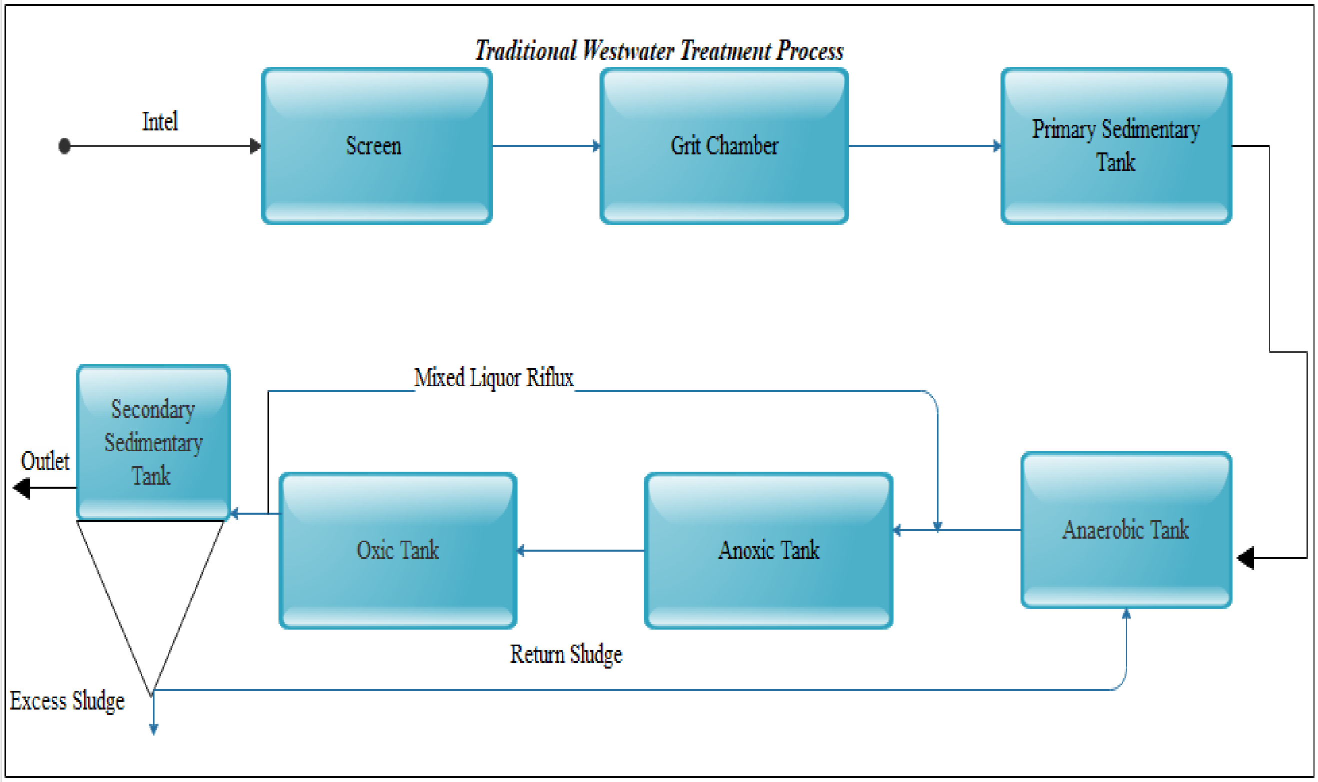

Figure 2.1: Anaerobic/ Anoxic/Oxic Process Flow Diagram [10].

2.4. The Overview of the Existing Wastewater Monitoring systems

Due to the lack of integration between Sensors and machine learning techniques to measure real-time effluent quality parameters, it is challenging to implement monitoring, control, and operational optimization in the wastewater treatment plants. On the other hand, existing on-line hardware sensors is not sufficient in accuracy and reliability due to significant investments, poor reliability, and difficult maintenance of devices. To overcome this problem the paper studied contributes data-driven software that increase popularity with the availability of recorded and stored data, and apply the most machine learning techniques, such as Random Forest multivariate statistical method, artificial neural network, and support machine learning (SVM) method, are successfully used to monitoring and model data-driven software sensor. It is traditional ways of modeling which are based on a single model and faced with many problems, such as complex process with strong nonlinearity's, time-variants, and highly uncertainty measurement [13].

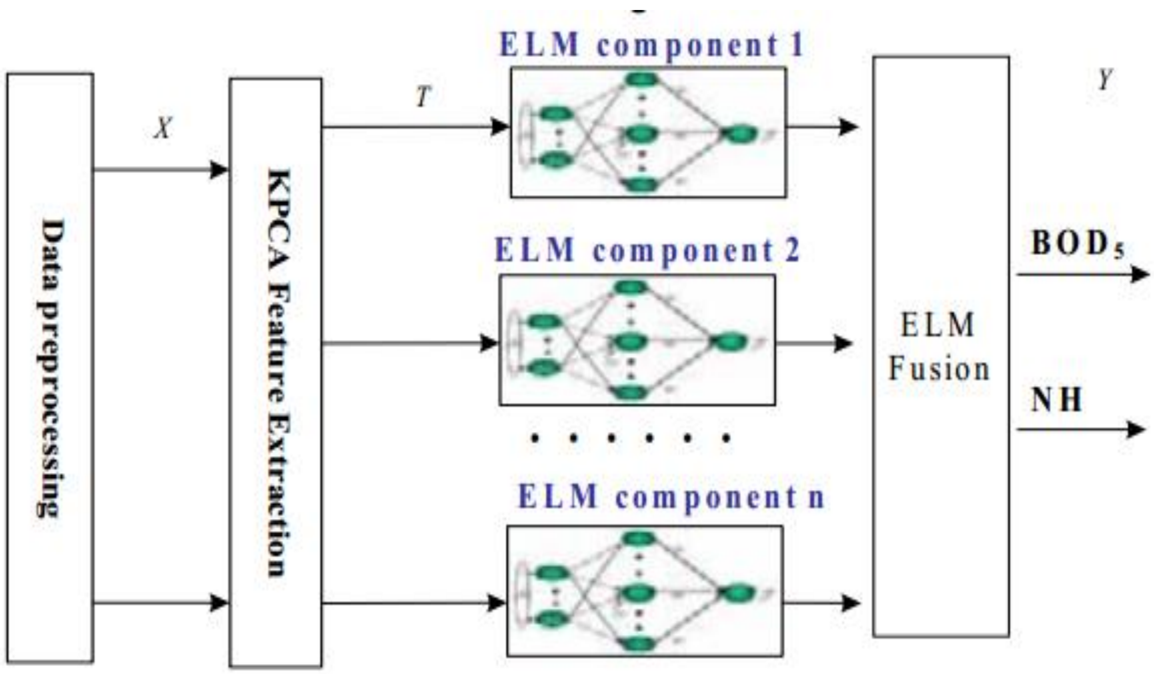

Figure 2.2: Strategy of the Ensemble Soft Sensing Based on KPCA and ELM [13].

The structure of the soft ensemble sensing based on KPCA and ELM on the above figure consists of the following parts, such as data preprocessing, KPCA feature extraction, ELM component, and ELM fusion [20].

2.4.1. Data Preprocessing

The prediction capacity of the data-driven classification and prediction model strongly depends on the standard quality of the training data set. So, gross errors and missing data contained in the history data set are first processed based on a robust EMPCA method [21].

2.4.2. KPCA Feature Extraction

KPCA is used to eliminate the collinearity and noise in the data set. The principal component analysis is one of the machine learning techniques used to take an extensive list of interconnected variables and choose the ones that best suit for a model building. This process of focusing on only a few variables is called ***dimensionality reduction*** and helps to reduce the complexity of our dataset.

2.4.3. ELM Component and Fusion

Its algorithm was inserted into the ensemble frame as a component model since ELM runs much faster and provides better generalization performance than another popular learning algorithm tool, which used to overcome variations in different trials of simulations for a single ELM model [21].

[Biswajit Paul] (2018) Traditional methodology of water pollution monitoring or west water treatment requires collecting data sets from various sources manually. Afterward, samples of data set have been sent to a laboratory for testing and analyzing. Sensors based technique used to save time consumption and decrease manual effort; the study testing equipment would be placed in any water pollution source. As a result, this model can detect pollution remotely and take necessary actions. The main goal of this paper to build a Sensor-based Water Quality Monitoring System [4][22].

Yazhini and Maruthi (2017) proposed model showed us how the internet of things platform could be used for water management. Remote Sensing techniques and the Internet of things can be applicable for a broader spectrum of research domains for monitoring, collecting, and analyzing data. The proposed a water quality monitoring system based on the internet of things. Water properties can be physical (temperature and turbidity), chemical (PH and dissolved oxygen), and biological (algae and phytoplankton). In the proposed system, the physical and chemical properties of water in different water sources such as drinking water, swimming pool water, water bodies inside Dhaka city, and industrial wastewater ware investigated.

2.4.4. The General Architecture for Sensor Base Water Pollution Monitoring System

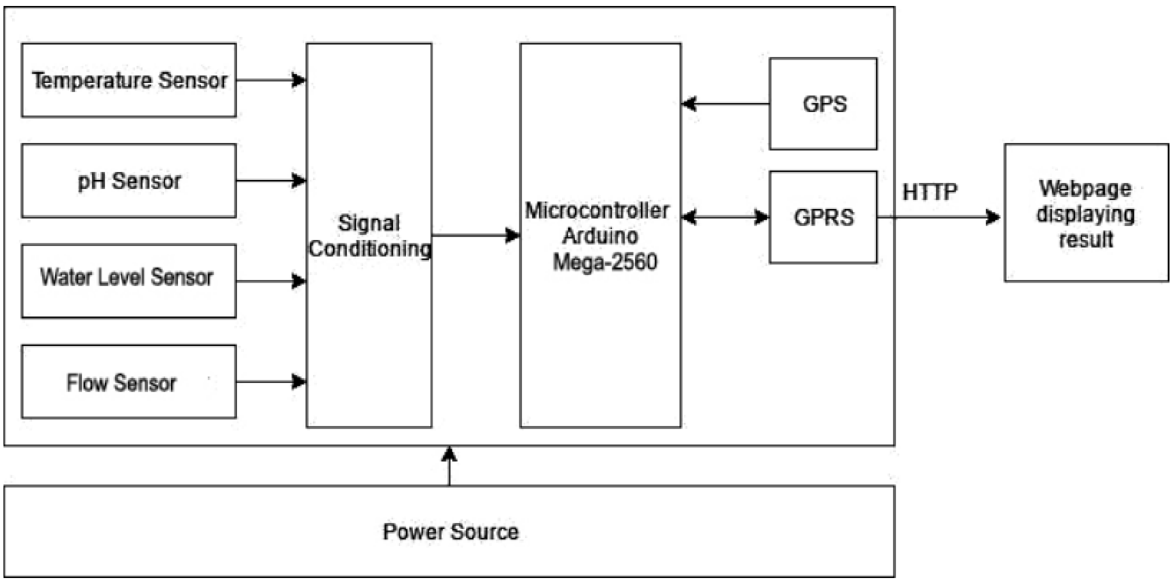

Figure 2.3: General Architecture for Sensors Based Water Quality Monitoring System [13].

A microcontroller used as a sensor node that stores real-time data and sends the data to the cloud server storage via the Wi-Fi module. ***GSM shield*** connected to the microcontroller, which sends the text notification to the mobile device. ***Http protocol*** transfer the data to the cloud and a webpage running on a localhost server displays the result.

2.5. Condition of the Water Quality and Wastewater for reuse

During the use of wastewater for irrigation or reuse, the effect or factor of Wastewater on the environment or human health must be considered. Additionally, when using wastewater for agricultural centers, they must be based on world health organizations and food and agricultural organization standards to reuse wastewater for irrigations. [18].

2.6. Water Management and Controlling

Water management and control can be seen in a different angle of views and address of different direction of function, among those angles of views, the following approach is listed:

- The imperative societal needed to make the best use of a scarce a good
- The geographical characteristic that requires tail made approaches

The technology trends to improve water quality, ensure water access, enhance better water management and control, and ensure fair pricing [23].

2.6.1. Significance of IoT in Smart Water Management

The use of IoT and sensors for water management and distribution combining with optimizing technique, as well as predictive analytics, can bring efficiency and cost reduction for water management operation gives the following improvement water management solution, such as water leakage detection, efficient, systematic water management, water quality, and safety monitoring, quality control on water reserves, transparence, and prescriptive maintenance on research infrastructure [23].

Currently, the integration of popular algorithms with the IoT sensing domain plays a significant role in polluted water monitoring sensing by predicting mechanism. One of the best algorithms applicable in smart water pollution monitoring and controlling is machine learning classification algorithms, such as random forest, Decision Tree, KNN, support vector machine, and linear regression were the most supervised classification algorithms. Most of the developed countries' governments, such as the US and EPA governments, are based world health organization guideline standard to set a polluted water standard for reuse in agricultural centers.

2.7. Related Work of the Study

Many types of research have been done in the area of water pollution monitoring or west water treatment in the last years. However, some papers covered some concepts about traditional west water treatment. Some papers also covered some concepts of sensor-based real-time west water treatment at the different river and lake centers. The gaps of those theses, their limitation, the methodology used, and the result obtained in previous work were revealed in the blow paragraphs. In general, a transparent understanding methodology, workflow, and architecture or a framework of west water treatment, and its characteristics, mostly IoT based, most of the researcher discussed more related work to polluted water treatment [4].

In the smart water quality monitoring system, when the sensor board is exchanged on, the sensors are playing a role in distinguishing the individual water parameter information.

The information on the water level, pH, turbidity, carbon dioxide, and temperature were shown on the water quality monitoring dashboard on the PC utilizing Python codes.

The consequences of water parameters were shown on the water quality monitoring dashboard. The perusing of water level is changed when the separation between the water surface and the water level sensor is changed. The waterproof temperature readings change, as indicated by the expanding and diminishing of the water temperature by utilizing warm water and cold water.

The scope of the esteem was shown for the observation of pH, temperature, turbidity, carbon dioxide, and level of water. In this paper, an adequate, low power, ongoing water quality observing system based on remote sensor systems is displayed. As the sensor nodes are utilizing sun powered boards and because of various methods of a sensor hub, the fundamental power issue is understood in this work. Future works incorporate the utilizing of more productive directing calculations to stretch out the system to the vast region and lie on the Integration of turbidity sensor, dissolved oxygen sensor, and color sensor to the sensors utilized as a part of this work [24].

The aims of this research paper are an aquatic form for water quality monitoring to develop the abilities to execute system to provide the information to different sensing parameters for an extended area. The gaps of this study are to develop the android application and improvement of their futures on the software side. In this paper studies or research study, the hardware is developed for water surface drone in the water quality monitoring [23].

In this study, they were two hypotheses tested the first stage of hypotheses on the quantity of recyclable material generation in Kisi town. The test material does not constitute constituent; an essential part of the town's waste was rejected and the optional acceptance. However, it was searching that the equipment is not generated in equal quantities in all the classified areas of the cities. The quantity of recyclable equipment generated is in proportion to the size and population density of the respective classification areas.

By its large area and high population density, the residential area generated the most waste. Most of this place's recyclable waste is scattered; these it making it hard to estimate its total quantity. The industrial, institutional, and commercial places are relatively small in size, accounting for their limited generation of recyclable wastewater. Except in the commercial place, the district does not play an essential role in solids wastewater monitoring of the urban area. This time problem is unrecognized taken to the district collects of user's service charges; district SWM service in the urban is improved. There is a reason for the state of the urban environment has been decreased rapidly in recent years.

In response to this time problem, 48% of the urban residents felt the district mandate of managing solid waste ought to be transferred to other agencies like private companies. 52% of the resident's ware in favor of the retention of the district as the leading player in solid west management. Many thought those private companies would be more accountable, reliable, and responsible, thereby leading to better services. The main area for ongoing research is water, air, and noise pollution in the public place. Water pollution is particularly urgent, given the ongoing pollution needed to treated sewage from town [25].

The ways of study to develop a wastewater quality monitoring system that incorporates a via UV/Vis spectrometer and a turbid meter to monitor some parameters like - COD, TSS, and O&G concentrations of the effluents of the Chinese restaurant on campus and a pilot ECeEF wastewater treatment plant. The sensor fusion technique is more used to fuse the signals from these two sensors/instruments. An improved boosting method, Boosting-IPW-PLS, was developed here to handle the noises and information unbalance of the fused information to model and predict the water quality. The system was evaluated in field trials. Satisfactory results were obtained, as seen subsequently.

Why I ignore local paper is, in our country, most of the researchers considered water quality and pollution monitoring systems for environmental science, agricultural and biological science, chemical engineering, biochemistry, genetics, and molecular biology area of researchers. So, some water pollution monitoring system which is related to my work is an area of agricultural environment and health centers. But they were mostly focused on- lab test improvement of water pollution monitoring system and lack of IoT device technology integration with machine learning approach for water pollution monitoring system. So, under computing college, there are no researchers exercise on water pollution monitoring systems using sensors and machine learning approaches. However, the following papers did on water pollution monitoring systems, wastewater treatment under agricultural environment and health center of research areas, those full lack of IoT technology, support of IoT device and machine learning integrations.

As I review the papers focus on rivers water pollution status and policy on the Nile, Omo-gibe, Tekeze, and awash river basin in Ethiopia. They state river water pollution is a growing challenge and needs a day to day activities to apply intersectoral collaboration on managing water resources. Generally, they found finding that enhance the awareness and participation of stakeholders to improve the ecological quality of rivers [26]. So, they limited only on finding policy provisions and increase awareness for society. Also, there are knowledge gaps to consider IoT technology and machine learning approach to enhance mechanisms to control river water pollution management. In addition, there is another paper which focuses on drinking water quality test for Shambu Town to improve health risk. Generally, this paper focus to assess the quality of water for drinking [27]. This paper also cannot consider IoT technology rather than using Physico-chemical and Biological to improve water quality using a lab-test. Also, there is no more consideration to reuse wastewater to overcome the scarcity of water.

As all most of the paper on wastewater treatment, water pollution monitoring systems were have done by collage of healthy, agricultural and environmental science. But they focus on lab-test improvements that need high cost, time, and human labor. So, it is important to participate from computing collages to overcome the above-listed problems with the support of IoT device and machine learning techniques.

Table 2.1: Related Work Summary.

No	*Contents*	*Description*
[28]	**Title**	Prediction of Water Quality and Smart Water Quality Monitoring System in IoT Environment
	Author	[Karthick. T], [Gayatri Dutt], [Tarunjot Singh Kohli], and Snigdha Pandey
	Scope of the Study	The esteem shows for the observing of pH, Temperature, turbidity, carbon dioxide, and level of water.
	Gaps	Directing calculations to stretch out the system to broad region, and lies on Integration of turbidity sensor, dissolved oxygen sensor and color sensor to the sensors utilized as future work
	Year	2018
[29]	**Title, Author &Year of Publication**	Air, and Water Quality Monitoring Using IoT By Using Aquatic Surface Drone, Ch. Pavan Kumar, Praveen Kumar & 2018.

