A WATER QUALITY MONITORING SYSTEM USING WIRELESS SENSOR NETWORKS

NAHSHON MOKUA OBIRI

A Thesis Submitted in Partial Fulfillment of the requirement for the Award of the Degree of Master of Science in Telecommunication Engineering, in the School of Engineering, Dedan Kimathi University of Technology (DeKUT).

NOVEMBER, 2021

DECLARATION

This thesis is my original work and has not been presented in any university/institution for a degree or consideration of any certification.

Signature: ……………………………………….Date: …………………………

Nahshon Mokua Obiri

E224-01-2005/2018

This thesis has been submitted for examination with our/my approval as the University Supervisor(s).

Signature: ……………………………………….Date: .…………………………….

Dr. Ciira wa Maina, PhD

Electrical and Electronic Engineering Department

Dedan Kimathi University of Technology, Kenya

Signature: ……………………………………….Date: .……/11/2021………

Dr. Henry Kiragu, PhD

Electrical and Communication Department

Multimedia University of Kenya, Kenya

DEDICATION

I dedicate this work to my dear mother (Ms. Milcah Mogoi), who I am highly indebted to, my spouse Lilian Nyakara, for their love, support, and encouragement during the entire time of my research alongside other life matters.

ACKNOWLEDGMENTS

First, I would like to thank the Dedan Kimathi University of Technology for their sponsorship for the Master of Science in Telecommunication Engineering. I am grateful to the Electrical and Electronic Engineering Department members and the Center for Data Science and Artificial Intelligence (DSAIL) in the Dedan Kimathi University of Technology (DeKUT) for their mentorship, support, and inspiration. I will not forget the Printed Circuit Boards (PCB) lab for the provision of necessary equipment for the prototype development of this research.

I acknowledge Dr. Ciira wa Maina, DeKUT, alongside Dr. Henry Kiragu, Multimedia University of Kenya, for their valuable comments and supervision of this research. You are my mentors, and you are the main inspiration for this work. Thanks for believing in me and changing and bettering my life over the difficult times in this endeavor.

Also, a remarkable vote of gratitude to the Nyeri Water and Sanitation Company (NYEWASCO) water quality laboratory staff, more so Mr. Mohamed Yunis Ali (Head of the Lab), Mr. Paul Michuki (Head of Operations), and their lab technicians, who assisted and guided in sensor node deployment for data collection. Without them, this research would have been something complex, not just close to how it looks.

Additionally, I would like to thank the Advanced Reduced Instruction Set Computing (RISC) Machines (ARM) for hardware support. I could go on and on to acknowledge all those who worked with me, but your name written here won't express the great feeling I have for you.

Finally, for love, patience, and great support, I sincerely thank my dear mother, who showed me what life is all about. This is the culmination of the things you have instilled in me, and I am very grateful for your unwavering support during the study and thesis research process.

Above all, Thanks to The Almighty.