	Scope of the Study	In this study, the hardware part is developed for water surface drone in the water and air quality monitoring.
	Gaps	Limited to develop the android application and improvement their futures on the software side
[30]	**Title, Author & Year of Publication**	Urban Environmental Problems in The Case of Solid Waste Pollution and Management in Kisi Town: - [KEFAM.OTISO] & 2017
	Scope of the Study	Solid West Management
	Gaps	The main area for ongoing research is water, air, and noise pollution in public place. Water pollution is particularly urgent given the ongoing pollution needed to treated industrial sewage from the town
[31]	**Title, Author & Year of Publication**	Inlet Water Quality Forecasting of Wastewater Treatment Based on Kernel Principal Component Analysis and an Extreme Learning Machine: - Tingting Yu 1,2, Shuai Yang2, Yun Bai2, *, Xu Gao2,3 and Chuan Li 1,2 1 and **2018**
	Scope of the Study	They only focus on the combination of KPCA-ELM algorithms based on two parameters like COD chemical oxygen demand and BOD biochemical oxygen demand.
	Gaps	In some of the models or methods, the model can be found to excel in water quality forecasting of wastewater treatment in ways that are complex, nonlinear, and uncertain.
[26]	**Title, Author and Year of Publications**	River Water Pollution Status and Water Policy Scenario in Ethiopia: Raising Awareness for Better Implementation in Developing Countries Aymere Awoke1,2 • Abebe Beyene1 • Helmut Kloos3 • Peter L.M. Goethals4 • Ludwig Triest2 19 June 2016
	Scope of the Study and Gaps	Limited to water pollution status and investigating policy for raising awareness and lack of IoT device technology and machine learning approach for water pollution monitoring systems.
[27]	**Title, Author and Year of Publications**	Drinking-Water Quality Test of Shambu Town (Ethiopia) from Source to Household Taps Using Some Physico-chemical and Biological Parameters

		Belay Garoma1, Girmaye Kenasa2* and Mulisa Jida3, on 31 December 2018.
	Scope of the Study and Gaps	Limited to Water Quality Test lab improvement using Physico-chemical and Biological Parameters & lack of sensor and machine learning technique integrations.
[32]	**Title, Author and Year of Publications**	The Pollution Profile of Modjo River Due to Industrial Wastewater Discharge, in Modjo Town, Oromia, Ethiopia, Amde Eshete Gebre*, Hailu Fekadu Demissie, Solomon Tejineh Mengesha, and Mesfin Tafa Segni, 2016
	Scope of the Study and Gaps	Limited to pollution profile enhancement of Modjo river that causes due to wastewater from industrial by lab-test improvements and no exercise on water pollution monitoring using sensors with the integration of machine learning approach.
[33]	**Title, Author and Year of Publications**	Virological Quality of Urban Rivers and Hospitals Wastewaters in Addis Ababa, Ethiopia, Tesfaye L. Bedada1, *, Teshome B. Eshete2, May 22, 2019
	Scope of the Study and Gaps	Limited to Virological Quality of Urban Rivers and enhance the mechanisms to control wastewaters released from hospitals to city.

Therefore, based on the above gap listed under related work in the table above, machine learning techniques and Sensors based mechanisms using python language is the best choice for real-time water pollution monitoring and sensing. Mostly several papers considered to kept quality water monitoring rather than west water reuse for irrigation. Because most researcher considers minimizing and avoiding factors comes due to water pollution. However, the proposed solution is not only considered to avoid factors respect to water pollution (west water from malty factor), the study considered both to avoid or minimize problems that come due to water pollution and reuse of west water for irrigation to overcome water scarcity and to the safe water resource.

CHAPTER THREE

3. Research Methodology

Assela Malt Factory (ASM) was explicitly selected for this research study. Assela Malt Factory (ASM) was established in1984 and situated 75 km southeast of Adama. Malt is the primary input for beer production, and the company currently supplies around 40% of this growing Ethiopian market. However, besides this, water pollution is the current issue problem as overall, the globe among those essential elements faced with environmental pollution [34].

3.1. Overview

Under this chapter, the workflow of research methodology; studying or examining the current system in order to deeply understand the existing, system development, define the proposed methodology, and describe the system train and testing method, compare and contrast the available proposed system in order to select the appropriate tools and method for proposed system studied. Method Are the ways used to achieve the study in different ways of data collection, data analysis, system architecture or designing a framework, and Development tools (Hardware and Software) used for system development. Generally, to obtain the objective and answer the research questions found in chapter one, the following methods and techniques followed figure 3.1 depicted the research procedure of this study, we mention the following steps as a research methodology.

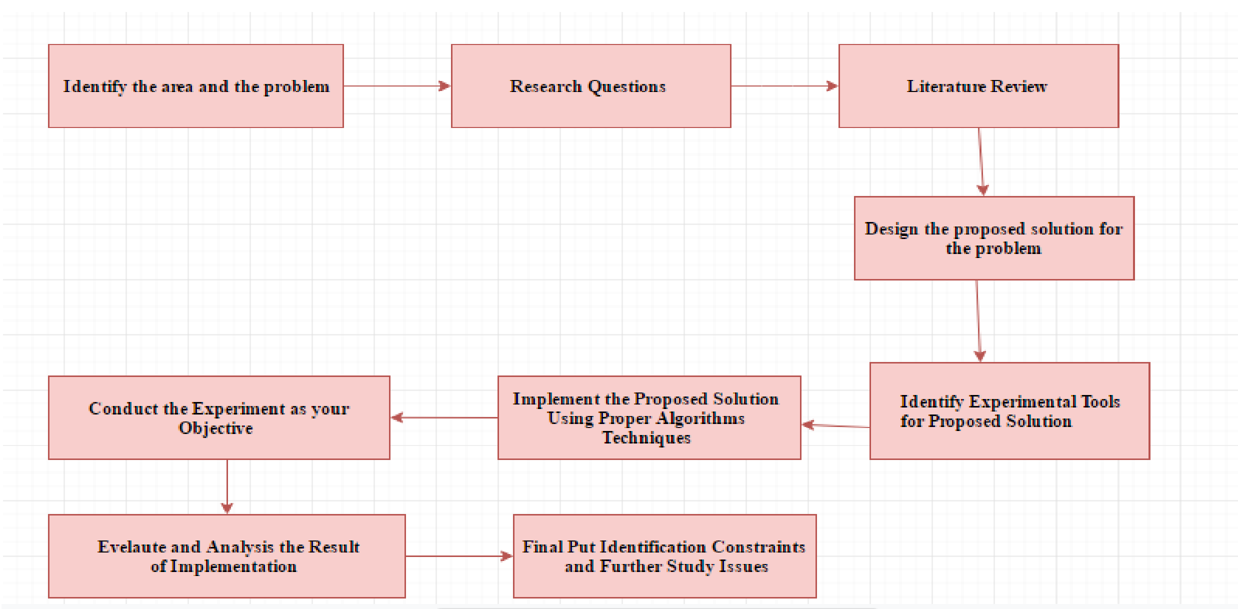

Figure 3.1: Research Study Procedure.

3.1.1. Description of the Study Area

The study conducted at Assela Malt Factory of the Arsi zone. The factory is located at 75 km northeast of Adama, the capital city of Oromia. Assela malt factory is located at an altitude of 2300 m above mean sea level with geographical coordinates of 3907'E of latitude and 7057'N longitudes. The average annual highest and lowest temperature of the Assela malt factory is 220^{c} and 100^{c} respectively. The mean annual rainfall varies between 800-1200mm. Humidity ranges from 43%-93%, the wind speed was three kilometers per hour from east to west. Soil type around Assela malt factory is composed of sandy loam, and clay soil (Assela Agricultural research center). By using rain; Farmers around Assela malt factory practice rainfed agriculture, and they mainly grow barley, bean, maize, wheat, potato, onion, and carrot.

(Source Assela Malt Factory and Assela Agricultural Research Center)

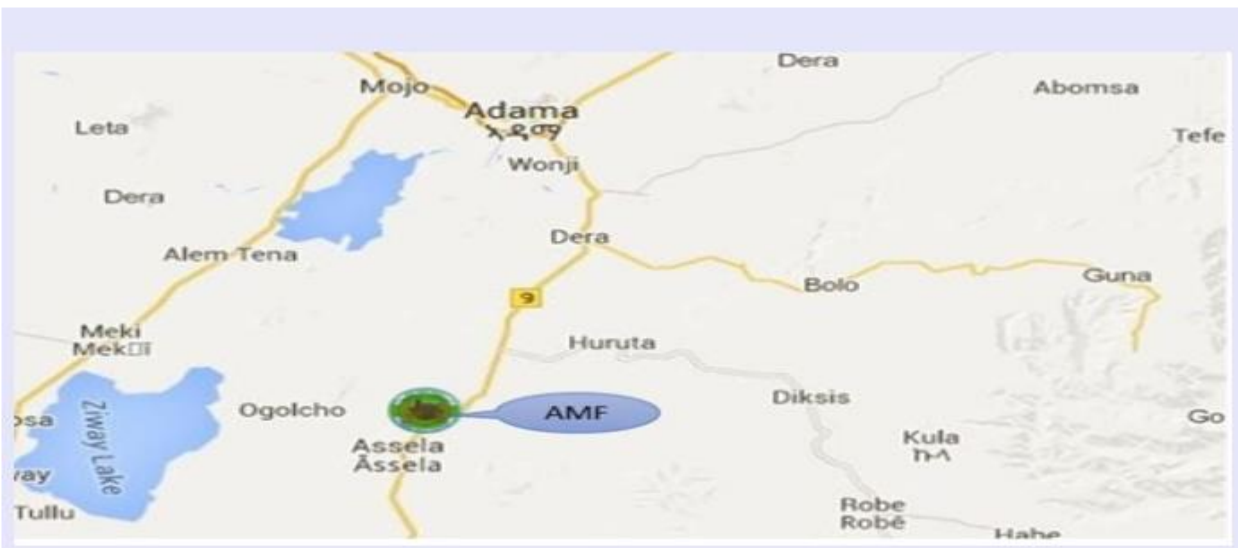

Figure 3.2: Location and Map of Assela Malt Factor Source: www.asmalt.gov.et

3.1.2. Description of the Study Site (Treatment tank, Steeping and Germination box)

Waste Released from Treatment Tank: **-**Before water goes to the steeping box, some chemicals are added to raw water to remove some waste from water. Such chemicals are Aluminum sulfate (AlSO4), Sodium carbonate (NaCO3), chlorine (Cl2), silicon oxide (SiO2), and charcoal. These all together remove some waste from the water. By the transportation of water, this waste is released from the factory.

***Wastewater from Steeping box:* -** The water from the treatment tank goes to steeping boxes for the steeping process. The steeping process is working for 480 minutes. During this time, the process is working for 15min and takes rest for 15 min to protect overflows. After 480 min of wastewater from the steeping process is released.

Germination box*:* - After the water transports, the steeping process barley from the steeping box into germination boxes. This barley stays in a germination box for 5- 6 days to make the germination process. Based on laboratory results, water is added to germination boxes when it is required to protect barley from dry. Within these days, the water is released from these boxes.

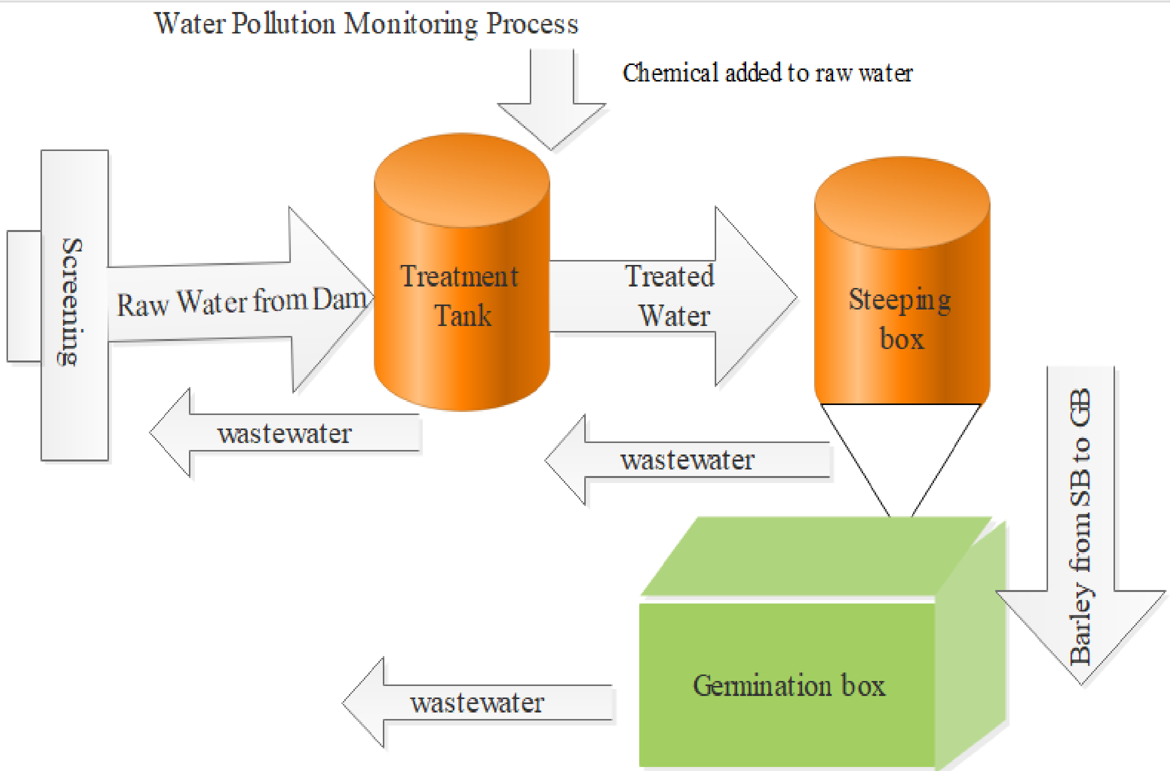

Figure 3.3: Assela Malt Factor Path Approach of Wastewater Treatment Process [2].

3.1.3. West Water Treatment or Monitoring Method

Currently, they are two basic ways of west water monitoring; those are the traditional Anaerobic/Anoxic/Oxic (A/A/O) process is applied to treat domestic sewage in WWTP, which exhibits excellent performance for nutrient removal, [9]. Moreover, Sensors based west water monitoring can classify water properties into physical (temperature and turbidity), chemical (PH and Dissolve Oxygen), and Biological (Algae and phytoplankton), which used to monitor and control west water within the real-time situation. So, in this proposed system, the physical and chemical properties of west water investigated in an industrial center. Mostly water quality monitoring techniques can be divided into three main parts: -

3.1.4. Physical Analysis Method

It is used for finding transparency, color, temperature, turbidity, and odor.

3.1.5. Chemical Analysis Methods

The chemical analysis method used for analyzing the following most parameters, such as total dissolved solids, total suspended substantial, biological oxygen demand, PH, total hardness, and electrical conductivity. Additionally, total solid particle, nitrate, phosphate, sulfate, and dissolved oxygen are also analyzed using the chemical analysis method.

3.1.6. Biological Monitoring Method

The biological monitoring technique includes all the qualitative analysis of planktons (zooplankton and phytoplankton).

As a whole, currently, water quality monitoring has become a problematic task became of global warming and an increasing population to concentrated location. The issue deal with it is essential to develop a better system that monitors water pollution or water quality prediction. The turbidity feature is used to measure water purity obtained by analyzing the number of particles in the water. If the turbidity is high, there is a chance for the highest presence of pollutants that may cause different bacterial diseases such as chloral and typhoid. The temperature of water increases with an increase in turbidity. The water quality is predicted based on Turbidity, PH, Temperature, and Dissolves oxygen [25].

Proposed System framework: the overview system of the prototype described in the figure below in section four (***Figure 4.1)*** with the following three main module categories. The proposed system framework used for west water monitoring and managing for reuse at plantation or agricultural center. According to the World Health Organization (WHO) studied and Assumption by 2025, half of the world's population may be living in the area of water scarcity or high-water crises as a result of water contamination and lack of plantation. So, for the developing country like Ethiopia, where people cannot afford high-cost water purifier, this system can provide an affordable solution to water crisis or Wastewater Treatment and Reduce Health problems Came Result Wastewater [4].

3.2. Dataset Description

Dataset is the core elements needed in our works. The area to get the dataset is challenging because of many reasons. As aggregate data set is got in two ways on this area, by taken the existing data that stored manually or electronically in the company by direct contact with the admin and the second way is using an electronic device or IoT based device like the sensor in real-time from sample polluted water in the area. However, both the method used to collect this data is complicated in our country, not only in our country, almost all in the developing countries. The core issues with this method are, one the security matter since there work is full of challenges and problems in the area, they are trying to secure themselves. A good example is a mojo thinning industry, which is not voluntary to give research area on it and finally stop there works by the influence of societies in our country.

Moreover, secondly, there is no fully accessible IoT device in our university for research works as a sensor, even it is not available in the market in our country. So, to overcome those challenges for my work, it read systematically and get some available from the website by reviewing different papers, journals, and interviews for the professionals at the water environmental department and chemistry departments. Then it prepares a blow dataset with the following parameters.

Table 3.1: Dataset Show with Common Parameter Needed.

No	***Parameter Name***	***Data type***
1.	Station Code	Object
2.	Location	Object
3.	State	Object
4.	Temperature	Float
5.	Dissolve Oxygen	Float
6.	PH (Potential of Hydrogen)	Float
7.	Conductivity	Float

8.	Turbidity	Float
9.	Total Dissolved Solids	Float
10.	Dependent Variable/Class	Object
11.	Year	Int

As a whole website and literature reviews are the primary resources of the dataset for this research work to design and develop a framework for west water pollution monitoring with relevant parameters. Important information is using the current information standard, and west water pollution monitoring is focused.

3.2.1. Data Collection

The study gathers water pollution data that is available on-line for sample data set and interview from domain experts of the west water treatment of Assela malty factor and Adama science and technology science university West water treatment domain experts for data labeling with the standards. Having this sample data set for model prediction, the real data used for testing parts that collected using ph. Sensor and waterproof sensor with Arduino Uno support. The aims of the study to have a real dataset for water pollution monitoring and prediction. They are several criteria for collecting dataset for model prediction. This study considers the following criteria for selecting ways of collecting datasets.

- PH and waterproof temperature sensor ware used to collect real data.
- The existing dataset ware used for model prediction and labeled with the support of domain experts.

Table 3.2: The Selected Parameters Categories.

No	***Categories Name of Parameters***	***Description***
1.	PH of Water	PH is the measurement of electrical charge particle substance used for the identification of acid and primary substance. So, based on the definition of PH, the PH of the water is the core parameter for identification weather; it is bases, acids, or regular.

		In other words, the PH of the water is to identify water is polluted or non-polluted.
2.	Waterproof Temperature	Temperature describes the degree of hotness or coldness in degree Celsius. So, water temperature needs to be monitored regularly as outside tolerable temperature range, disease, stress can be become more prevalent. As temperature consciously changes, the following issues become dissolved. Those issues are photosynthesis activity, dissolve rate or gases, and amount of oxygen in environments and water temperature. Generally, water temperature parameters used to distinguish the effects of polluted water on the solubility of dissolved oxygen and the rate of plant growth. ***The solubility of dissolved oxygen***: as more and more gas can be dissolved in cold water; the water became warm. ***Rate of plant growth:*** water temperature and rate of plant growth is the parallel relationship in the case of photosynthesis So, water temperatures used to identify a plant growth is normal or abnormal.[35]
3.	Dissolve Oxygen	DO describe the concentration of oxygen molecules in the water. In other words, it depends on water temperature and BOD directly or indirectly. So, do is used to identify the rate of gases in water with its advantage and disadvantage of plant growth.
4.	Turbidity	Turbidity describes a measure of coldness or clarity of the water that is polluted or not- polluted. Because of the coldness of water caused by different solid. Mostly soil particles (sand, silt, clay), macro plants, and animals that ware suspended in water column or area. So, the higher level of water turbidity causes several problems for stream systems because higher turbidity of water level is associated with a higher level of virus, parasites and some bacteria that attach themselves to dirty water and cause health

		problems while moderate turbidity level indicates healthy environments. Generally, the level of turbidity in water used to indicate whether water is polluted or suitable for loving things.

3.2.2. Dataset Preparation

After the data collected, a preparation process proceeded, which are follow collecting, cleaning, normalizing, filtering, and consolidation data into one file or data table (with excel more of time). This process used preparation tools, MS Excel, and other related tools. Handling missing values and normalizing the data used for water pollution prediction in dataset preparation while designing a training model for polluted water monitoring. The ***following tasks*** perform for preparing the dataset to be protected or prediction.

- ❖ We are handling missing values for water pollution parameters.
- ❖ Normalizing and standardize water pollution parameters by scaling the scalar standard.
- ❖ Join the data (parameters) with the target data.
- ❖ We are removing duplicate data in the stored dataset.
- ❖ Parameters were used for water pollution identification with the help of Machining learning. All the above preparation of the dataset considers for behavior and nature of polluted water in a malt factory.

3.2.3. Feature Selection and Extraction Method for ML Classification

In most of the real-world datasets, not all of the attributes contribute to the definition or determination of class labels. So, there is a challenge to selecting a subset of features within the minimum risk or significantly essential to determine the class/ pattern of classification. Because features were characterized with the following behaviors, ***Relevant*** which influence the class levels, and the role cannot be assumed by the rest of parameters, ***Irrelevant*** variable that has no any influence in class levels, ***Redundant*** variables that can take the role of others and makes confusion with the class of classification.

So, based on the above challenges, the feature needs to the feature selection technique that targets to find the best optimal subset consisting of $\boldsymbol{m}$ feature chosen from the total $\boldsymbol{n}$ feature [36]. Generally, the target of feature selections is removed subset from the garbage data, that is irrelevant, reducing dimension and irrelevant feature that can be used to produce a comprehensive model for classification. Additionally, this method used to decrease the overfitting problem and the size of data storage and can decrease the cost to train to obtain the highest accuracy [37]. Feature selection method for classification can be classified basically into three groups, filter, wrappers, and embedded method of feature selection.

Filter Method: This method of feature selection is considered as preprocessing steps. The filter method is based on the score gained from various statistical tests correlation between dependent and independent variables, feature selection without any dependency on machine learnings.

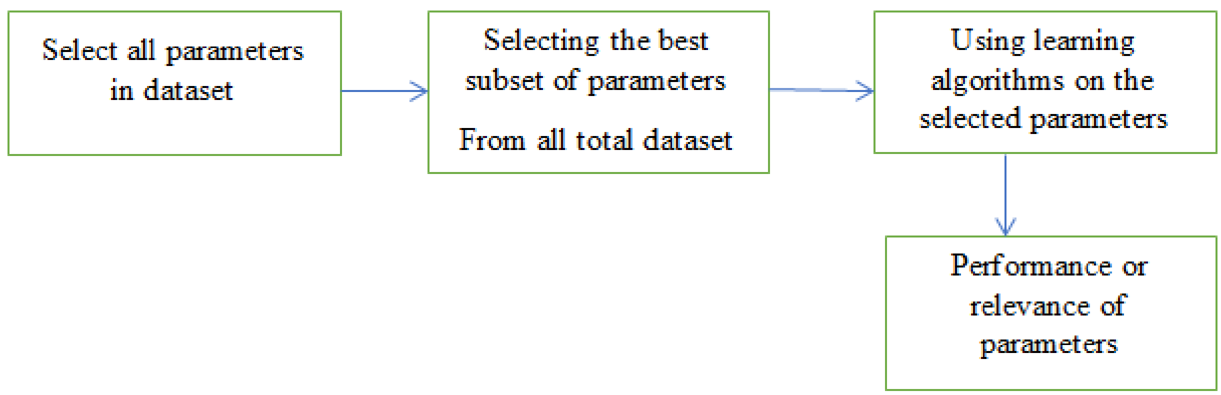

Figure 3.4: Filter Method Procedure.

The correlation has excellent guidance between dependent and independent variables identifications. The table below shows the correlate targets on the relation between these two categories of variables.

Table 3.3 Correlation Co-efficient Descriptions.

Feature /response	***Continuous***	***Categorical***
Continuous	Pearson's correlations	LDA
Categorical	ANOVA	Chi-Square

Pearson's correlations: it used to measure for quantifying linear dependence between dependent and independent variables represented by X and Y, respectively, with reading value -1 to 1.

LDA: linear discrimination analysis is an essential correlation in a combination of a linear feature that classifies or separates binary or multiple classes (or levels) of categorical variables.

ANOVA stands for analysis variance. It is similar to LDA excepts for the facts that are applied using one or more categorical independent parameters and single continuous dependent features or binary class identifications.

Chi-Square: it is statistical tests operated on the group of independent categorical parameters to evaluate the likelihood of correlation of association among them based on the number of frequencies.

Wrappers Method: In wrappers methods of feature selection, we try to use selected parameters and train the model using those selected parameters. Then, based on the trained models, we decide to add or remove parameters from these variables also. This method is essential to reduce searching problems, and it is usually costly. The table below shows some standard wrappers method used in feature selection.

Table 3.4: Common Types of Wrappers Method for Feature Selections.

No	***Common Types of Wrappers Method***	***Descriptions***
1.	Forward Selection	In the forward random iteration method, iteration starts without any features in the given models. In each iteration of forwarding selection, the method considered the summation of features till to improve the best models and until the addition of new parameters does not affect the performance of the models.
2.	Backward Eliminations	To improve the model's performance, the backward elimination method applies entire features and remove the least important feature from a given entire feature at every iteration of the models. This method iterated until no improvement observed in the removal of unimportant features.

3.	Recursive Feature Eliminations	The main goal of recursive feature elimination is to found the best performing feature subset with the help of greed optimization algorithms.

Embedded method: it is the combination filter and wrappers method of feature selections. It implemented by algorithms that have their built-in feature selection method.

Comparison between filter and wrappers method of feature selection: Both methods are used in feature selection, having different functions specifically. Below tables shows different function between filter and wrapper method in feature selection.

Table 3.5: Different Function of Filter, and Wrappers Method in Feature Selections.

Filter Method Function	***Wrappers Method Function***
It used to measure the relationship between features and target variables based on correlations. Are much faster parallel to the wrapper's method. It might fail to find the best parameters of features on many occasions. ***Filter Methods*** Uses statistical methods for evaluation of a subset of features.	It is essential to measure the significance of the subset features based on the actual training model. It does not include the activities of the training models. ***Wrappers Methods*** Are computationally costly parallel filter method. It can always provide the best subset of the feature. It used cross-validation for the evaluation of the subset.

So, based on the above comparison, the wrappers method of feature selection is more appropriate for this research works.

3.3. Development Tool.

The development tool discusses the tool used for system development to develop the framework for the proposed approach. There are different tools like Hardware (Arduino Uno, different sensor type, computer, and the other relevant Hardware), Software (Python tool, Edraw max, online drew max, Mendeley software, Tera Term, Arduino software (IDE), sensor, and other most relevant software. Those tools are discussed below with their function in this work deeply.

3.4. Designing Tool.

When designing the system based on IoT device and machine learning technique we have two primary or core component in case of water pollution monitoring or controlling at different location, those are hardware (any sensor, Arduino platform and mostly related hardware for water pollution monitoring) and software (mostly any program developed using different language like python language, etc.

3.4.1. Python Language

Python language developed since 1991 by GUIDO VAN ROSSUM and an open-source programming language that includes much supporting the library. The library is the best feature of python, making it one of the extensible platforms. Python is a dynamic programming language, and it uses an interpreter to execute code at runtime rather than using a compiler to compile and create executable bytes code. The unique identification of python language to accessible high level- language is that support to functional, imperative, and object-oriented programming with automatic memory management [38].

Why we use python: python is considered to be one of the most natural languages to learn for first-time programmers. Because when compared with other accessible objects- oriented languages such as C++ and Java, python has the following benefits for programmers:

- ❖ It is easy to read and understand
- ❖ It enables rapid prototyping and reduces development times.
- ❖ It has a humongous amount of free library package

HTML (hypertext markup language): - one of the more compelling ideas in computer technology is that a file can include code from perspective can be seen as data from which programming language. These means write programming python language and manipulating by HTML file is possible. Manipulating HTML was done when storing HTML tags in multiple python string and saving the content to a new file. So, based on this relation, HTML is used to call developed machine learning models by python language environment. Additionally, HTML is to enhance the user application platform with the integration of flask.

Flask: - is a small framework by most standard, small enough to be called "macro framework. "Flask is small enough that once the system becomes familiar with it, you can be able to read and understand or used all of the source code. Additionally, it is used to deploy the model on it since it used as a server-side platform [39].

Edraw Max tool is a type of design tool which enables the software designer to reliably create and publish many kinds of diagrams to represent an idea.

Tera Term: is the terminal emulators for Microsoft window that supports serial ports, telnet, and SSH connections. Tera term usually used to automate related tasks to a remote connection-initiated pc with serial, telnet, and another related supported device.

3.5. Implementation Tools

3.5.1. PH Sensor Device

The PH sensor is a type of sensor that used to measure the content of acid, base and natural based on PH standard values, such as below 7, 7, and above 7, those calls acid, natural, and base respectively. PH sensor can operate on 5-volt power and integrates with Arduino Uno to read real-time data. As standard, the normal range of PH is 6-to 8.5 [40].

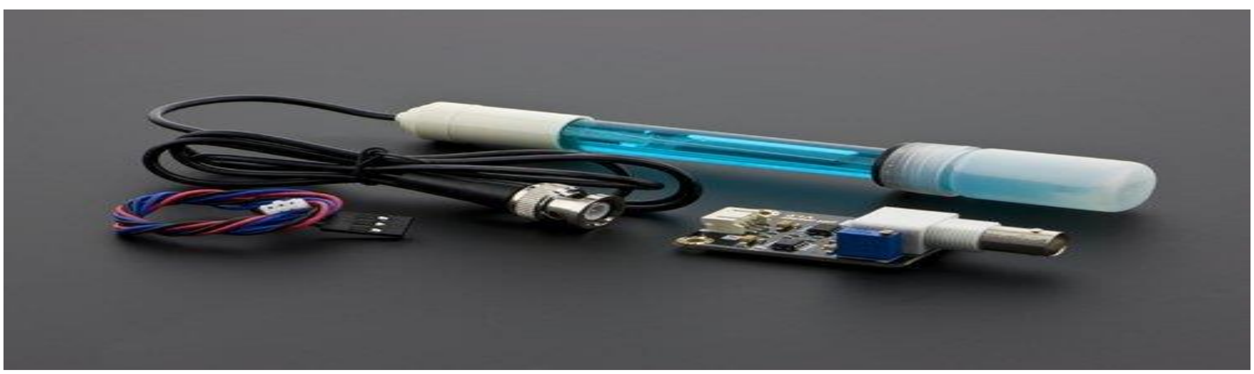

Figure 3.5: PH Sensor SEN0161 Device [41]**.**

A PH sensor is an electronic device used for measuring the PH level in the water. It consists of three types of probes (i) Glass electrode (ii) Reference electrode (iii) combination of gel electrode ph value described as the "negative logarithm" of hydrogen ion concentration in water. $pH=-log[H+]$ A pH meter consists of unique probes that connected to an electronic meter that would display the reading. If the pH level is greater than 7, then it is alkaline, if the pH level is less than 7, then it is acidic, and generally, the range of pH is 0-14pH [42].

$$\text{Ph\_val} = \frac{PH\_read * 14}{1023} \ldots\ldots (3.6.1)$$

Equation 3.1: PH Analog Read Formula.

3.5.2. Turbidity Sensor Device

The turbidity sensor device is also another popular sensor device used to measure the cloudiness of water in water bodies. The main aim of measuring the turbidity of water are used to indicate the degree at which water loses its transparency. Primarily, it is good to measure the turbidity of water for aquatic vegetations. As turbidity in water increases, the surface water temperature also increases above average due to suspended solid particles near to surface water facilitate observation of heat from regular sunlight [40].

3.5.3. Temperature Sensor Device

The temperature sensor is a sensor used to measure water temperature and indicates how water is hot or cold. The range of DS18B20 sensor is -55 to +125 ^{0}c. this temperature is a digital type that gives an accurate reading [40].

Figure 3.6: Waterproof Temperature DS18B20 Sensor Device [43].

3.5.4. Arduino Uno

Arduino Uno is the most microcontroller board based on the A Tmega328P. Does it have 14 digital input or output pins are the most selected boards (of which pins 6, used as PWM outputs, 6 pins, analog inputs, with a 16 MHz quartz crystal, and a USB connection, a power jack, an ISCP header, and the reset button [44]? It handles everything needed to support the microcontroller. Arduino Software (IDE) was the reference version of Arduino, now evolved to a newer release. The Arduino Uno board is the first in a series of USB Arduino boards and the reference model for the Arduino platform; for an extensive list of current, past old boards, see the Arduino index of aboard [40].

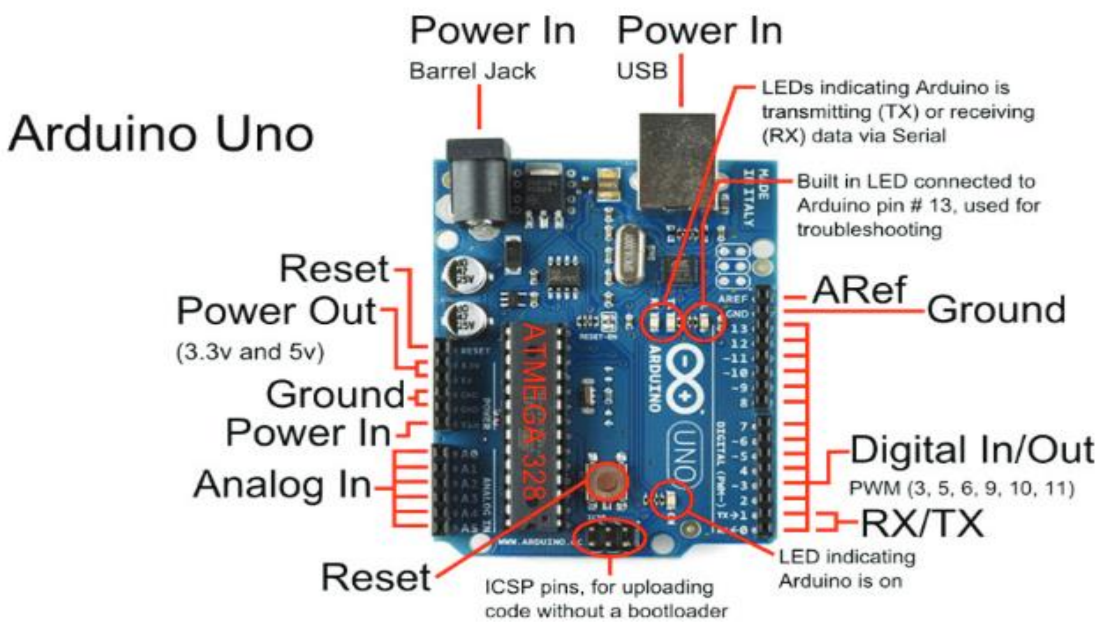

Figure 3.7 Arduino Uno Device [45].

3.6. Machine Learning

Machine learning algorithms are a type of artificial intelligence that enables digital learning from patterns in data without explicit programming, creating a model that can be applied to new data. Machine learning algorithms tools can be classified into three main categories (such as Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning). Among these machine learning algorithms, ***supervised learning*** is the appropriate machine learning algorithm for the proposed systems because the classification technique applied to the proposed system. In classification, a methodology used for partitioning data instances into a set of categories which were also referred to as classes or labels (such as ‘Polluted’ or ‘Suitable’). The classification methodology can be based on using supervised learning, which learns directly from relationships among variables in available data instances or patterns (usually referred to as the training set).

3.6.1. Supervised Learning

Supervised learning algorithms are an algorithm to generate a function that corporate (join) input to output based on labeled data. In this algorithm techniques, the learner requires to predict the given input into appropriate classes based on the given input to output [46]. The most common and relevant supervised learning techniques to solve such a problem are classification and linear regression.

$$Y = f(x) \qquad (3.6.1)$$

Equation 3.2: Linear Classification for Supervised Learning Algorithms.

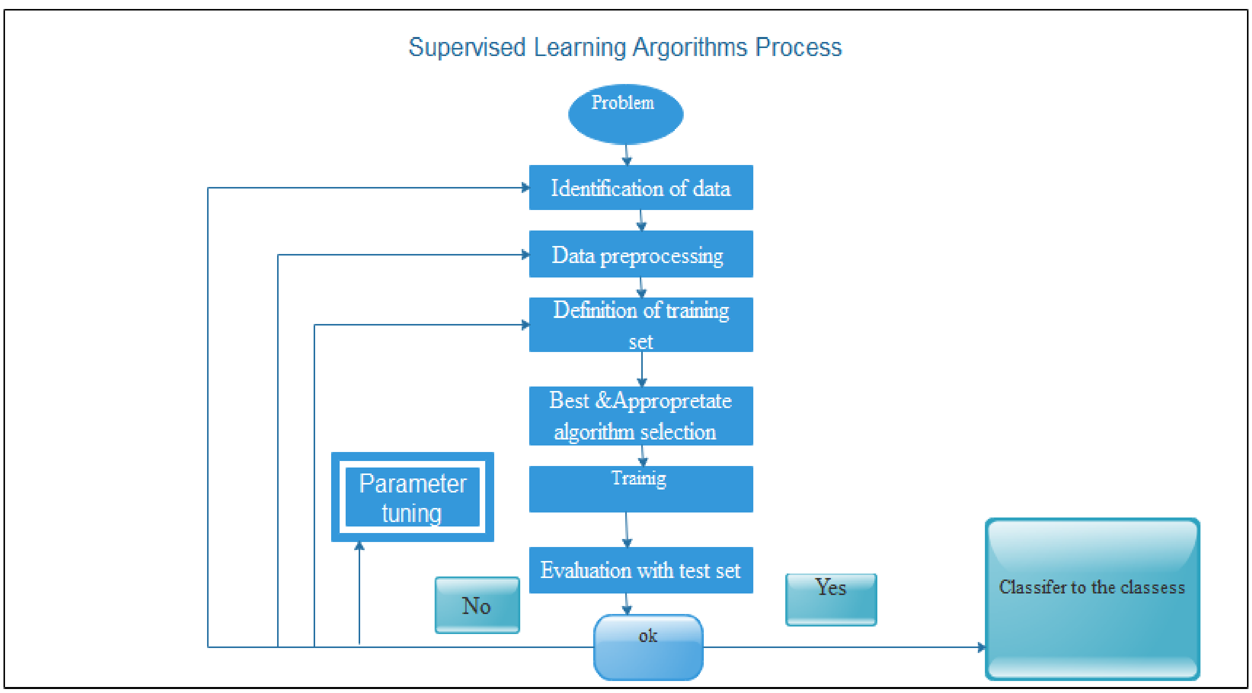

Figure 3.8: Supervised Learning Techniques Processes [32].

In a supervised learning technique, they are three things considered: the dependent variable, independent variable, and function that were used for cooperating (generating) and predict the effect of the required output.

3.6.2. Unsupervised Learning

Unsupervised learning is another type of machine learning tool which can search for invisible pattern in data without labeled data examples. Most of the time, unsupervised learning used for clustering techniques rather than classification techniques. It tries to disconnect information from the input with an intrinsic feature extraction or the total amount of input data. In unsupervised learning, the data can be categorizing without any training data or column name; only input data are available to learn data.

3.6.3. Semi-Supervised Learning

In Semi-supervised algorithms, both labeled and unlabeled data were combined in order to improve the performance of generalization and to perform an otherwise supervised and unsupervised learning task. Since these learning types use both labeled and unlabeled training data, predictor learned, that predict future test data better than predictor learned from labeled one.

3.6.4. Reinforcement Learning

Reinforcement learning is also a type of machine learning which determines the interaction of the agent with the environment based on feedback that might be positive or negative. The objective (aims) of this learning is to determine action within the specific environment is improving positive feedback is called a reward. Unlike other machine learning or (supervised, unsupervised, and semi-supervised), reinforcement learning does not depend on the training data; instead, it is based on action or interaction that happened between agent and environment. The interaction happens with the environment but not with data [47].

Data analysis: in the machine learning model, they are many analyses used to predict the models. The main steps used to data analysis in machine learning models are: -

I. ***Understanding of required data*****:** before move forward to work directly with data, it is necessary to analyses original data first. Identifying measurement used in data analyses like - any missing data required, the possible model used to data analysis, target (objective of the research is defined.

II. ***Data preparation:*** **-** formulating the data as required (merging of related data, filling of missing data, excluding too many missing values or variables in the data, and finally sorting the data as the target.

III. ***Model training:*** **-** based o the given data, train the model and analysis of the data with fitting models.

IV. ***Evaluating the result:*** **-** it is the most significant step of data analysis, which means an understanding of results that facilitate the adjustment of models and correction of the initial targets of the research objective. Complete data analysis for the machine learning model is the precondition used for model correction and predictions [48]**.**

3.7. Classification Algorithms

Classification is a supervised learning concept in which the problem is identified and classified into their appropriate categories, on a given trained dataset containing instances whose category membership is specified. In supervised learning, the classification approach was based on a given data (or input data) to the computer program that learns from it. A linear classifier algorithm has been used to achieve the objective of the study. A linear classifier is a subtype of supervised algorithms that classify or group the items into their class type. This type of classifier performs its classification based on linear predictor function combining weight with the feature vector.

Supervised learning aims to build a machine learning classifier that is used to classify unknown samples with proper labels, polluted or not -polluted(suitable) class in case of this study. They serve as the most critical areas of water pollution monitoring and predictions with the features. Machine learning is the most popular technique rather than a data mining technique for classification. In table 3., 8.1 Several machine learning-based classification algorithms were briefly described. For more detail, it can refer to [49].