TABLE OF CONTENTS

DECLARATION..i

DEDICATION..i

ACKNOWLEDGMENTS ii

LIST OF TABLES vii

LIST OF FIGURES viii

ABBREVIATIONS AND ACRONYMS..x

ABSTRACT..xi

CHAPTER ONE..1

INTRODUCTION..1

1.1 Background of the Study..1

1.2 Problem Statement 3

1.3 Objectives..4

1.3.1 Main Objective..4

1.3.2 Specific Objectives..4

1.4 Scope of the Study..4

1.5 Outline of the Research 4

1.6 Contributions of this Research 5

CHAPTER TWO 6

LITERATURE REVIEW..6

2.1 Water Quality Parameters..6

2.1.1 Turbidity..7

2.1.2 pH..7

2.2 Water Quality Management..7

2.3 Conventional Water Treatment 8

2.4 Previous Works 8

2.5 WSN-Based WQM Framework..9

2.5.1 Data Acquisition..10

2.5.2 Filtering/Processing..10

2.5.3 Network Communication..10

2.5.4 Energy Management..10

2.5.5 Data Processing, Storage, and Retrieval..11

2.6 Emerging and Trending Issues..11

2.6.1 Data Computation, Analysis, and Reporting..12

2.6.2 Data Communication and Transmission..12

2.6.3 Energy Management..12

2.7 The Intended Approach..13

2.8 Overview of LoRa..13

2.8.1 Long Range Wide Area Network (LoRaWAN)..13

2.8.2 LoRa Network Architecture..13

2.8.3 Parameters of the Physical Layer and Network Capacity..14

2.8.4 Why LoRa Technology..15

2.9 Anomaly Detection and Machine Learning..15

2.9.1 Anomalies..15

2.9.2 Machine Learning Algorithms..17

2.9.3 Anomaly Detection Algorithms..17

I. Local Outlier Factor..19

II. Isolation Forest..21

III. Extended Isolation Forest..24

IV. Robust Random Cut Forest..25

2.10 The Research Gap..26

CHAPTER THREE 28

RESEARCH METHODOLOGY 28

3.1 LoRa Connectivity and Range Evaluation 28

3.1.1 Statement of Purpose..28

3.1.2 Measurement Setup 28

I. Base Station 29

II. End Device 30

3.2 Water Quality Parameters..31

3.2.1 Turbidity..32

3.2.2 pH 34

3.3 Sensor Node Designing and Fabrication 36

3.4 System Energy Management..38

3.5 Data and the Analytical Tool Determination...39

3.6 General System Design Overview..41

CHAPTER FOUR..43

RESULTS AND DISCUSSION 43

4.1 LoRa Connectivity and Range Evaluation 43

4.2 The WQM Sensor Network System Development 44

4.2.1 The pH Sensor..44

4.2.2 Turbidity Sensor 45

4.3 Anomaly Detection and Machine Learning...46

4.3.1 Performance Evaluation based on a Subset...46

A. Turbidity Dataset 46

I. The Local Outlier Factor Algorithm..47

II. Isolation Forest and the Extended Isolation Forest Algorithms 48

III. The Robust Random Cut Forest Algorithm 50

B. pH Dataset 52

I. The Local Outlier Factor Algorithm..53

II. Isolation Forest and the Extended Isolation Forest Algorithms 54

III. The Robust Random Cut Forest Algorithm 55

4.3.2 Performance Evaluation Based on the overall Datasets..57

A. Turbidity Dataset 57

I. The Local Outlier Factor Algorithm..57

II. The Isolation Forest and the Standard Isolation Forest Algorithms 58

III. The Robust Random Cut Forest Algorithm 59

B. pH Dataset 60

I. The Local Outlier Factor Algorithm..60

II. The Isolation Forest and the Extended Isolation Forest 61

III. The Robust Random Cut Forest...62

CHAPTER FIVE...64

CONCLUSIONS AND RECOMMENDATIONS 64

5.1 Conclusions 64

5.2 Recommendations 65

REFERENCES 66

APPENDICES 73

Appendix I – Anomaly Detection Algorithms 73

Appendix II – List of Conference Papers & Journal Publications 76

Appendix III – Permission and Approval Letter for Sensor Node Deployment at NYEWASCO Treatment Plant..77

LIST OF TABLES

Table 2.1: Semtech LoRa Receiver Sensitivity in dBm at Different Bandwidths and Spreading Factors [48] 15

Table 3.1: The LoRaWAN Industry Gateway Specifications (Subject To Environmental Factors and Placement of Nodes/Sensors and Gateways) 30

Table 3.2: The LoRaWAN Industry Gateway Operation (Subject To Environmental Factors and Placement of Nodes/Sensors and Gateways) 30

Table 3.3: Properties of the turbidity and water pH dataset....................40

Table 4.1: The Mean RSSI (in dBm) for the Ten (10) Test Locations 43

Table 4.2: pH validation results for the three standard solutions....................45

Table 4.3: Turbidity Validation Results using the 7 Standard Solutions....................45

Table 4.4: Turbidity Dataset Subsection....................47

Table 4.5: Turbidity outliers for the subset data as detected by the LOF algorithm....................48

Table 4.6: Turbidity outliers as detected by the RRCF algorithm for the subset data....................51

Table 4.7: pH Dataset Subsection 52

Table 4.8: pH outliers as detected by the LOF algorithm on the subset data....................53

Table 4.9: pH outliers as detected by the RRCF algorithm on the subset data....................56

Table 4.10: pH subset data algorithms performance evaluation 57

Table 4.11: Turbidity subset data algorithms performance evaluation....................57

Table 4.12: pH and Turbidity whole datasets algorithms evaluation....................63

LIST OF FIGURES

Figure 2.1: A WSN structure unit components [12] 9

Figure 2.2: LoRa Network Architecture [43]..14

Figure 2.3: Types of anomalies [48] 16

Figure 2.4: Machine learning algorithms categories: a) Supervised learning, b) Unsupervised learning [49] 17

Figure 2.5: Anomaly detection techniques...18

Figure 2.6: Local Outlier Factor (LOF) degree of outlier-ness..19

Figure 2.7: Example of reachability distance for k=4 [50] 20

Figure 2.8: The schematic diagram of a single tree (a), and the forest (b) [52]..........................22

Figure 2.9: Branching process for an anomalous data (a) and a nominal point (b) [52] 23

Figure 2.10: Branching in the EIF [52]...24

Figure 2.11: Random Cut Tree [53]...26

Figure 3.1: Test points geographical locations. [Extracted from Google Maps]............................29

Figure 3.2: The LoRaWAN industry gateway (based on the MultiTech Conduit).........................30

Figure 3.3: STM32 Nucleo board, equipped with a LoRaWAN transceiver shield 31

Figure 3.4: The DFRobot Gravity Arduino turbidity sensor interfaced with Arduino UNO R3 Board 32

Figure 3.5: The DFRobot Gravity Arduino turbidity sensor Trimmer 33

Figure 3.6: The KEBS calibrated turbidimeter 34

Figure 3.7: DFRobot's Gravity Analog pH sensor interfaced with Arduino UNO R3 Module........35

Figure 3.8: KEBS Calibrated pH meter at NYEWASCO Water Quality Lab................................36

Figure 3.9: The schematic design of one sensor node board 37

Figure 3.10: The PCB Design of the sensor node board...38

Figure 3.11: the L293D Motor Driver IC Pin-out Diagram..39

Figure 3.12: Experimental evaluation for the anomaly detection algorithms.................................41

Figure 3.13: A Picture of the sensor node of the WQM System deployed at NYEWASCO 42

Figure 3.14: General system operation flow chart...42

Figure 4.1: The Received Strength Whisker and Box Plots for the 10 Test Locations 44

Figure 4.2: A plot of probe sensor mean values against the pH-meter mean values.......................45

Figure 4.3: A plot of probe sensor mean values against the turbidimeter mean values...................46

Figure 4.4: Turbidity Dataset for the Sixty (60) days..47

Figure 4.5: A plot of LOF turbidity outliers for the subset data 48

Figure 4.6: A plot of the IF and EIF turbidity outliers for the subset data.......................................50

Figure 4.7: A plot of RRCF turbidity outliers for the subset data..51

Figure 4.8: pH Dataset for the Sixty (60) days 52

Figure 4.9: A plot of LOF pH outliers on the subset data...54

Figure 4.10: A plot of IF and EIF pH outliers on the subset data..55

Figure 4.11: A plot of RRCF pH anomalies and their outlier scores on the subset data 56

Figure 4.12: A plot of LOF turbidity anomalies on the whole dataset 58

Figure 4.13: A plot of IF and EIF turbidity anomalies on the whole dataset……..……….……. 60