Selecting the appropriate classification algorithm according to water pollution monitoring and prediction is very important since it predicts accuracy and performance enhancement. When the training dataset is too large, linear regression (LR) is selected. If we need to observe the accuracy possibility of each class, logistic regression is better to choose. If they are many types of extracted feature to be considered, random forest classification could be a good classifier. In particular work of machine learning classification, a large number of experiments were needed for setting appropriate parameters and algorithms for finding the most effective result.

The class labels of the instances in the training set are known, and the aim is to build a model to label new instances that were presented. An instantiation of the particular algorithm (such as SVM (Support Vector Machine), KNN (K-Nearest Neighbors, Linear-Regression, and RF (Random Forest) for a specific training set are called a classifier [50].

Support Vector Machine is a binary supervised classification algorithm that is mostly used in classification problems as extensive as possible through a hyperplane, which is called a support vector. The hyperplane is the decision boundaries that help to classify the data point. The hyperplane in n-dimension is an affine subspace of dimension n-1 [51]. Especially, the rate success of the support vector machine is high when the training dataset is good enough.

$$W*x+b=0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.7)$$

Equation 3.3: Calculate the Form of Hyperplane in Support Vector Machine.

K-Nearest Neighbors: - k-nearest neighbors are supervised classification algorithms that used to classify datasets based on the most similar data point at the training stage and simple method used. As the name indicate that it considered the nearest neighbor data to assign the class for classification based on the most K nearest positive integer numbers.

Linear-Regression, as the name, represents that it is a linear relationship between independent and dependent predictor variables. They are many types of linear regression algorithms for variable prediction. Mostly simple linear regression and multiple linear regressions are applicable in independent and dependent variable prediction purposes. ***Simple linear regression:*** is a statistical technique for obtaining a formula to predict [52] the value of one variable from another when there is a causal relationship between the two variables. Generally, statisticians generate a preferred formula for simple linear regression involving beta: [52].

$$Y = \beta_0 + \beta 1x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{equ } (3.7)$$

Equation 3.4 : Simple Linear Regression.

Multiple linear regressions are also a form of linear regression used for more than one prediction variables and require multiple inputs. The blew equation represents the relationship between many independent and dependent predictor variables through multiple linear regression [53].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + £ \ldots\ldots \text{ equ } (3.7)$$

Equation 3.5 : Multiple Linear Regressions.

Naïve Bayes (NB) is a probabilistic classifier technique based on Bayes theorem with an assumption of independence among predictors. In a simple description, the NB classifier tool considers that the presence of a specific feature in-class categories is unrelated to the presence of any other feature. The equation for Bayes Theorem [51].

$$P(c/x) = p(x/c)\, p(x)/p(x) \ldots\ldots (3.7)$$

Equation 3.6 : Calculation for Bayes Theorem.

Random Forest Classification is a supervised classification that sets or generates forest. It is the most classification algorithm that uses to create the forest with several decision trees that operate by constructing a multitude of the tree at training time. Random forest uses a node of the tree to use the best possible tree among a subset of predictors of features selected at a random node. The building blocks of the random forest node are decision trees (DT). It uses bagging and feature randomness when building each decision tree.

Table 3.6: The Most Frequently Used ML Classification Algorithms in Water Pollution Monitoring and Prediction.

Algorithms	***Advantages***	***Disadvantages***
Support vector machine	High accuracy and speedy, deal with high dimensional, handle imbalance class and non-linear problem;	Heavy storage burden, does not perform very well when the data has more noise or overlapping class, require to choose kernel function and repeatedly call parameters accurately;
K-Nearest Neighbors	High precision and accuracy, non -linear classification, and no assumption of features;	Sensitive to unbalanced sample set, heavy computing burden;
Random forest	Overcome overfitting, deal with high dimensional data;	Accuracy depends on tree number, sensitive to unbalanced samples set

Logistic regression	Output has an excellent probabilistic interpretation, and the algorithm can be regularized to avoid overfitting;	Weak to handle high dimensional data;
Naïve Bayes (NB)	Even though the conditional independence assumption rarely holds, NB models perform surprisingly well in practice, especially for how simple they are. They are easy to implement and can scale with datasets.	Due to their sheer simplicity, NB models are
Decision tree	They are robust to outliers, scalable, and able to model non-linear decision boundaries thanks to their hierarchical structure naturally	Easy to overfitting

Scikit-learns

Scikit- learns is the free software ML library, which is used for implementing machine learning algorithms including, Support vector machine, linear regression, KNN, and RF. The data set and CSV files are trained to give appropriate prediction results for the given sample water parameters for a specified location [25].

3.8. Proposed System Design Method

The system design has been designed to input dataset based on designed system parameter evaluation and give output at the end. Specifically, the designed system is to evaluate parameter or variable on the given input dataset, identify the water contamination, and gives some guidelines for it.

3.9. Proposed System Implementation Method

A framework for water pollution monitoring using Sensors and machine learning techniques is based on the proposed solution framework designed at the proposed solution section. The designed architecture or framework has been implemented using a sensor device to Arduino, Arduino to computer device integration in machine learning technique and connection server with python.

3.10. Proposed System Evaluation

This proposed system verification and validation evaluation run on python with machine learning algorithms. A developer can perform the training and testing of the dataset during development. First, the developer import dataset and run the code on the code editor platform. Finally, the system generated the result. After the researcher has applied the above-listed steps, the performance of the expected predicted values was gained or evaluated. The evaluation is essential and mandatory to analyze our work result very well. In this study, the researcher train, test, and evaluate the performance of the proposed system.

3.10.1. Cross-Validation

Cross-validation is a commonly known model evaluation method for reducing scales of overfitting problems and increasing the accuracy of the models. It is an additional method of training model technique for evaluating, and the effectiveness of the result is based on the best-chosen models for available data samples. They are several types of cross-validation techniques, which all have their pros and cons, among that technique: -

Hold- Out Validation: the concept of this validation technique is dividing the data into training and test set (more of the time 70% is for training, and 30% is for the test). The trained models are based on the first 70% of data, and then the result of the prediction for the remaining 30% is compared with the actual result. It is known techniques to estimate performance and have cons since the amount of data used for training were reduced.

K- Fold cross-validation: - the consideration of this technique is in dividing (categorizing) the data in k parts. The model trained using n-1 parts of data, and the rest of the data is used for testing. The procedure is repeated n times; as a result, each of n pieces of data used for testing. This method is commonly applicable and used for this research as well.

Leave-one-out- cross-validation: - is a unique case technique of k- fold validation, where on each iteration, all data except for one observation is used for training, and assessment is done using this one observation.

Table 3.7 : Comparison of Cross-Validation Technique.

Validation method	***Pros***	***Cons***
Hold – out validation	Independent training and test	Reduce data for training and testing Large variance
K-fold cross-validation	Accurate performance estimation	A small sample of performance estimation; overlapped training data; elevated type I error for comparisons; underestimated performance variance or overestimated degree of freedom for comparison
Leave one – out cross-validation	Unbiased performance estimation	Huge variance

3.11. Evaluation

Evaluation of this research study on a framework of water pollution monitoring and prediction models development for polluted water, all in all, the evaluation conducted the results discussed and concluded by recommending the best machine learning model for water pollution monitoring and prediction.

3.11.1. Prediction Model Evaluation

The water pollution prediction or monitoring must be evaluated using validation technique to see whether the model fits with the dataset and to validate the most model fit correctly for work when the new data/ real data were coming to the model for prediction. So, we used a cross-validation method to make sure that models got the pattern for the training dataset. Cross-validation is the most resampling procedure used to evaluate the machine learning model on a particular dataset sample. In order to validate the model, we used the K-fold cross-validation is selected because of its popularity and its simplicity for implement on the selected machine learning classifier, and it avoids overfitting as well as an underfitting problem [54]. The model testing technique used K-fold CV to evaluate the machine learning model on a particular dataset. The entire dataset with the feature is split into K part and K-1 used for training and one used for the test set, and also helps to prevent overfitting and underfitting problems.

3.11.2. Model Performance Evaluation, and Prediction Metrics

Prediction model evaluation is required to quantify the prediction model performance because of classifier algorithms give a biased result. Besides the CV testing, this study used different matrices appropriate for model performance evaluation. Metrics used, namely are accuracy, precision, recall, F-score, and confusion metrics. Terms related to performance evaluation are: -

TRUE POSITIVE (TP): is the class when the actual class of the data point was True, and the predicted value is also True.

TRUE NEGATIVE (TN): This was the case when the actual class of the data point was False, and the predicted value also false.

FALSE POSITIVE (FP): are the cases when the actual class of the data point was False, and the predicted values are accurate.

FALSE NEGATIVE (FN): False-negative was the case when the actual class of the data point was correct, and the predicted value is False.

3.11.2.1. Accuracy

Accuracy shows the classification problem correct prediction value and calculates the total number of the model prediction divided by all number of data instances used for the model [54]. Accuracy calculated using the following equation:

$$\text{Accuracy} = \frac{TN+TP}{For\ all\ number\ of\ instance\ data} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.12.2.1)$$

Equation 3.7: Calculate the Accuracy of Model Prediction.

3.11.2.2. Precision

Precision measures the predicted value actual, and it shows how many times the model predicts true. Precision answer what proportion of identification was correct? Moreover, precision calculated using the equation:

$$\text{Precision} = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.12.2.2)$$

Equation 3.8: Calculate the Precision of Models.

3.11.2.3. Recall

Recall answers what proportion of actual positives was correctly identified. Recall calculating using the below equation:

$$\text{Recall} = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.12.2.3)$$

Equation 3.9: Calculate Recall of Models.

3.11.2.4. F-Measure

The F- measure (F1- score or F-score) is measured of testes accuracy and is defined as the weighted harmonic average of the precision and recall of the test. F1 is more useful than accuracy when the dataset contains uneven class distribution.[54] This F1 -score is calculated using this equation:

$$\text{F-score} = 2 * \frac{(recall * precision)}{(recall + precision)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.12.2.4)$$

Equation 3.10: Calculate the F1-Score of Models.

3.11.2.5. Confusion Matrix

A confusion matrix is one of the most techniques for summarizing the performance of classification algorithms [55]. The number of correct and incorrect predictions is summarized with count value and broken down each class. It gives insight not only into the error being made by the classifier but, more importantly, the type of errors that are being made. So, this study uses a normalized confusion matrix for binary class models classification (polluted or non-polluted/ suitable) class. The following table 3.12 shows a sample format of a confusion matrix, with 2-classes or binary classification [56].

Table 3.8: Sample Format of Confusion Matrix for Binary Classification.

		Predicted Value	
		Suitable	***Polluted***
Actual Value	***Suitable***	True suitable	False polluted
	Polluted	False polluted	True polluted

CHAPTER FOUR

4. Designing Proposed Solution for Polluted Water Monitoring Model

This chapter describes the proposed architectural solution for real west water monitoring and sensing using a machine learning approach and Sensors based. The research question was defined to solve the problem mentioned inside the statement of the problem under chapter one for this works. System architecture or framework is proposed based on the literature discussed under chapter two, with related work to succeed in this framework development.

4.1.1. The Proposed Framework for Polluted Water Monitoring

The proposed frameworks used to water pollution monitoring and prediction into suitable and polluted classes. The proposed solution is based on the framework shown in figure 4.1. It takes samples of polluted water dataset as input and then the preprocessed based on the nature of polluted water and another essential necessary preprocess. After all preprocessing came over, then feature selection and extraction takes place to extract the parameters using PCA, ICA, LDA, LLE, t-SNE, and AE. The output of these tasks is a crucial feature suffering from the overfitting of the dataset for training the model.

After feature extraction, models developed using RF, SVM, LR, and K-NN machine learning algorithms and training set with feature regularizations data frame of which dataset. The model evaluated using k-fold cross-validation or k-fold model selections; the result used to select the best monitor and prediction, model. The outcome of these tasks is a model used for suitable and polluted predictions classes. Finally, the prediction model is evaluated and selected based on the results obtained using the model evaluation method discussed in chapter three. The final selected prediction model used to develop a prototype that can take real-time water pollution datasets as input, and it classifies the input, whether it is suitable or polluted.

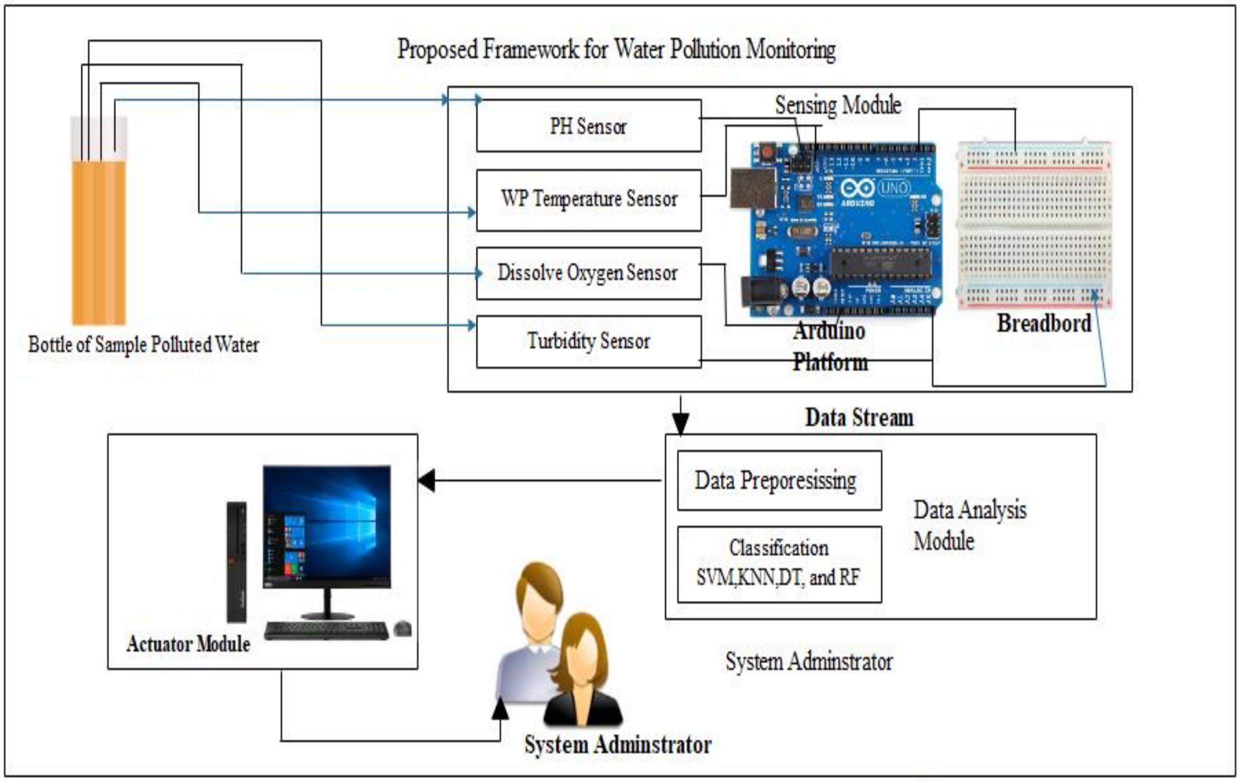

Figure 4.1: Overview of the *Proposed* Framework

4.2. Overview of the Proposed Framework

The proposed system framework for Sensors and machine learning approach-based water pollution monitoring are the integration of hardware and software component used in water pollution monitoring. The hardware component used in this proposed system is Microcontroller (Arduino Uno) and sensor to sense real data (such as PH sensor, Dissolve oxygen sensor, Turbidity sensor, and Temperature sensor), and Wi-Fi module for data transmission from Arduino to cloud server, from Arduino to data analytical module, from data analytical module to cloud server and finally from cloud server to actuator module or end-user applications.

The software component used in this proposed architecture is a computer application, Arduino soft software (IDE), the dashboard for water pollution monitoring, and prediction. The real data collected using a sensor is transmitted to the cloud server for storage, and analytical prediction based on the model developed using machine learning algorithms and stored on it. The Wi-Fi protocol is used as a channel for data communication between sensors collected data through Arduino and to the cloud server.

The developed model data compared against WHO and FAO standards for comparison of the developed machine learning model and real-time data prediction. The end-user assesses water pollution monitoring and sensing using the mobile or computer device application. The data analysis is performed on a dataset consisting of 1992 samples gathered from different paper reviews and interviews from three water pollution sources of Ethiopia and use for the machine learning model. In all, the proposed system framework consists of these three basic modules (sensing module, analytical data module, and end-user application) for water pollution monitoring and sensing in real-time.

4.3. Sensing Module

Under these categories, there are three most influencing water pollution monitoring and measuring sensors. Such as PH sensor, Turbidity sensor, and Water Temperature sensor. These sensors are connected to the Arduino board for collecting real data, which further send to the cloud server with an interval of 1 minute. Arduino wifi shield has been used for communication medium or channels when data is sent to the cloud server from the sensor node. The connection diagram of the hardware is shown in the figure below.

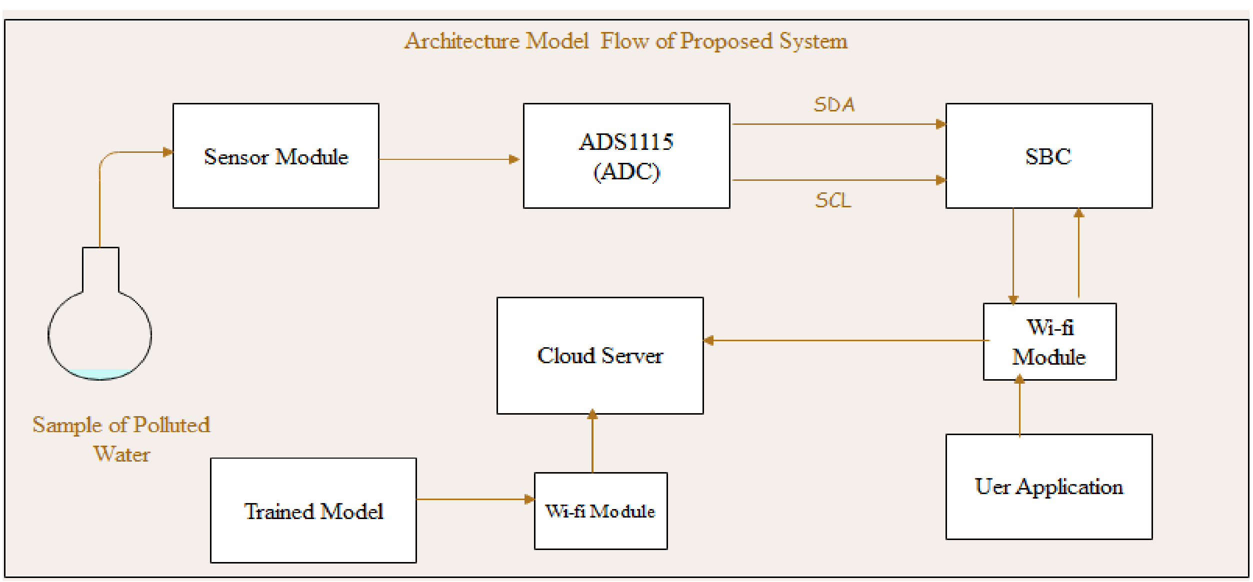

Figure 4.2: Framework Model Flow for Proposed System

4.3.1. Sensor Module

For this proposed system, around three types of sensors applied to sense or collecting real data of polluted water:(such as Analog PH Sensor, Dissolve Oxygen Sensor, Turbidity Sensor, and Temperature Sensor) with detail function of the are discussed or listed below. One is an analog PH sensor which has high response time (1min) during real data collecting, and capacity to monitor and measures water PH values in all range (0-14) for PH values of polluted water.

SKU: SEN0161 sensor provides the output with high accuracy of PH water and operates with an input voltage of 5VDC. It is also a secure device to integrate with both Pi 3 B models and ADS1115. Mostly it is a low-cost device with broad technical support freely available on the internet. So, for the above-listed reason or factors, SKU: SEN0161 is selected among another PH sensor device used for prototype development [57].

The PH sensor is used to measure the acidity or alkaline nature of water. As referring to WHO and FAO standard and study of this system development, if PH > 8.5 or 9, the water is considered as alkaline, if PH < 6 or 6.5 the water is considered too acidic and both of them classified to polluted class while the PH > 6 OR 6.5 >X<9 the water is considered to reuse for irrigation, then it classified under suitable class.

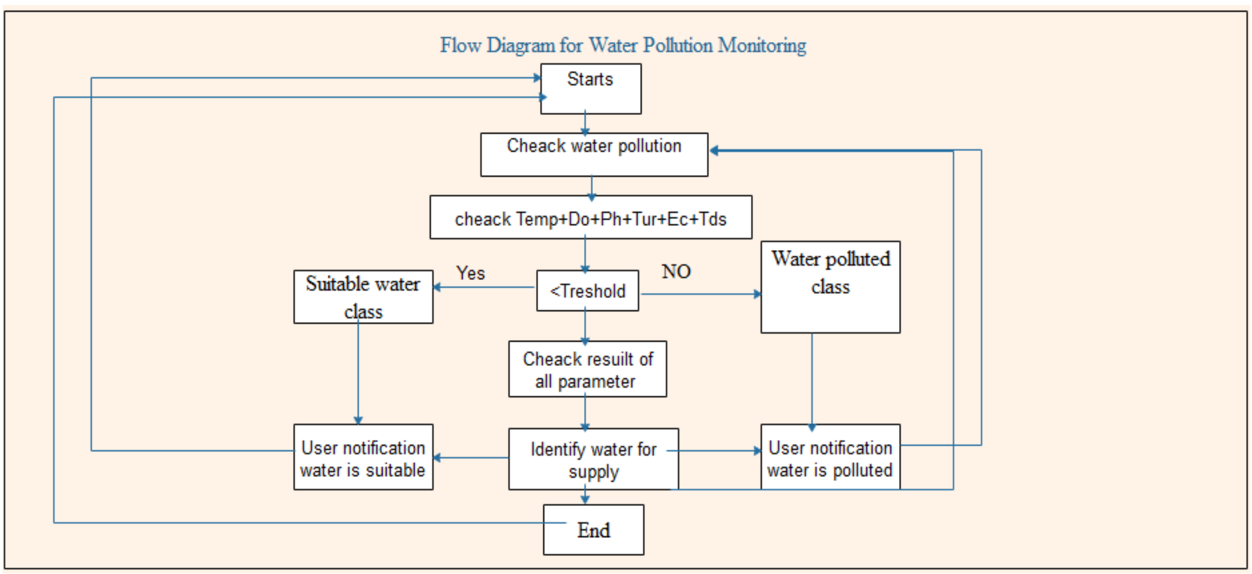

Figure 4.3: Flow Diagram for Water Pollution Monitoring Systems

Another most crucial parameter to determine the water pollution level is turbidity, which is used to measure water transparency and the level of dissolved oxygen in the water. WHO states that the turbidity level is between 0-15NTU or below 15NTU for plantation, and on another hand, the turbidity level is less than 2NTU to 5NTU; it is preferable; for drinking water? Temperature also another type of significant parameter which is directly water quality. Therefore, water temperature is monitored when the PH water values and the turbidity level of water were calculated.

Arduino Uno (SBC) there are several SBC platforms are available real analog data reading converting to digital data format with the development of the data acquisition unit, and the digital signal is carrying by ADS1115 channels to python library. Among SBC platforms, the essential feature is the Raspberry Pi 3 B+ model for the following reasons. Such as it uses python as the primary programming language hence bring simplicity, scalability, and richness to the library [58].

It is also high processing speed, memory, and support to high external storage up to 32GB, low power consumption, and less expensive when we compare with other boards like Nucleus-F410re, and ODROID-XU4 Boards [59] [60][61][62].

4.3.2. Data Analytical Module

The other categories are analytical data module, which predicted analysis is performed on the dataset acquired from sample imputed dataset. The data set consist 1992 sample size and six factors or parameters gathered from the currently available sample data set relative to three areas of water pollution center with the support of detail studied journal paper and interview with exact and relative department and institution on some multiple water pollution measurement parameters.

Most of the researcher states that manually lab analysis of water pollution is not estimated at the exact standard index. The proposed model improves those notions, and the procedure followed in the blow figure 4.4 steps.

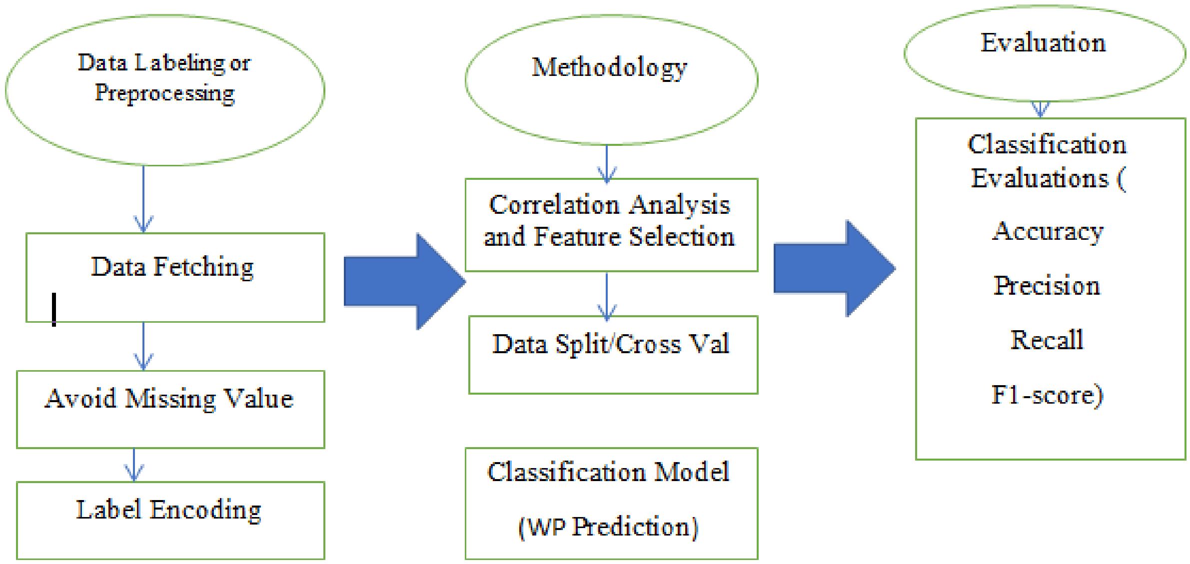

Figure 4.4: Procedure Flow in Model Evaluation Steps.

4.3.3. Propose Feature Selection

The proposed feature selection component of the prediction and monitor system perform the selection of important feature from a sample dataset. These methods take input of preprocessed and label datasets to perform selection, as shown in figure 4.5. Then the selected feature used for ranking the importance of each feature for the target class and contribution of each feature for target classes. Then, the selected feature used for training and tests the model to predict the class of water pollution as polluted and suitable.

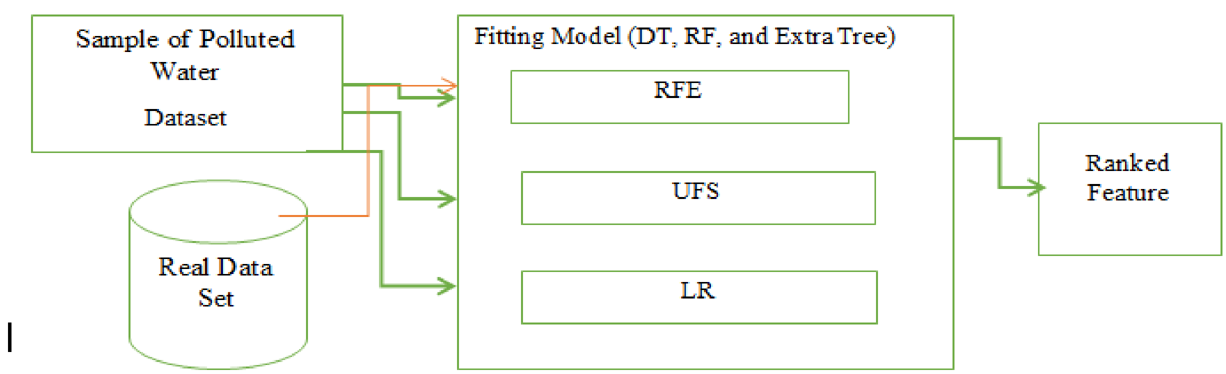

Figure 4.5: Feature Selection Flow Diagram.

This component uses RFE, UFS, LR, and CMA feature selection methods are well known in the water prediction approach, as described in chapter three. Each method gives selected features used to train a machine learning classifier method.

4.3.3.1. Recursive Feature Elimination

The recursive feature elimination method is the most feature selection method. RFE takes the machine learning model as an input instance and, finally used to identify the desired number of features for use. In this study, RFE takes a random forest model as input with all features in this dataset. They are six features in this dataset and, the most four features selected as the desired feature for use. Also, RFE is used to rank the most desired feature in order of importance. Examples, in this study, the desired feature are ranked or selected were temperature, conductivity, TDS, Turbidity as most four desired features among those six features using the RFE selection method.

4.3.3.2. Univariate Feature Selection.

UFS is the other most crucial feature selection method. UFS is a statistical method used to select features based on the relation between dependent and independent features. UFS has used Select K Best and Chi2 method for selecting the best feature based on the scoring function we need. In this study, UFS was based on the scoring scored by precision, recall, and f1-scores.

4.3.3.3. Lasso Regression

Lasso regression feature selection is also another essential feature selection method due to avoiding overfitting. Lasso regression used the regularization method to discharge unnecessary features for model based on their coefficient of the input features. Generally, they are several feature extraction and selection methods. However, the above-listed method is the most appropriate feature selection in this study. Additionally, correlation matrix analysis is also used in this study.

4.4. Machine Learning Model Building

In these subcomponents of the proposed frameworks for water pollution monitoring, pollution prediction performs machine learning classifiers training on all-important features constructed by feature selection component method. This study builds classification models mainly by using machine learning and collected real data using sensors.

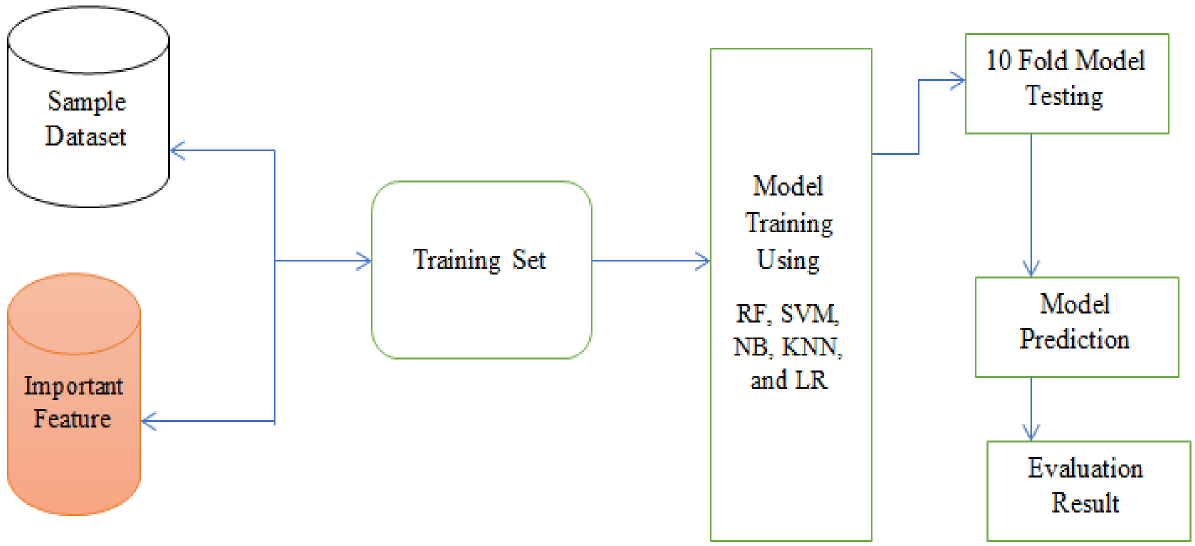

Figure 4.6: Model Building Flow Diagram.

4.5. Model Definition

The model definition is a primary component for a machine learning model in which a data source is created for training the model. They are three primary activity under the model definition component have been defined.

Get Data: the first thing to perform machine learning is data. In this study, the dataset is obtained from Assela malty factor; some samples downloaded from the west water treatment project and final real-time data from sensor device data for monitoring and predictions. This dataset includes the number of both suitable and polluted water indications, including a feature that differentiates whether water is suitable or polluted.

Prepared Data: - a dataset is usually constraints; some preprocessing earlier in it can be analyzed. Data preparation is the process of obtaining data ready for machine learning. They may be noticed with the missing value and scaling presented in the column of varies rows. In this study, the row that missing values have been filled using the mean and median method.

Define Feature: - in machine learning, features that are used as an independent variable is the individual measurable attribute of water content. In our dataset, each row represents one category of water class, and each column is the feature of the water content.

4.6. Model Training

In this section, the algorithm performance concerning feature selected constructed from the need; the given polluted water represented request. Here, our objective is to select the best classification algorithms for water pollution prediction and monitoring system.

The training model is the process of constructing the predictive machine learning model that decides the unknown water, whether it is polluted or suitable for irrigations. It is mathematical modeling with parameters that maps the input variable into the target variable, i.e., output. The entire or label dataset is divided into two categories, one which is used in training the model, i.e., the training set, and the other is that in testing the model after training, i.e., testing set.

***Training set:* -** Training set is used to build a model. It consists of the correct answer, which is used to train the system. The learning algorithm finds the pattern in the training set that map the unknown water attribute or feature values to the target values (answer the water is polluted or suitable), and it output the machine learning models that predicts or monitors these patterns. Training rules and algorithms used to give relevant information on how to associate input data with output decisions or target results. The framework or model is trained basically by applying four machine learning tools on the dataset; all the data and results were obtained. Generally, 70% of the data of the dataset is taken for training data [63].

***Testing Set:* -** the testing set is used to test the classifiers. It is a set of data that was used to verify whether the algorithm is producing the correct output after being trained. Generally, 30 % of the data among the dataset is used for testing. Testing the data is used to measure the accuracy of the classifiers. All in all, the accuracy of the model mostly evaluated using the holdout method and K-fold method of evaluation technique.

4.7. Models Evaluation and Testing

In order to test and estimate the learning ability of machine learning models trained on the polluted water prediction and monitoring dataset. The fundamental concerns of the machine learning models are to find an accurate estimation of the generalization error of the trained models in finite datasets. Because of the reason this study is used, proposed k-fold cross-validation (CV) and different performance evaluation matrix appropriate for models. These matrixes are confession matrix, accuracy, precision, recall, and F-measure, as discussed deeply in chapter three.

CHAPTER FIVE

5. Experimental Set-Up and Implementation

5.1. Experimental Setup

This section discusses the experimental environment used to conduct this research work. Conducting the experimental for this proposed system or real-time water pollution monitoring and controlling is time-consuming due to complexity. So, Sensor device and machine learning approaches are used.

5.1.1. Working Environment

- ***Laptop:***

  -Intel(R) Core (TM) i5-7200 u CPU @2.50GHz 2.70GHz

  - Installed memory (RAM): 4:00GB

  - System type 64bit operating system, x64 based processor

- ***Anaconda***

Anaconda is a packaged set of programs, including the python language, a huge number of libraries, and several tools [64]. These include the spider development environment, Jupiter notebooks, RStudio, orange3, and Jupiter. Among these three of them are applied or used while comprising the research project for python language. Such as Jupiter lab, Jupiter notebook, and spider.

JupyterLab provides flexible building blocks for interactive and exploratory computing [65]. While JupyterLab has many features found in traditional integrated development environments (IDEs), it remains focused on interactive, exploratory computing [65]. ***JupyterLab*** is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. JupyterLab is extensible and modular: write plugins that add new components and integrate with existing ones [66].

Jupyter Notebook, the Jupyter Notebook (formerly I Python Notebook) allows you to execute, store, load, re-execute a sequence of Python commands, and to include explanatory text, images, and other media in between [38].

Spyder is the Scientific Python Development Environment: a powerful interactive development environment for the Python language with advanced editing, interactive testing, debugging and introspection features and a numerical computing environment thanks to the support of Python (enhanced interactive Python interpreter) and popular Python libraries such as NumPy (linear algebra), SciPy (signal, and different types image processing) or mat plot lib to graph (interactive 2D/3D plotting) [38].

5.1.2. Programming Language

- ***Python:*** for implementation of supervised learning algorithms and data learning strategies because it has better numerical and categorical classification and scientific libraries that provide an environment of data analysis in terms of number-based class classification for the specific task.

Implementation or Working Environment: this study used several tools, machine learning tools, and packages to implement the proposed solution for water pollution monitoring and prediction.

5.1.3. Implementation Environment

This study used several development tools and packaged to implement the proposed solution of water polluted predictions. This study uses the python programming language for implementing and experimenting with each proposed solution from the data preprocessing phase. Moreover, to evaluate the implemented proposed classifies models. Python is mostly used because of its programming language of choice for developers, researchers, and data scientists who need to work in the machine learning model with the integration of Sensors technology. ***Table 5.1*** shows the lists of tools, Arduino Uno package tools, and python package with their version and description, which is used in this study.

Table 5.1: Description of the Tools, Arduino Uno, and Python Package Used During the Implementation

Tools		
Name of Tools	***Version***	***Description***
Arduino Uno	1.8.9	Arduino Uno is the most popular among other Arduino boards for prototype platforms. Because Arduino Uno is open-source based on easy to use the integration between hardware and software platforms. Also, Arduino Uno consists of a ready-made software called IDE (integrated development environment), which is used to write and deploy the computer code to the physical boards of Arduino Uno [67].
Waterproof Temperature	DS18B20	Used to measure temperature in a wet environment with the simple one-wire interface. It provides 9 to 12-bit (configurable) temperature readings over a 1-wire interface. So, that needs only one wire VCC (and ground) needs to be connected from a central microprocessor (Arduino Uno).
PH Sensor[41]	SKU: SEN0161	It is used to measure water quality and some related parameter values. Because it is designed for Arduino controller and built-in simple, convenient, and accessible, practical connectional designed features. Also, it can read the 0-14 range of PH water and 0-60^{0c} of temperature values.
Anaconda Navigators	1.9.6	Allow to launch development applications, and easily manage Conda packages, environments, and channels without the need to use command-line commands.
Tera Term		Tera terms (windows) are one of the current, more related windows terminal programs. It is used to store data from another device to a computer device with any formats. Tera term can communicate with other devices through serial ports. So, in this study, Tera term used to store real-time data read by the sensor on a device with CSV formats.

Jupyter Notebooks	5.7.8	An open-source web application that allows you to create and share a document that contains live code, equations, visualizations, and narrative text. Its uses include data cleaning, transformation, numerical simulation, statistical modeling, data visualization, and machine learning.
Python	3.7.4	Easy to learn, powerful programming language to develop a machine learning application.
Microsoft Excel	2016	Used data preparation tasks in cleaning, fitting, and labeling the gathered data. Moreover, to manage the classification task.
Python Packages		
Scikit-Learn	0.21.3	A set of python modules used for machine learning algorithms and data mining [68]. This study uses it for feature extraction, feature selection, training, and testing of the models.
Pandas	0.25.0	High-performance, easy to use data structures and data analysis tools. This study used it for data reading, manipulation, writing, and handling the data frame.
NumPy	1.17.0	Array processing for number, strings, and objects. This study uses it for handling converting the categorical to numerical data for features, training, and testing models.
Matplotlib	3.1.1	Publication quality figures in python. This study uses it for data and results visualization.
Flask	1.1.1	Microframework for making web service in python. This study uses it for implementing for selected models.
Arduino Uno Library		
Name of Library	***Function of Each Library***	
Software Serial Library	It is used to allow serial communication on other digital pins of the Arduino Uno, using serial communication software. Because it is possible to replicate the functionality (which is called software serial communications).	

	It can have multiple communication software serial ports with speed up to 115200bps. So, in this study, this library plays a significant role in communication between the computer and Arduino Uno hardware.
One Wire Library	One wire library Provides routines for communication using Dallas, one wire protocols, and slave protocols. Because all communication cabling required single wire. So, in this study, the waterproof temperature is communicated with Dallas temperature using one wire library.
Dallas Temperature Library	It is the most required package for the Dallas temperature sensor. Because the communication reading of waterproof temperature is using the Dallas temperature library and one wire protocols. So, in this study, the Dallas temperature library and one wire protocols are the core package tools to read real-time data using a waterproof temperature sensor.
DFRBot –PH Library	PH Library is also the package tools to support PH sensor, during reading real-time data. Because this library contains all PH sensor function and the ability to communicate PH sensor hardware with related devices to give the full function of reading PH values.
Time Library	This library also another package tool used to control time starts and delay during the reading of PH values and Waterproof temperature values.