Figure 4.14: A plot of RRCF turbidity anomalies and their outlier scores on the whole dataset......61

Figure 4.15: A plot of LOF pH anomalies on the whole dataset…………………..………………62

Figure 4.16: A plot of IF and EIF pH anomalies on the whole dataset…………………..……….63

Figure 4.17: A plot of RRCF pH anomalies and their outlier scores on the whole dataset……..…64

ABBREVIATIONS AND ACRONYMS

BW	Bandwidth
DeKUT	Dedan Kimathi University of Technology
EIF	Extended Isolation Forest
GCP	Google Console Protocol
IF	Isolation Forest
ISO	International Standards Organization
IoT	Internet of Things
KEBS	Kenya Bureau of Standards
LPWANs	Low Power Wide Area Networks
LOF	Local Outlier Factor
LoRa	Long Range technology
LoRaWAN	Long Range Wide Area Network
MDGs	Millennium Development Goals
NTU	Nephelometric Turbidity Units
NYEWASCO	Nyeri Water and Sanitation Company
PCB	Printed Circuit Board
RRCF	Robust Random Cut Forest
TMLB	Traditional Manual Laboratory-Based approach
TMSI	Traditional Manual in Situ
TTN	The Things Network
VM	Virtual Machine
WQM	Water Quality Management
WSNs	Wireless Sensor Networks

ABSTRACT

Wireless sensor networks have gained popularity since remote water quality monitoring happens instantaneously with minimal human supervision, unlike in the current outdated manual methods. Therefore, this research focused on developing a real-time raw water quality monitoring system using wireless sensor networks. Four specific objectives started from *LoRa* technology studies for data transmission to sensor integration, deployment, and data analysis were pursued. The following methodology was adopted. First, experiments on *LoRa* technology connectivity and range evaluation for wireless sensor networks were conducted since *LoRa* technology performance varies with topographical features. This relied on the received signal strength indicator of signal transmission between an end device and the gateway. These studies were carried out at the Dedan Kimathi University of Technology. Secondly, the *DFRobot Gravity Arduino* turbidity sensor and the *DFRobot's Gravity Analog* pH sensor were integrated. During their calibration, significant consideration was given to obtaining linear responses, mitigating noise, high accuracy, and quality resolution. They are power-hungry, and therefore a mechanism to switch them off during times of no data sensing was developed; based on an *H-Bridge* motor control circuit. After that, the developed system was deployed at the Nyeri Water and Sanitation Company water quality treatment plant in the outskirts of Nyeri Town from the 4th of November 2020 for 60 days. The sensed data values of these parameters were relayed to a gateway by a wireless LoRaWAN transceiver installed at the plant. The gateway then forwarded the received data to The Things Network platform, interfaced with a Google Cloud Platform Console, containing an influx dB virtual machine database. A web-based application (Dash *Plotly* app) was deployed for real-time visualization of the acquired data. A total of 2,658 records containing turbidity and pH were collected. A subset of 291 records was extracted and manually examined as the ground truth. This subset was also verified with a comparison of the data collected manually by the treatment plant technicians. Lastly, analyses based on machine learning anomaly detection algorithms were performed for the evaluation of each parameter. The techniques analyzed included the Local Outlier Factor, the Isolation Forest, the Extended Isolation Forest, and the Robust Random Cut Forest algorithms. The Local Outlier Factor was the easiest to use as long as optimum parameters were selected. With little or no training, it emerged as a powerful tool compared to the other three algorithms. The overall results demonstrated that a successful low-cost and real-time water quality monitoring system was developed and deployed. The framework is more suitable for large-scale implementation to collect and analyze raw water quality data in water supply firms and water authorities. The developed water quality management system can be installed in multiple locations in water distribution networks to gather water quality data and compare sensor values in practical deployment. Moreover, more water quality sensors can be incorporated into the developed system, like temperature, for robustness.