5.2. Water Pollution Monitoring and Prediction Model Implementation

This model prediction applied to water pollution prediction and monitoring having binary class in the target column using machine learning and python environments. We have prepared and labeled dataset by an expert; however, there are some missing values in the dataset. The sample of the dataset contains 1991 rows and 11 columns. Still, the dataset contains missing values, string, and NaN values. Pandas read this dataset, and every operation is performed using the Jupiter lab. For this model, the study applies four techniques of machine learning algorithms. Such as KNN, SVM, Linear Regression, and Random Forest Classification. After comparing these four techniques, in the end, linear regression and random forest classification are applied to predict water pollution monitoring, having the following essential steps of learning in ***figure 5.1***.

5.2.1. Best Approach for Sample Dataset Model Prediction

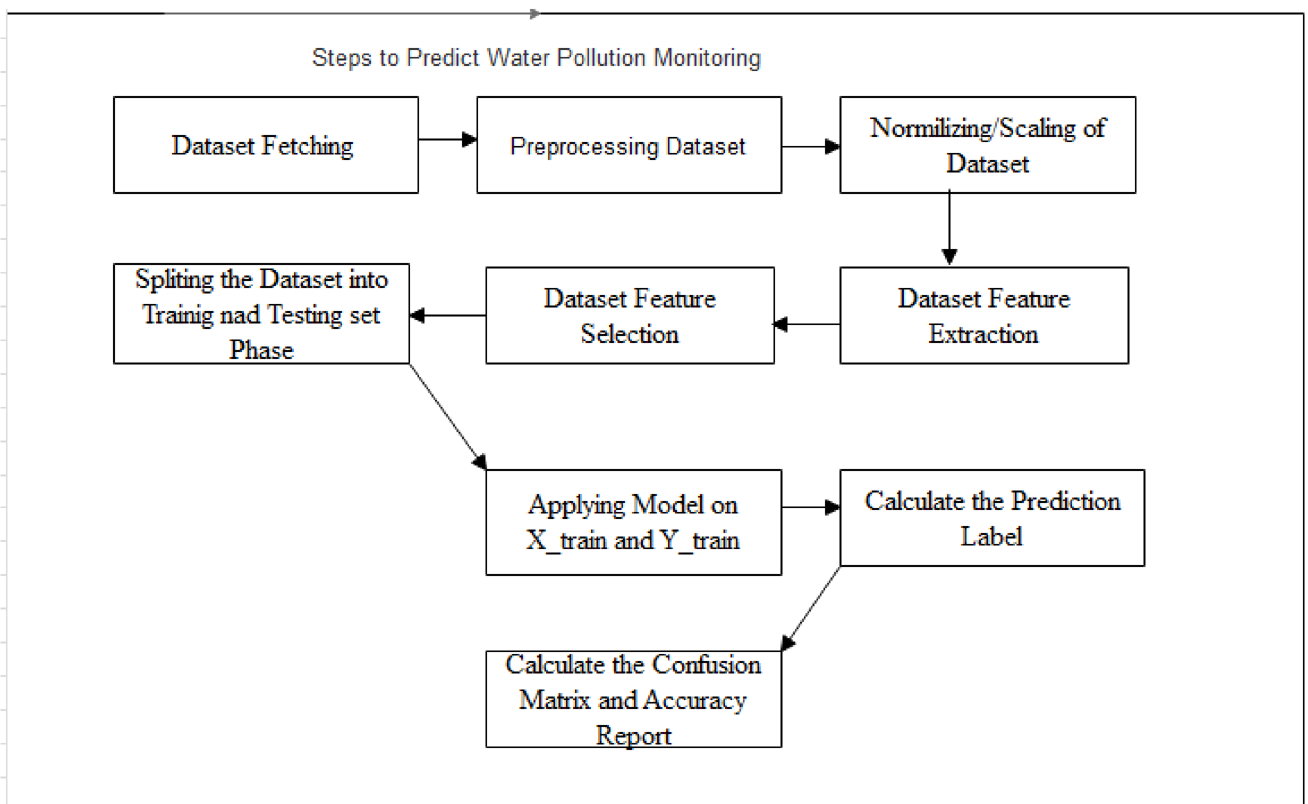

Figure 5.1: Steps Undergo for Prediction of Water Pollution Monitoring

5.2.2. Feature of Dataset

A feature identification is typically a specific representation of top row data, which is an individual measurable attribute typically depicted by a column in the dataset [69]. Consider a generic two-dimensional dataset; each observation is depicted by rows and features by rows, which can behave specific values for an observation.

Column /Feature

Rows / Observation

	Station Code	Location	State	Temp	DO(mg/l)	PH	Conductivity(µS/cm)	TDS(mg/l)	Turbidity(NTU)	Class	year
0	A2233	Assela Multi Factor	Assela	30.60	6.940049	7.5	203.0	166.67	3.0	polluted	2011
1	A2234	Assela Multi Factor	Assela	29.80	2.000000	7.2	189.0	166.68	4.0	polluted	2011
2	A2235	Assela Multi Factor	Assela	29.50	1.700000	6.9	179.0	166.69	2.0	suitable	2011
3	A2236	Assela Multi Factor	Assela	29.70	3.800000	6.9	64.0	166.70	1.0	suitable	2011
4	A2237	Assela Multi Factor	Assela	18.62	1.900000	7.3	83.0	166.71	8.0	suitable	2011

Figure 5.2: Snapshot of Feature Dataset

5.2.3. The approach of the Water Pollution Prediction and Monitoring

Step: 1 prepare dataset labeled by a domain expert and prepare a sample of the dataset for model prediction, saved the data in CSV file extension, and the CSV file dataset read using panda's library.

```
# Load data
df = pd.read_csv('water_dataX.csv',encoding='unicode_escape')
x = df.iloc[:,:].values
y = pd.DataFrame(x)
```

Figure 5.3: Snapshot for Water Pollution Dataset Reading

Step: 2 then, apply appropriate algorithms, which is used to remove some irreverent data, missing values, string values, and non-values filling to the dataset.

```
# Taking care of missing data
from sklearn.preprocessing import Imputer
imputer = Imputer(missing_values="NaN", strategy="mean", axis=0)
imputer = imputer.fit(X[:,3:9])
X[:,3:9]= imputer.transform(X[:,3:9])
y =pd.DataFrame(X)
```

Figure 5.4: Snapshot of Taking Care of Missing Values

The filled space with the values gains on the above algorithms using blew algorithms.

```
#fill missing values with mean or median
dataset_mean_imputed = dataset.fillna(dataset.mean())
dataset_median_imputed = dataset.fillna(dataset.median())
dataset_mean_imputed.head(5)
```

Figure 5.5: Snapshot for Fill Missing Values with Mean or Media

Step3: feature scaling /normalization are used to minimize the difficulty of distance among data in columns to machine understanding in a secure way. As soon as feature scaling is applied, all the dataset is bounded between [0, 1] except the target variable /dependent parameters so that the machine can understand the data in the best ways.

```
# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.fit_transform(X_test)
```

Figure 5.6: Snapshot for Feature Scaling Dataset

5.2.4. Real-Time Reading Using PH and Waterproof Temperature

This research study used PH and waterproof temperatures for real-time reading from a sample of polluted water. During reading PH and waterproof temperature, the system used Arduino Uno for communication between devices and Tera Term (window terminal) for saving real-time data on devices with excel or CSV formats. The following two figures, ***5.7*** and ***5.8,*** show Arduino Uno reading result and Tera term reading result from sensor data, ***respectively.***

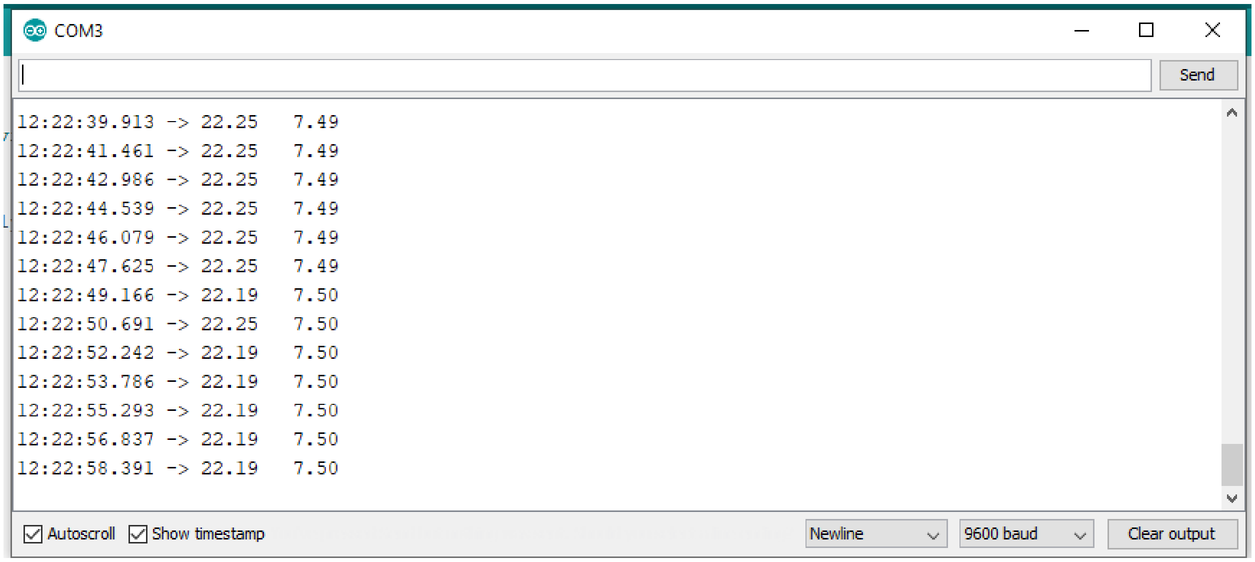

Figure 5.7: Waterproof Temperature and PH Sensor Data on Arduino IDE

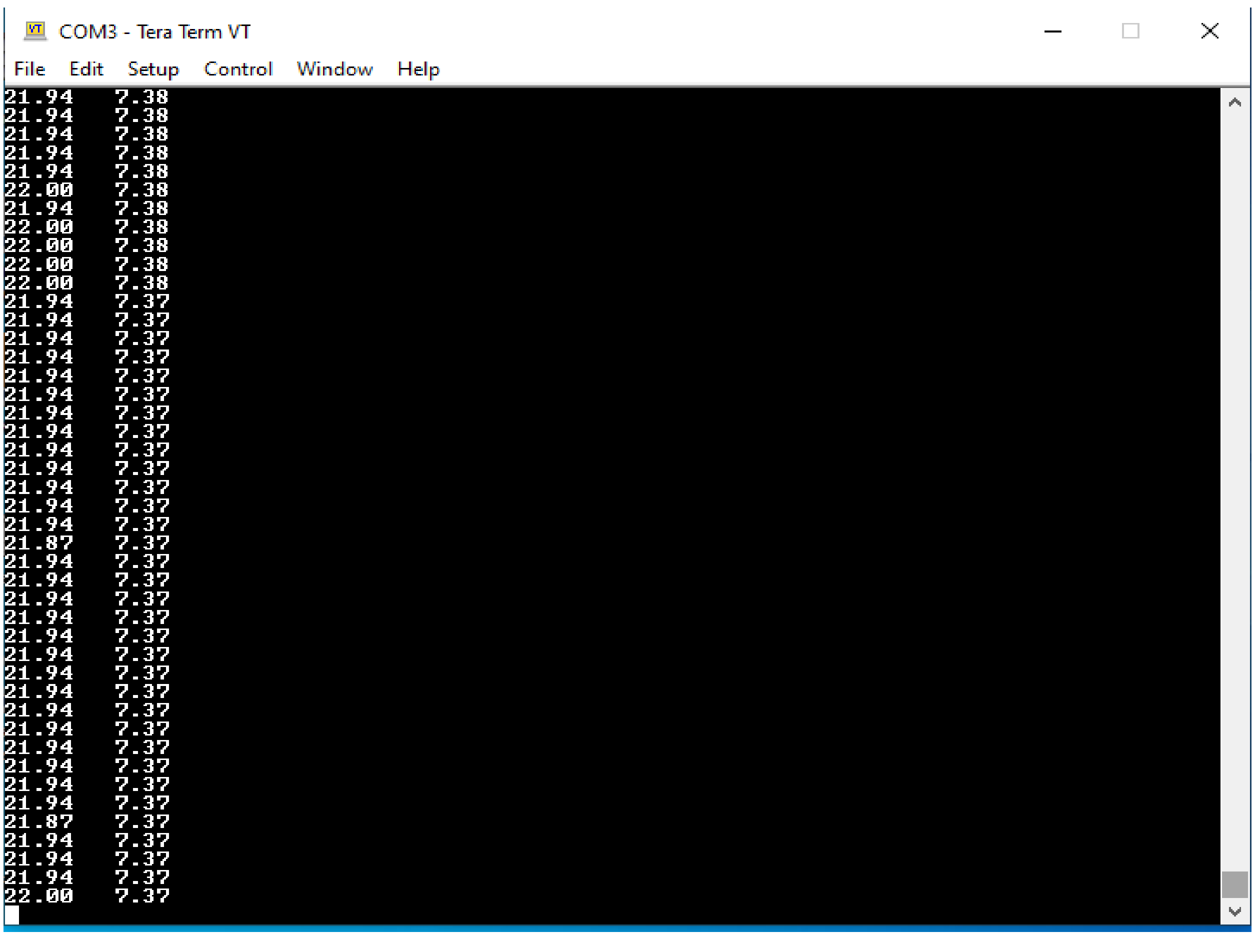

Figure 5.8: Waterproof Temperature and PH Sensor Data on Tera Term

5.3. Machine Learning Model Implementation

To build machine learning models using the proposed algorithms, we used python Sci-kit learning library package. We followed the convention method that is importing the essential library package for modeling and matrix for modeling evaluation.

In [ ]:

```
# Importing the Libraries for Modeling and Model Evaluation
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.naive_bayes import GaussianNB
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn import datasets
from sklearn import svm
from sklearn.neighbors import KNeighborsClassifier
%matplotlib inline
from sklearn import datasets, linear_model
from sklearn.model_selection import cross_validate
from sklearn.metrics.scorer import make_scorer
from sklearn.metrics import confusion_matrix
from sklearn.datasets import make_classification
from sklearn.ensemble import RandomForestClassifier
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import accuracy_score
from matplotlib import colors as mcolors
np.random.seed(0)
from sklearn.decomposition import PCA
from matplotlib.colors import ListedColormap
import seaborn
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.preprocessing import FunctionTransformer
from sklearn.preprocessing import Imputer
```

Figure 5.9: Importing the Important Package for Modeling and Model Evaluation.

After the feature selection process, the polluted and suitable sample dataset with the selected feature used as X_train training and the target class, which contains the class of polluted and suitable belongs set as Y_train. Moreover, then the classification train by using the X_train and Y_train of the entire dataset. To train the fit. () the method with an appropriate set feature for each classification instantiates the class.

```python
# Samle Feature Selection Code with Random Forest Model
start = time.process_time()
trainedforest = RandomForestClassifier(n_estimators=100).fit(X_Train,Y_Train)
print(time.process_time() - start)
predictionforest = trainedforest.predict(X_Test)
print(confusion_matrix(Y_Test,predictionforest))
print(classification_report(Y_Test,predictionforest))
```

Figure 5.10: Code for Selecting Important Features for Modeling with An RF Model.

In this study, mostly four machine learning tools used for model training and evaluation. Linear SVC () classifier of the Sklearn package used to build the SVM model. This classifier is instantiating with the selected feature to classify binary- class dataset. To implement the DT classifier, the DecisionTreeClassifier (,) method of the Sklearn ensemble package is used to train the classifier. This classifier fits several decision tree classifiers as declared in n_estimators features.

Although to implement the KNN classifiers, the KNeighbor's Classifiers () method of the Sklearn ensemble package is used to train the classifiers. The classifiers fit several KNighborsClassifiers () with n_neighbors of the feature. Moreover, finally, the most appropriate tool used for this study is a random forest model. To implement an RF classifier, the random forest classifier () method of the Sklearn ensemble package is used to train and test the classifier. This tool is the most appropriate classifier due to; it fits a number of the classifier as declared in n_estimators' parameters. Generally, the instantiating code for SVM, DT, KNN, and RF fit models are found in appendix A, respectively. Additional PH, Waterproof Temperature, and combined of both PH and waterproof temperature Sample Code Found in appendix C, D, and E, respectively.

5.4. Model Testing and Evaluation

The training that is performed is validated using both holds out and k-fold method techniques for implementing. However, k-fold CV is the most appropriate technique for implement using the Sklearn cross_val_score () and cross_val_predictcs method using K values Ten due to K of K-fold classification is recommended. These methods use the models, and the dataset is divided into ten different sets and nine sets to train, and one set used to test models each 1 to K iteration.

```
In [152]:
from sklearn.model_selection import cross_val_score
import numpy as np

#create a new KNN model
rf_cv = RandomForestClassifier(n_estimators = 100, criterion = 'entropy', random_state = 50)

#train model with cv of 10
cv_scores = cross_val_score(rf_cv, X_test, Y_test, cv=10)


#print each cv score (accuracy) and average them
print(cv_scores)

```

Figure 5.11: Performing 10-Fold CV on Models.

The study used a classification report (), accuracy score (), and confussion_matrix () methods, which are used to evaluate the performance of the built models using predict the values of 10-fold CV for the entire dataset. These methods give a result of evaluation- matrices such as accuracy, precision, recall, and F1-score value of the models under testing.

Finally, A prototype for predict water pollution is implemented using the selected best model then deployed using flask web service for it to be able to accept real-time data and then return prediction the result.

CHAPTER SIX

6. Evaluation, Result, and Discussion

In this chapter, the result of the experiment of the proposed solution for polluted water prediction and monitoring using machine learning techniques and Sensors based are discussed. Here the result of polluted water prediction and monitoring process and the experiment to build the classifications models for prediction based on the binary class dataset is discussed. Finally, we discuss the significance of the result obtained by the experiment in this study.

6.1. Dataset Analyses and Exploratory Result

To classify water as polluted and suitable, water polluted to be analyzed; the researcher utilizes a simple random sampling technique. This technique gave an equal chance for all the selected features and extracted features of polluted and suitable to be analyzed. We then decide to analyze the 1990 sample of water pollution datasets with six features. The size is limited to the 1990 sample of dataset due to data security for malt Factory Company, time for this research, and resource for the data gathering, expert for the data labeling process for the collected data.

The analyses were given all 1990 samples of the dataset to be labeled using WHO and FAO guidelines with the support domain experts. The most important feature was given to calculate the factor of water pollution by domain experts on the sample of datasets. The whole process of labeling the dataset was tedious, challenging, and time-consuming due to the labeling is manually ways of mechanisms. The researcher decides with a domain expert; the final class classification is using the standard thresh hold method for water pollution identifications — the result of two class distribution in dataset showing in table 6.1. The dataset used to train the machine learning algorithms consisting of 1393 unique analyzing training dataset and 597 unique analyzing for a testing dataset (using hold out method) decided by majority researchers. A total of 1990 sample of the dataset for given polluted water is in the analyzed dataset.

Table 6.1.: The Two-Class Distribution of the Dataset

Labels or Class	***Number of Sample Dataset***
Polluted (1)	1393
Suitable (2)	598
Total = 1991	

The hold out the method used to split data into 70% train and 30% test dataset respectively. Then, we get a 1393 sample of train datasets and 597 samples of the test dataset shown in table 6.1 above. In hand, the binary classification of distributed dataset showed in table 6.2 blows with its percentage of classification.

Table 6.2: The Binary Class Distribution of the Dataset

Label or Class	***Number of Sample Dataset***	***Percentage***
Polluted (1)	1187	59.62%
Suitable (2)	804	40.38%
Total	1991	100%

Finally, the analysis result shows that the dataset contains some miss-classification because of some related weighted of features confused for classification. However, having a guideline that specifies how to analyze the two classes as polluted or suitable classes. The result is shown in figure 6.1 that a moderate and right agreement of analysis for binary classification datasets.

```
1    1187
2     804
Name: Class, dtype: int64
Percentage of polluted water: 59.62
Percentage of suitable water: 40.38
```

Figure 6.1: Binary Classification Distributed Result.

6.2. Feature Analyzes Result

Each feature is analyzed to target variables for classification based on weight mean value. In this study, all six features analyzed with target variables and the vital feature selected for model prediction and monitoring. The below figure 6.2 – 6.5 shows each feature analysis result with target variable classes. This method is called a feature analysis method.

```
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x1915c25f1c8>
```

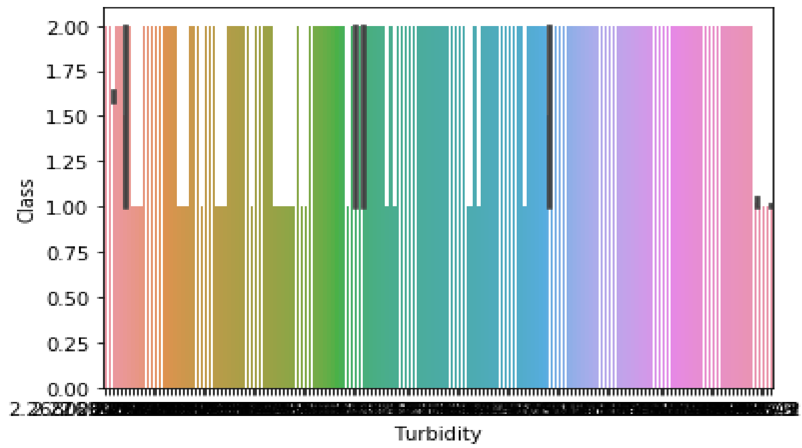

Figure 6.2: Turbidity Analysis Result with Target Class

So, as a result, indicated in figure 6.2, turbidity has a more correlated relation with the target variables.

```
Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x1915c8df8c8>
```

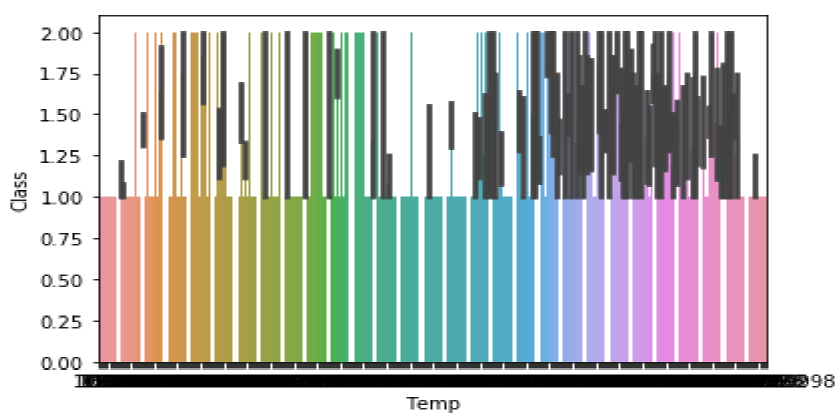

Figure 6.3: Temperature Analysis Result with the Target Variable

Here, the temperature is one of the most features for model building and prediction, because the analysis result shown in the above figure 6.3, the black color parts represent the factor of temperature on water pollution.

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x1915cc3f5c8>

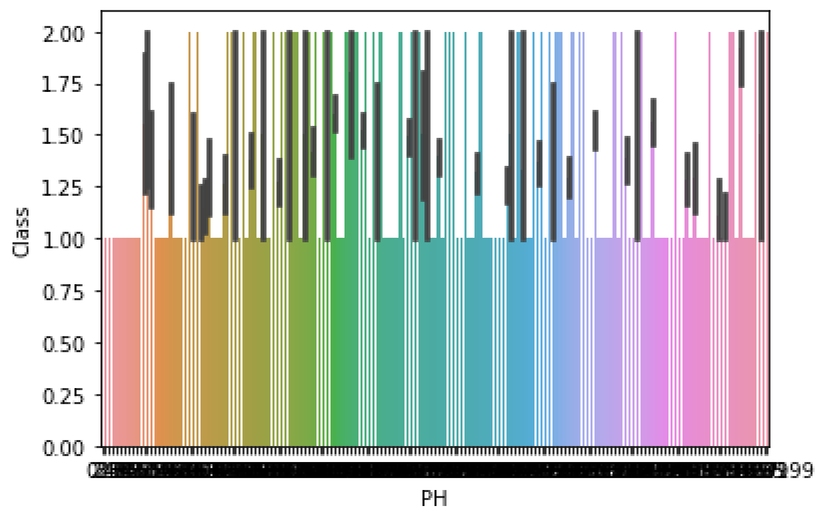

Figure 6.4: PH Analysis Result with Target Variable

Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x1915d9369c8>

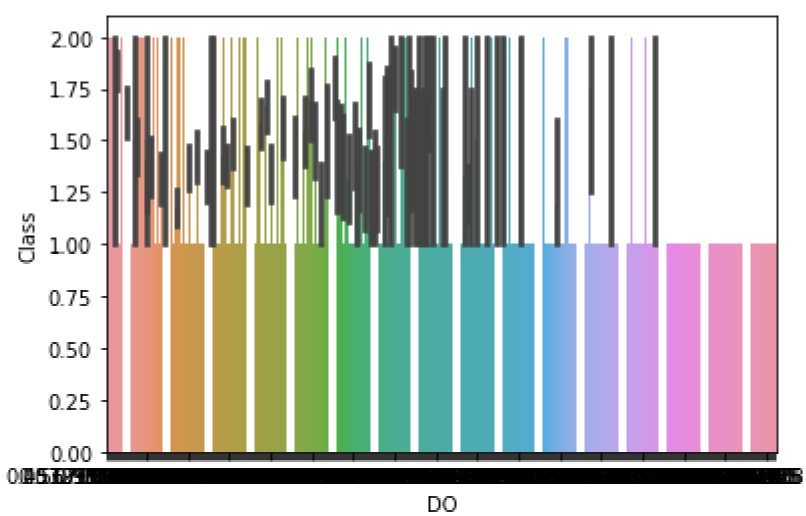

Figure 6.5: Dissolve Oxygen Analysis Result with Target Variable

In general, data exploratory analysis used for handling missing values, testing variables, predicting variables, identifying important factors, and the relation among features and target variables.

6.3. Binary Classification Models Evaluation and Evaluation Result

To evaluate the performance of the machine learning classifier, K-fold cross-validation is usually used. The right metric for this problem is the F1 score since it combines precision and recall. Therefore, for each classifier that has been used, we apply K-fold cross-validation with K = 10, i.e., the dataset is categorized ten times into different sets. This way, each time we use 90% of the data for training and 10% for testing.

These classification models are built using the two-class dataset that means the target class is polluted and suitable. The trained models are tested using the 10-fold CV. The result of each test accuracy score for DT, KNN, SVM, and RF models based on feature selection are presented in tables separated below;

Table 6.3: DT Models Accuracy for Each Feature Selection Using Binary Class Dataset.

Features	***10-fold accuracy (%) for DT models***										**Average Accuracy (%)**
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
RFE	96.7	95.0	95.0	96.6	1.00	96.6	96.6	96.6	96.6	94.9	96.4
UFS	93.4	93.4	96.6	96.6	96.6	96.6	94.9	1.00	98.3	94.9	96.1
LR	95.0	93.4	95.0	95.0	1.00	94.9	94.9	93.2	93.2	94.9	94.9

The result in table 6.3 shows the CV accuracy scores for the DT models based on feature selection with corresponding each fold test. The lower average accuracy 94.9% result recorded on LR the feature selection model, and the higher accuracy 96.4% resulted using RFE feature selection.

Table 6.4: SVM Models Accuracy Score on Feature Selections Using Binary Class Dataset

Features	***10-Fold Accuracy (%) For SVM Models***										***Average accuracy (%)***
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
RFE	90.1	83.6	78.6	83.3	85.0	84.7	88.1	88.1	89.8	84.7	85.6
UFS	85.2	83.6	78.3	85.0	86.6	85.0	81.3	84.7	91.5	77.9	83.9
LR	91.8	85.2	78.6	85.0	86.6	84.7	86.4	88.1	89.8	86.4	86.2

The results in table 6.4 show the prediction accuracy scores for the SVM models on each feature selection with corresponding each fold test. The lower average accuracy of 83% result recorded on the UFS feature selection model and slightly higher accuracy of 86% recorded using LR feature selections.

Table 6.5: KNN Models Accuracy Scores on Each Feature Selection Using Binary Class Dataset

Features	***10-Fold Accuracy (%) For KNN Models***										***Average accuracy (%)***
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
RFE	59.0	59.1	59.0	58.4	58.3	59.3	59.3	59.4	59.4	59.0	60.1
UFS	88.5	86.8	88.3	90.0	91.6	93.3	89.8	88.1	93.2	77.9	88.7
LR	59.0	59.0	59.0	58.3	58.3	59.3	59.3	59.1	59.3	59.3	59.0

The result in table 6.5 shows the prediction accuracy for the KNN models on the selected feature with corresponding each fold test.

The lower average of 59% result recorded on the LR feature selections. The higher accuracy of 88% recorded model using UFS feature selections.

Table 6.6: RF Models Accuracy Scores on Each Feature Selection Using Binary Class Dataset.

Features	***10-Fold Accuracy (%) For RF Models***										***Average accuracy (%)***
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
RFE	96.7	95.0	98.3	1.00	96.6	98.3	96.6	98.0	98.3	98.3	97.6
UFS	95.0	95.0	96.6	98.3	1.00	96.6	96.6	1.00	96.6	98.3	97.3
LR	98.3	96.7	95.0	96.6	98.3	98.3	96.6	98.3	94.9	1.00	97.3

The result in table 6.6 shows the prediction accuracy scores for the RF models on the selected features with corresponding each fold test. The lower average accuracy of 97.3% results recorded both on UFS and LR feature selections. The higher accuracy of 97.6 recorded model using RFE feature selections. The 10-fold CV models testing for each selected feature result for binary classification are summarized and illustrated in the graph below. The graph in figure 6.8 shows the visual representation of the CV average accuracy of each model on the selected feature in the above four tables for binary class experiments. The navy blue represents the average classification of KNN models, the red-orange represents the average classification of SVM models, the forest green represents the average accuracy classification of DT models, and the red-orange represents the average accuracy classification of RF models. The X-axis represents the selected feature accuracy, and Y-axis is based on the average results. These show the average accuracy of the RF models slightly higher than those of the DT, KNN, and SVM models based on both RFE and UFS feature selection methods.

```
The accuracy score achieved using K-Nearest Neighbors is: 95.65 %
The accuracy score achieved using Support Vector Machine is: 91.3 %
The accuracy score achieved using Decision Tree is: 95.82 %
The accuracy score achieved using Random Forest is: 97.66 %

Out[357]: <matplotlib.axes._subplots.AxesSubplot at 0x1916edcaa08>
```

Figure 6.6: Binary Models Comparisons Using CV Avg Accuracy.

Besides, the result accuracy scored obtained by Tenfold CV; the study also used other model performance evaluation matrix, as discussed in chapters three and four. These metrics are precision (p), Recall(R), and F-score (F). In table 6.7 shows the results of evaluation metrics for each model based on feature selections.

Table 6.7: Classification Performance Result of Binary Models.

Features	***SVM***			***KNN***			***DT***			***RF***		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1

RFE	89	86	87	89	86	86	89	86	87	98	98	98
UFS	88	85	85	94	94	94	94	94	94	97	97	97
LR	87	83	83	90	88	88	96	96	96	98	98	98

Base on the F1-score, the RF models on both RFE and LR, obtained a higher score of 98% than SVM, KNN, and DT models. However, the accuracy of the RF models is 97.6% higher than SVM, KNN, and DT using RFE feature selection see in table 6.6. F1-score metrics selected to compare the models based on the feature selection, because it is more useful than accuracy when the dataset contains even class distributions. Finally, the result of binary class models demonstrates the RF models based on both RFE and LR models show better accuracy than SVM, KKN, and DT. Also, RF models based on RFE shows higher accuracy than all three models and give a better classification result than three models.

6.3.1. A Framework Use Case Evaluation

To evaluate and identify the right use case for the framework in this study, an evaluation is done based on A Framework for water quality monitoring [70]. Figure 6.7 illustrates the use case and steps used to evaluate the proposed framework of this study. As the evaluation framework shows clearly, using IoT with machine learning techniques in polluted water classification and prediction is an optimal solution. The applicability of the use case is analyzed and assessed against the framework core characterized.

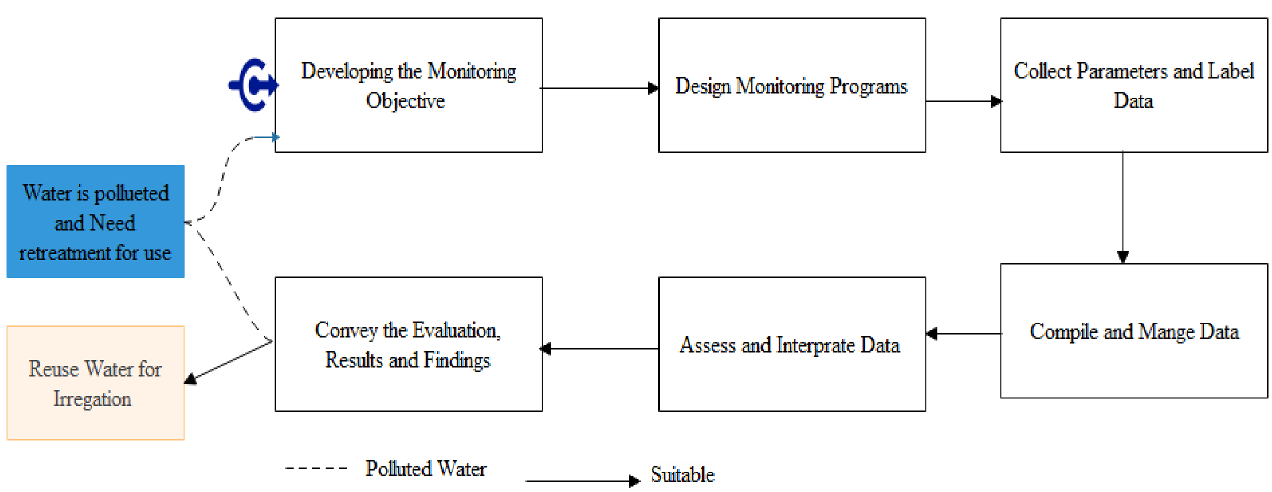

Figure 6.7: Used Use Case for Evaluation Framework.

This unique framework of the national water quality, monitoring council process is based on communicating, coordinating, and collaborating systems.

Developing Monitoring Objective: developing a monitoring objective is the first step to develop a water quality monitoring system based on reviewing information goals to respond to specific water resources and scarcity of water quality management.

Designing Monitoring Program: developing a monitoring program is the second procedure support to achieve the monitoring objective or goals. Under this step, the specific area for study, data collection procedure, tools, and method of data measurement are identified.

Collect Fields and Laboratory Data: the collect parameter stage used to translate water properties into quantitive data that distinguish water states with the help of laboratory testing.

Compile and Manage Data: under this, based on measured data, the system suggests users use waters and do not to use if the water is polluted.

Assesses and Interpret Data: at these points, having all the above information and collected data, the system achieves the monitoring objective with the support of analysis methodology.

Convey Finding and Evaluate Program: finally, the developed framework gives functions for stakeholders, and the framework itself evaluated based on its functions. Generally, the strength of the monitoring framework is that emphasis feedback in every stage with the support of communicating, coordinate, and collaborate system.

6.4. Discussion

Both of IoT and machine learning techniques are one of the special keys for predicting and monitoring water pollution. According to the contribution obtained from this research, IoT based mechanism can provide to read real-time data and the weight of different variables; which must be need association with other techniques to reduce cost, such as machine learning comparison, to make complete analysis for water parameters in polluted water; this result is justified and evaluated with interesting performance indicators. The two best evaluation indicators of machine learning techniques in this study are the HOLD OUT method and the K-FOLD method. Among them, the K-FOLD method is the most interesting performance indicators based on precision, recall, and F1 scores with total mean accuracy. For the work on the evaluation technique comparisons, the solution helps the user to get a more accurate classification and avoid Mis-classifications. On the other hand, IoT based solution helps the user to get real-time data for model predictions. They are around four to five machine learning algorithms applied to identify the most accurate algorithms for classification, avoid Mis-classification, prediction of water pollutions. Random forest algorithms are the most important and accurate performance for the proposed systems because of all requests based on both polluted and suitable data combined to develop classifiers. Because based on this study, data shows that random forest has better classification accuracy. The first experiment of the proposed study shows that the random forest better classification algorithms compared with logistic regression, KNN, DT, and SVM. Mostly, this study experiment with three different feature selection (RFE, UFS, and LR) and machine learning algorithms SVM, KNN, DT, and RF to model polluted water classification and predictions using dataset evaluated by 10-fold CV.

Feature selection methods are an essential process to select the best parameters based on the weight from all the datasets to model using a machine learning tool. The result of this study shows that RFE performs better accuracy than UFS and equal to LR for binary classification models. In this thesis work, three research questions were answered by using both IoT and machine learning techniques. ***In our first research question,*** we get real-time data using the sensor device through Arduino Uno and IDE; Tera Term (window terminal) to save real-time data on excels files and sends it to the model. However, the real-time data cannot directly send to models for prediction. Because as machine learning algorithms operate for prediction and classification, and the sensor device read real-time data continuously. So, the system may crash. In this study, the real-time data saved, then send to models for prediction, and the result is reported whether water is polluted or suitable for irrigations. ***In the second question,*** the best tool in water pollution classification and prediction answered, that discussed in chapter five tables, 5.1 details. ***In the third question***, we are quantifying the quality of classification and prediction tools in order to compare the classifier and prediction algorithms. ***In the last question,*** the performance of the model is around 97.6 and improved using model and framework evaluation systems. Machine learning classifier algorithms have different classification and prediction performance. The classification accuracy, one type of metric for classifier performance, was calculated for four different classifiers such as; SVM, KNN, DT, and RF. Among those algorithms, RF has better classification accuracy than SVM, KNN, and DT.

CHAPTER SEVEN

7. Conclusion, Recommendation and Future Work

In this chapter, the proposed of polluted water monitoring and prediction using machine learning techniques and IoT based is the finding of this research is concluded. Also, recommendation and future work of the research direction is described in this chapter.

7.1. Conclusion

This research study proposed to develop a solution to water pollution monitoring and predicting using IoT and machine learning techniques. The study conducts to develop, implement, and compare appropriate machine learning model, and essential feature selection methods specifically for polluted water prediction.

To successfully execute the study, it was essential to understand and define the effect of polluted water on human life and plantation, explore existing various techniques used to tackle the problems and understand the polluted water risk, as discussed in chapter two. Moreover, they are different method followed to implement (using python language) and designed frameworks that have the ability for water pollution prediction and monitoring. These methods include collecting sample datasets for labeling and build the standard dataset, developing the standard guideline, preprocessing, feature selection using; recursive feature elimination, Univariate feature selection, and lasso regression; also, the most training model used were; SVM, KNN, DT, and RF models testing. Finally, perform model's comparison based 10-fold CV evaluation matrices.

In this research study, the domain expert labels polluted water datasets into the binary class of polluted (p) and suitable(s) water for reuse in irrigation. The labeled dataset converts into two classes based on all sample datasets and classified each as polluted and suitable. This result dataset with 1990 instance and six feature s converted into two class datasets. Based on this sample dataset and real-collecting data (using PH sensor and waterproof temperature sensors with the help of Arduino Uno communication between window terminal and sensors), the model develops using SVM, KNN, DT, and RF with three feature selection method implemented. The experiment performs using dataset resulted in 12 models to identifies the most and best models based on the accuracy and capability of prediction models.

The binary models using RF with recursive feature eliminations resulted in better accuracy than SVM, KNN, and DT. Moreover, RF with RFE and LR perform better in classification result with 98% F-1 scores, 97% RF model accuracy, and it shows better performance than the models based on SVM, KNN, and DT. Finally, the models based on RF using RFE slightly better performance than SVM, KNN, and DT in this research study. Finally, the developed framework successfully classifies and predicts the polluted water class into polluted and suitable.

7.2. Recommendations

Since the prediction and classification models needed to show some capability for polluted water using a minimal dataset with a lot of vague labeling domain experts of a sample dataset. As the results, the researcher recommended that educated the best expert on database specialization that support to stores and labels data with WHO and FOA, and environmental condition experts. Finally, we recommend adopting the behaviors of testing west water during reusing for irrigation.

Also, the researcher recommended that the university must have to enhance the relation or linkage with the different industrial companies to solve the problem of researcher scarcity with the dataset.