CHAPTER ONE

INTRODUCTION

Human activities have various effects on the environment. This has adversely affected human health in many ways [1]. Developing countries have been the most affected by the growth of numerous slums, poor sanitation, and post-mining effects. The collective impact leads to a deteriorating environment. Environmental monitoring programs and systems have, therefore, been established globally to promote ecological sustainability. For example, governments and international organizations have put in place air quality monitoring systems as a concerted effort to enhance environmental sustainability [2], [3]. Water quality monitoring systems have also been proposed [4], [5]. Another notable effort is animal tracking towards environmental sustainability [6], [7]. Interestingly, it has also been noted that monitoring earthquakes will improve the quality of the environment [8], [9]. This thesis focused on water quality systems.

1.1 Background of the Study

Freshwater management has faced severe challenges in many world economies. These challenges are due to escalating competition for freshwater from many quarters of the ecosystem and human activities. The overexploitation of freshwater has reduced its availability for agricultural uses. Consequently, poverty alleviation has become more difficult because agricultural development is a critical contributor to its reduction. Since water is an integral contributor to food security [10], the 2002 World Summit on Sustainable Development focused on water management and its relation to the Millennium Development Goals (MDGs) [11]. The summit's consensus was that for water resource sustainability to be achieved, these resources should be exploited with care, bearing in mind their importance to future generations. The current exploitation trends and competition for water resources fail to guarantee that the envisioned sustainability will be achieved [11]. Therefore, all stakeholders were tasked to make rational decisions, projections, and plans to sustainably exploit and manage water resources. It was agreed that a universally accepted approach must be employed at all levels of society if water management goals are achieved.

Water Quality Management (WQM) between the 1960s and early 2000s depended on manual sampling and analysis of water. Researchers collected samples from water sources and tested them in laboratories. The main focus was on the strategies and particular methods of water analysis. In that framework, the network design was crucial. The researchers would define water quality parameters to be studied, the water sites to be sampled, and the water samples' rates [12]. In the early 2000s, new technology was integrated into WQM to remedy some limitations in the manual methods employed in the previous few decades. Notably, microelectronic mechanical sensors, fibre optics, laser technology, biosensors, among other sensors, revolutionized water quality analysis [13], [14]. These sensors identify various aspects of water quality in situ.

Moreover, the advanced technology introduced water telemetry, which enhanced the acquisition of water quality data and accompanying monitoring procedures. Satellite technology also facilitated the acquisition of water images used to approximate different water quality parameters [14]. Lakes, rivers, springs, and seas, among other water bodies, could also be monitored using visualization architectures for water quality courtesy of modern technology [15]. All the advancements remedied the undoing of manual water sampling and analysis as they introduced automatic selection and monitoring points where water could be analyzed periodically [12], [14], [16], [17].

The introduction of Wireless Sensor Networks (WSNs) in the early 2000s further bolstered WQM due to improved communication systems. Their ease of operation has made them increasingly popular. These networks have promoted quick capture, transmission, and analysis of data relating to the environment. The application of WSNs in WQM procedures has lowered the sensing costs and increased the amount of data and sampling points analyzed at any particular moment. The WSNs also have an in-built capability to transfer data by utilizing low-power techniques. This capability enhances the easy remote transmission of data from numerous data sensors and points. Therefore, the new technology is more appealing than the previous manual methods.

1.2 Problem Statement

Many parts of the world have continually faced challenges in sustaining the supply of safe drinking water. The challenges include inadequate resources and weak water systems that cannot promptly identify water contamination, leakages, and blockages. These operational failures are a result of insufficient planning, which excludes environmental considerations. Thus, the most affected regions are likely to incur severe water-related hazards. To solve some of these issues, technologically enhanced water management methods should be used to quickly identify and address system failures and minimize water-related problems, wastage, and losses.