7.3. Future Work

The proposed solution limited to a supervised learning algorithm with the sample, sample dataset, and real data using a feature selection method to build model prediction. So, it is better to see the difference in performance results using unsupervised or deep learning algorithms model. Also, the researcher recommends on-line real-time water pollution prediction having a database on the server.

REFERENCES

[1] J. Z. G. X. Gaganjot Kaur Kang, "Data-driven Water Quality Analysis and Prediction: A Survey," Data-driven Water Qual. Anal. Predict. A Surv., vol. At: San Fr, no. Conference: The 3rd IEEE International Conference on BIG DATA COMPUTING SERVICE AND APPLICATIONS, p. pp 224-232, 2017.

[2] FIREW GETACHEW MARCH, "ASSESSMENT OF PHYSICAL QUALITY OF WASTEWATER AND ITS," vol 1, pp. 1-80, 2017.

[3] R. Rosly *et al.*, "The Study on the Accuracy of Classifiers for Water Quality Application The Study on the Accuracy of Classifiers for Water Quality Application," vol 1, pp. 1-9. January 2016, 2015.

[4] B. Paul, "Sensor-Based Water Quality Monitoring System," BRAC University, vol 1, pp 1-33, 2017.

[5] P. Vic Shayne, "Can we separate the health of our planet & the health of our bodies?" https://www.elephantjournal.com, 2011. [Online]. Available: https://www.elephantjournal.com/2011/06/can-we-separate-the-health-of-our-planet-from-the-health-of-our-bodies-vic-shayne-phd/.

[6] J. W. S. Longhurst, J. Barnes, and C. A. Edokpayi, "Towards a new framework for air quality management in Nigeria," vol 3, pp. 1-7, May 2017, 2011.

[7] UNESCO-IHE, "Design of a Water Water Quality Monitoring Network for the Limpopo Basin, Mozambique,", vol 1, pp. 1-4, 2007.

[8] UNDESA, "International decade for Action 'water for life,'" no.3, vol 3, pp.1-13 March 2014.

[9] By Inyinbor Adejumoke A., Adebesin Babatunde O., Oluyori Abimbola P., Adelani-Akande Tabitha A., Dada Adewumi O., and Oreofe Toyin A. "Water Pollution: Effects, Prevention, and Climatic Impact," Water Challenges an Urban. World, 2018.

[10] JRC, Minimum quality requirements for water reuse in agricultural irrigation and aquifer recharge, vol 1, pp.1-5, 2017.

[11] S. K. Sanyal and B. P. Bhatt, “Water Pollution and Agriculture,” vol 1, pp.1-9, January 2015.

[12] J. Jeffrey Peirce P Aarne Vesilind Ruth Weiner, “Environmental Pollution and Control,” https://www.elsevier.com, 1990. [Online]. Available: https://www.elsevier.com/books/environmental-pollution-and-control/vesilind/978-0-409-90272-3. [Accessed: 26-Jan-2020].

[13] P. Monitoring and A. Branch, “WATER QUALITY MONITORING PROGRAMME Objectives and Structure of Presentation : • Present Background and Objectives of Water Quality Monitoring Programme • Define water quality and rationale for monitoring • Water quality parameters • Highlight NEPA ‘ s role,”. March 2014.

[14] J. Mateo-Sagasta, S. Marjani, H. Turral, and J. Burke, “Water pollution from agriculture: a global review,” FAO y IWMI, vol. 1, p. 35, 2017.

[15] M. Chilundo, P. Kelderman, and J. H. Ókeeffe, “Design of a water quality monitoring network for the Limpopo River Basin in Mozambique,” Phys. Chem. Earth, vol. 33, no. 8–13, pp. 655–665, 2008.

[16] “Turbidity,” https://www.lenntech.com. [Online]. Available: https://www.lenntech.com/turbidity.htm.

[17] T. P. Lambrou, C. C. Anastasiou, and C. G. Panayiotou, “A nephelometric turbidity system for monitoring residential drinking water quality,” Lect. Notes Inst. Comput. *Sci. Soc.* Telecommun. Eng., vol. 29 LNICST, pp. 43–55, 2010.

[18] H. J. K., “Irrigation Water Quality Standards for Indirect Wastewater Reuse in Agriculture,” https://www.mdpi.com, 2016. [Online]. Available: https://www.mdpi.com/2073-4441/8/4/169/htm.

[19] Microverse, “What is the difference between total dissolved solids and total suspended solids,” https://www.answers.com, 2010. [Online]. Available: https://www.answers.com/Q/What_is_the_difference_between_total_dissolved_solids_and_total_suspended_solids.

[20] L. J. Zhao, D. C. Yuan, T. Y. Chai, and J. Tang, “Procedia Engineering KPCA and ELM ensemble modeling of wastewater effluent quality indices,” vol. 15, pp. 5558–5562, 2011.

[21] L. J. Zhao, D. C. Yuan, T. Y. Chai, and J. Tang, “KPCA and ELM ensemble modeling of wastewater effluent quality indices,” Procedia Eng., vol. 15, pp. 5558–5562, 2011.

[22] B. Paul, “Sensor-Based Water Quality Monitoring System,” BRAC University, vol 1, pp. 1-33, 2017.

[23] AIOTI, “Research and Innovation Priorities for IoT. Industrial, Business and Consumer Solutions,” vol 1, pp 1-13, August 2018.

[24] R. Ramya, “The Real-Time Monitoring of Water Quality in,” 2015 Int. Conf. Circuits, Power Comput. Technol. [ICCPCT-2015], pp. 1–4, 2015.

[25] K. Ashwini, J. J. Vedha, and M. D. Priya, “Intelligent model for predicting water quality,” vol. 5, no. 2, pp. 70–75, 2012.

[26] A. Awoke, A. Beyene, H. Kloos, P. L. M. Goethals, and L. Triest, “River Water Pollution Status and Water Policy Scenario in Ethiopia : Raising Awareness for Better Implementation in Developing Countries River Water Pollution Status and Water Policy Scenario in Ethiopia : Raising Awareness for Better Implementation in De,” Environ. Manage., vol. Vol 2, June, pp. 1–14, 2016.

[27] B. Garoma, G. Kenasa, and M. Jida, “Shambu Drinking water quality Drinking Water Quality Test of Shambu Town (Ethiopia) from Source to Household Taps Using Some Physico-chemical and Biological Parameters Ethiopian Institute of Biotechnology, Addis Ababa, Ethiopia,” J. Ecol. Environ. Sci., vol. Vol 6, December, pp. 1–8, 2018.

[28] T. Karthick, G. Dutt, T. S. Kohli, and S. Pandey, “Prediction of Water Quality and Smart Water Quality Monitoring System in IoT Environment,” vol. 118, no. 20, pp. 3969–3977, 2018.

[29] A. Mathematics, “Air and Water Quality Monitoring Through IoT By Using,” vol. 118, no. 22, pp. 51–55, 2018.

[30] N. K. Gakungu, an N. Gitau, B. N. K. Njoroge, and M. W. Kimani, “Solid waste management in Kenya: A case study of Public Technical Training Institutions,” ICASTOR J. Eng., vol. 5, no. 3, pp. 127–138, 2012.

[31] T. Yu, S. Yang, Y. Bai, X. Gao, and C. Li, “Inlet water quality forecasting of wastewater treatment based on kernel principal component analysis and an extreme learning machine,” Water (Switzerland), vol. 10, no. 7, 2018.

[32] R. F. Maheder, S. Teklu, K. Awoke, T. K. Feto, and K. D. Tullu, “Virological Quality of Urban Rivers and Hospitals Wastewaters in Addis Ababa, Abstract :” Open Microbiol. J., vol. 13, pp. 164–170, 2019.

[33] J. E. A. Toxicol, M. Town, A. E. Gebre, H. F. Demissie, S. T. Mengesha, and M. T. Segni, “Environmental & Analytical Toxicology The Pollution Profile of Modjo River Due to Industrial Wastewater,” Environ. Anal. Toxicol., vol. 6, no. 3, pp. 1–5, 2016.

[34] R. J. Mckee and J. L. Bells, “International Journal of Economics and Sociology,” vol. 4, no. 1, pp. 1–27, 2013.

[35] V. S. Kale, “Consequence of Temperature, pH, Turbidity and Dissolved Oxygen Water Quality Parameters.http://www.env.gov.bc.ca/wat/wq/BCguidelines/temptech/temperature.html#rec,” Int. Adv. Res. J. Sci. Eng. Technol. ISO, vol. 3, no. 8, pp. 186–190, 2016.

[36] S. B. Kotsiantis, “Erratum: Feature selection for machine learning classification problems: A recent overview (Artificial Intelligence Review (2011)),” Artif. Intell. Rev., vol. 42, no. 1, p. 157, 2014.

[37] H. Polat, H. Danaei Mehr, and A. Cetin, “Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods,” J. Med. Syst., vol. 41, no. 4, 2017.

[38] H. Fangohr, “Introduction to Python for Computational Science and Engineering (A beginner ’ s guide),” 2015.

[39] M. Grinberg, Flask Web Development_ Developing Web Applications with Python [Grinberg 2014-05-18]. 2014.

[40] D. JIAO and X. LUO, “Water Quality Monitoring System Based on LoRa,” DEStech Trans. Comput. Sci. Eng., vol. 10,. ie, pp. 1107–1116, 2018.

[41] DFRobot, “PH meter(SKU: SEN0161),” wiki.dfrobot.com, 2017. [Online]. Available: https://wiki.dfrobot.com/PH_meter_SKU__SEN0161_.

[42] V. Madhavireddy and B. Koteswarrao, “Smart Water Quality Monitoring System Using IoT Technology,” Int. J. Eng. Technol., vol. 7, no. 4.36, p. 636, 2018.

[43] H. Shan, “DS18B20 Waterproof Temperature Sensor Cable,” www.terraelectronica.ru, 2017. [Online]. Available: https://www.terraelectronica.ru/pdf/show?pdf_file=%2Fz%2FDatasheet%2F1%2F1420644897.pdf.

[44] Arduino Uno official, “Arduino Uno Rev3,” Store.Arduino.*Cc*, p. 1, 2018.

[45] C. C. By-Sadark, “Introduction to Arduino // Basic Arduino reference sheet : Support : Forums ,” 2012.

[46] T. Oladipupo, “Types of Machine Learning Algorithms,” New Adv. Mach. Learn., 2010.

[47] N. Abhishek, Reinforcement learning with Open AI, TensorFlow and Keras Using Python, vol. 3, no. 9. 2012.

[48] A. Shkurin, “WATER QUALITY ANALYSIS USING MACHINE LEARNING ALGORITHMS Environmental Engineering Water Quality Analysis Using Machine Learning Algorithms,”. December 2015.

[49] N. Settouti, M. E. A. Bechar, and M. A. Chikh, “Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classi cation Task,” Int. J. Interact. Multimed. Artif. Intell., vol. 4, no. 1, p. 46, 2016.

[50] “Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems,” Water Resour. Manag., vol. 31, no. 5, pp. 1575–1589, 2017.

[51] Y. Masutani, M. Nemoto, Y. Nomura, and N. Hayashi, Clinical Machine Learning in Action. 2012.

[52] A. R. Feinstein, “Simple Linear Regression,” Multivariable Anal., pp. 128–163, 2018.

[53] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, and R. Irfan, “Efficient Water Quality Prediction Using Supervised,” pp. 1–14, 2019.

[54] I. Pak and P. Lee Teher, “Machine Learning Classifiers: Evaluation of the Performance in

Online Reviews," Indian J. Sci. Technol., vol. 9, no. 45, 2016.

[55] J. Brownlee, "What is a Confusion Matrix in Machine Learning," https://machinelearningmastery.com, 2019. [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/.

[56] C. Manliguez, "Generalized Confusion Matrix for Multiple Classes," no. November, pp. 4–6, 2016.

[57] K. Mwemezi and A. Sam, "Development of Innovative Secured Remote Sensor Water Quality Monitoring & Management System: Case of Pangani Water Basin," Int. J. Eng. Manuf., vol. 9, no. 1, pp. 47–63, 2019.

[58] J. Li and L. Li, "Comparative research on Python speed optimization strategies," Proc. - 2010 Int. Conf. Intell. Comput. Integer. Syst. ICISS2010, pp. 57–59, 2010.

[59] J.-M. Chery, "STMicroelectronics Reports 2019 Second Quarter Financial Results," www.globenewswire, 2016. [Online]. Available: https://www.globenewswire.com/news-release/2019/07/25/1887706/0/en/STMicroelectronics-Reports-2019-Second-Quarter-Financial-Results.html. [Accessed: 20-Dec-2019].

[60] G. P. Berhad, "Sustainability Report Sustainability Report," pp. 27–35, 2013.

[61] T. Pc-engine, U. States, O. G. Kit, and P. Corporation, "BASH Basics : Introduction to BASH Linux Gaming : PC-Engine / TurboGrafx Object Tracking Using oCam and ODROID-XU4 : An Easy Step-By-Step Guide To Object Tracking ODROID-GO Game Kit : A Handheld Gaming Console To Celebrate Hardkernel ' s 10th Anniversary ."

[62] G. Halfacree, "The Official Raspberry Pi Beginner 's Guide," Raspberry Pi Press, 2018.

[63] A. Mishra, "Amazon Machine Learning," Mach. Learn. AWS Cloud, pp. 317–351, 2019.

[64] B. Gregor, "Introduction to Python Part 1 RCS Team and Expertise.," www.bu.edu/tech/files/2018/05/Intro-to-Python-Part-1, 2016. [Online]. Available: https://www.bu.edu/tech/files/2018/05/Intro-to-Python-Part-1.

[65] "The JupyterLab Interface," https://jupyterlab.readthedocs.io. [Online]. Available: https://jupyterlab.readthedocs.io/en/stable/user/interface.html.

[66] C. R. Mora, "JupyterLab Documentation," codingclubuc3m, 2019. [Online]. Available: https://codingclubuc3m.rbind.io/post/2019-05-08/. [Accessed: 19-Dec-2020].

[67] "Arduino Tutorial," https://www.tutorialspoint.com. [Online]. Available: https://www.tutorialspoint.com/arduino/index.htm.

[68] "sklearn," https://pypi.org, 2015. [Online]. Available: https://pypi.org/project/sklearn/.

[69] D. (DJ) Sarkar, "Continuous Numeric Data," https://towardsdatascience.com, 2018. [Online]. Available: https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b.

[70] C. Spooner, "A Framework for Water Quality Monitoring," 2014. [Online]. Available: http://water.usgs.gov/wicp/acwi/monitoring/.

APPENDIXES

Appendix A: The Instantiating Code for SVM, DT, KNN, AND RF, Respectively.

SVM

In [36]:

```
from sklearn import svm

sv = svm.SVC(kernel='linear')

sv.fit(X_train, Y_train)

Y_pred_svm = sv.predict(X_test)
```

Decision Tree

In [42]:

```
from sklearn.tree import DecisionTreeClassifier

max_accuracy = 0


for x in range(200):
    dt = DecisionTreeClassifier(random_state=x)
    dt.fit(X_train,Y_train)
    Y_pred_dt = dt.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_dt,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)


dt = DecisionTreeClassifier(random_state=best_x)
dt.fit(X_train,Y_train)
Y_pred_dt = dt.predict(X_test)
```

K Nearest Neighbors

In [39]:

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(X_train,Y_train)
Y_pred_knn=knn.predict(X_test)
```

Random Forest

In [125]:

```
from sklearn.ensemble import RandomForestClassifier

max_accuracy = 0


for x in range(200):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)

rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)
```

Appendix B: Final Function of Developed System

Finally, the study analyzed the result that gains from the implementation. Also, the study compares the performance of each algorithm by calculating the result of each parameter based on the feature selection method and input dataset for each model in a conventional manner. The importance of achieving the above-listed goals gives the best awareness for the end-users. The significance of achieving the above goal result found in appendix B in detail. This study puts more challenging work or uncovered work on the specific area of this study as future work.

- ❖ To determine west water, whether it used for reuse in irrigation or not.

- Performing a framework for water pollution monitoring and provide ways to implement real-time water pollution monitoring for the water pollution environment.
- To reduce the cost required for polluted water treatment, and time consumed for west water treatment.
- Enhance the accuracy of west water treatment and reduce waterborne disease.
- We are proposing a mathematical model for each core parameter used for water pollution identification in the ways of the machine learning technique.
- Developing the prototype for the proposed framework using machine learning algorithms in python tools. Finally, performance evaluation for the proposed framework.

Appendix C: Simple Code for Waterproof Temperature

temo_demo §

```
#include <OneWire.h>
#include <DallasTemperature.h>

// Data wire is conntec to the Arduino digital pin 4
#define ONE_WIRE_BUS 4

// Setup a oneWire instance to communicate with any OneWire devices
OneWire oneWire(ONE_WIRE_BUS);

// Pass our oneWire reference to Dallas Temperature sensor with dallas library
DallasTemperature sensors(&oneWire);

void setup(void)
{
  // Start serial communication for debugging purposes
  Serial.begin(9600);
  sensors.begin();
}

void loop(void){
  //Call sensors.requestTemperatures() to issue a global temperature and Requests to all devices on the bus
  sensors.requestTemperatures();
  Serial.print("Celsius temperature: ");
  // Why "byIndex"? You can have more than one IC on the same bus. 0 refers to the first IC on the wire
  Serial.print(sensors.getTempCByIndex(0));
  Serial.print(" - Fahrenheit temperature: ");
  Serial.println(sensors.getTempFByIndex(0));
  delay(1000);
}
```

Appendix D: Simple Code for PH Analog Reading

```
SEN0161_29_12§
#define LED 13
#define samplingInterval 20
#define printInterval 800
#define ArrayLenth  40    //times of collection
int pHArray[ArrayLenth];   //Store the average value of the sensor on CSV file
int pHArrayIndex=0;

void setup(void) {
  pinMode(LED,OUTPUT);
  Serial.begin(9600);
  Serial.println("pH meter experiment!");    //Test the serial monitor
}

void loop(void) {
  static unsigned long samplingTime = millis();
  static unsigned long printTime = millis();
  static float pHValue,voltage;
  if(millis()-samplingTime > samplingInterval)
  {
      pHArray[pHArrayIndex++]=analogRead(SensorPin);
      if(pHArrayIndex==ArrayLenth)pHArrayIndex=0;
      voltage = avergearray(pHArray, ArrayLenth)*5.0/1024;
      pHValue = 3.5*voltage+Offset;
      samplingTime=millis();
  }
```

Appendix E: Combined PH and Waterproof Temperature Sample Code

```
/*

 # Product: analog pH meter

 # SKU   : SEN0161

*/

#include <One Wire.h>

#include<Dallas Temperature.h>

// Data wire is connected to the Arduino digital pin 4

#define ONE_WIRE_BUS 4

// Setup a one Wire instance to communicate with any One Wire devices

One Wire one Wire (ONE_WIRE_BUS);

// Pass our one Wire reference to Dallas Temperature sensor

Dallas Temperature sensors (&one Wire);

#define Sensor Pin A0 //pH meter Analog output to Arduino Analog Input 0

#define Offset 0.00 //deviation compensate

#define sampling Interval 20

#define print Interval 1000

#define Array Length 40 //times of collection

Int pH Array [Array Length]; //Store the average value of the sensor feedback

Int pH Array Index=0;

Void setup (void) {

  Serial. Begin (9600);

  sensors. Begin ();

}

Void loop (void) {

  Static unsigned long sampling Time = millis ();

  Static unsigned long print Time = millis ();

  Static float pH Value, voltage;

sensors. request Temperatures ();

  //Serial.Print ("Celsius temperature: ");

  Serial.Print (sensors. getTempCByIndex (0));

  if(millis()-sampling Time > sampling Interval) {

      PH Array [pHArrayIndex++]= analog Read (SensorPin);

     If (pHArrayIndex==Array Length) {

PHArrayIndex=0;

      }

      Voltage = avergearray (pHArray, ArrayLenth) *5.0/1024;

      PH Value = 3.5*voltage+Offset;

      Sampling Time=millis ();

  }

  if (millis () - print Time > print Interval) {  //Every 800 milliseconds, print a numerical, convert the state of the LED indicator

    /*

    Serial.print ("Voltage:");

    Serial.print (voltage, 2);

    */

    Serial.print (“ “);

    Serial.print (pHValue,2);

    Print Time=millis ();

  }

  Delay (1000);

}
```