Whereas the local water quality treatment plants adopt the traditional manual lab-based water quality monitoring systems, the existing water quality management systems suffer several shortcomings. The traditional manual methods are outdated. Other than being costly, they can be timewasting and subjected to bias in terms of parameter monitoring. On the other hand, the newly developed systems based on wireless sensors have many setbacks, including a lot of power consumption, false anomaly detection, and inappropriate wireless technologies. The adopted technologies are understood to have constraints of coverage and connectivity range, let alone improper anomaly detection methods!

Therefore, this research focused on developing a real-time and low-cost water quality monitoring system based on wireless sensor networks and a more easy-to-use technology curbing the disadvantages of the strength of connectivity, power usage, and range of coverage. Additionally, an effective machine-learning anomaly detection algorithm was determined to minimize the issues of false alarms in the event of water contamination.

1.3 Objectives

1.3.1 Main Objective

To develop and implement a real-time water quality monitoring system based on wireless sensor networks.

1.3.2 Specific Objectives

i. To determine the range of coverage and the strength of connectivity for Long-Range (LoRa) technology in a rural setup surrounding DeKUT.

ii. To calibrate and integrate the pH and the turbidity sensors to be used in the sensor node.

iii. To design and fabricate a sensor node to be deployed for water quality parameter collection and transmission.

iv. To determine an effective machine-learning analytical algorithm for real-time anomaly detection in the monitored parameters.

1.4 Scope of the Study

Among the many technological service providers of the low power wide area networks (LPWANs), this study was limited to LoRa technology alone; and therefore, the LoRa connectivity and range of coverage studies. Consequently, since the performance of this wireless communication technology has varied performance in different topographies, a rural setup area of the Dedan Kimathi University of Technology (DeKUT) was chosen as the area of study.

The developed system was also limited to the section of raw water quality monitoring of the parameters for the treatment procedures: turbidity and pH. These studies were carried out at the Nyeri Water and Sanitation Company (NYEWASCO) raw water section. They were willing to share their data to verify the parameters collected with the developed system in this study.

1.5 Outline of the Research

This thesis is organized into five main chapters with their corresponding subsections. Chapter one provides the background to the research study and ends up setting out the aims and objectives of this study. Chapter two deals with the literature review describing the overview of water quality monitoring systems. It also describes various sections encompassing a water quality monitoring system and compares the multiple approaches to wireless sensor networks. Chapter three describes the methodology of developing the water quality monitoring system. This starts with LoRa technology studies, water quality sensors calibration, experimental set-ups, computations, sensor node deployment, and anomaly detection algorithms evaluation. Chapter four covers the results, analysis, and discussions of the developed water quality monitoring system. Finally, chapter five gives conclusions and recommendations based on the results obtained.

1.6 Contributions of this Research

Based on the technical characteristics of wireless sensor networks, one of the main contributions was to develop a water quality monitoring system based on *LoRa* technology, a low power, long-range wireless transmission technology. This led to *LoRa* technology studies in the Dedan Kimathi University of Technology to determine its connectivity. From the results obtained, this technology addressed the challenges of connectivity and range of coverage.

Moreover, there was a determination of an effective anomaly detection algorithm for time-series water quality parameters. The techniques analyzed included the Local Outlier Factor, the Isolation Forest, the Extended Isolation Forest, and the Robust Random Cut Forest algorithms. The Local Outlier Factor was determined to be easier to use as long optimum parameters are selected. With little or no training, it is a powerful tool for anomaly detection of water quality data compared to the other three analyzed.

CHAPTER TWO

LITERATURE REVIEW

Amongst the several environmental quality monitoring systems, water quality management is a crucial consideration in the sector. It is investable since water is a precious commodity to the sustenance of standard environmental quality. This chapter defines and discusses water quality systems and emerging technologies, while the vital intention was to identify the research gap.

2.1 Water Quality Parameters

The water condition in terms of chemical, physical and biological characteristics, usually concerning its suitability for a particular purpose (i.e., drinking, swimming, fishing, etc.), is described using water quality [18]. The presence of substances such as pesticides or fertilizers in specific concentrations impacts water quality, thereby negatively affecting marine life. A measure of water quality is provided by factors such as concentration of dissolved oxygen (DO); levels of fecal coliform bacteria from human and animal wastes; concentrations of plant nutrients, nitrogen, and phosphorus; the amount of particulate matter suspended in the water (turbidity); and amount of salt (salinity). The concentration of chlorophyll-a, a green pigment found in microscopic algae, is also filtered from water samples to measure the microalgae living in the water column in many bodies of water [19]. Moreover, determining water quality may be possible by measuring quantities of pesticides, herbicides, heavy metals, and other impurities.

Water quality can be classified into raw and treated parameters [20]. Whereas the raw water quality parameters determine the treatment procedures, the treated water quality parameters assess the safety of use in river flow maintenance, drinking, industrial water supply, water recreation, irrigation, and many other services, including being safely returned to the environment. This research discusses the main parameters of raw water: Turbidity and pH.

2.1.1 Turbidity

The level of cloudiness of water caused by suspended particles is measured. Turbidity is obtained using the ISO 7027 approach, where infrared light scatters at right angles to cross beams [21]. It is indicated in Nephelometric Turbidity Units (NTUs). Turbid waters are susceptible to escalated growth of microbes as they provide sufficient food and shelter for pathogens. A turbidity sensor measures transmittance and scattering rate, which varies with the total solids suspended (TSS).

2.1.2 pH

The pH of water is indicated by use a negative logarithm of the concentration of hydrogen ions in moles per liter. It shows the acidic or basic levels. Despite that requirement, the pH does not cause any health complications [22]. If the pH of water suddenly changes by a minimum of 0.5 pH units, there is a reason to suspect contamination. During the measurement of pH, a combined pH electrode is utilized to ensure accuracy.

2.2 Water Quality Management

Water quality management is defined as the idea of water constituents and conditions being sampled and analyzed. The monitored elements include naturally occurring ones (such as nutrients, oxygen and bacteria) that remain unaffected by human resources, and pollutants including metals, oil, and pesticides [23]. The extent to which their effect is felt depends on factors including temperature and pH. For instance, the quality of dissolved oxygen that water can contain is determined by temperature, whereas the level of toxins in ammonia is determined by pH.

Water quality has been monitored for several years by several groups ranging from researchers, volunteers, and professionals at the local and state levels. Until and up to the past decade, before the development and application of monitoring using biological protocols, water monitoring has been the fundamental way of establishing water pollution problems [24]. All the stakeholders in water quality management are currently focused on developing methods of uniting physical, chemical and biological modes of monitoring for pre-eminent conditions of water quality.

2.3 Conventional Water Treatment

Water treatment for any particular use entails a conventional method of two screening stages: A raw water section (primary screening) and the treated water section (secondary) [25]. Preliminary screening entails two main parameters; turbidity and pH. On the other hand, secondary screening entails several parameters, including residual chlorine, oxygen reduction potential, and electrical conductivity.

Primary screening determines the treatment process and, therefore, the most fundamental stage. The amount of treatment chemicals input for coagulation, flocculation, and sedimentation procedures is determined by the measurements obtained from these two parameters [25]. The process requires chemical knowledge of source water characteristics to ensure that a compelling coagulation mix is employed. Inappropriate coagulants make these treatment methods ineffective [26]. The most widely used coagulant is Aluminium Sulphate which is commonly called alum.

2.4 Previous Works

In the last two decades, various researchers have submitted that a WSN is the most suitable method for WQM [27], [28], [29], [30]. Online platforms have also been increasingly used to analyze data and automatically discern water quality-related problems in the past few years [12], [15], [31]. Research indicates that the WSNs method overcomes most of the limitations experienced in the traditional manual based in situ (TMIS) and traditional manual lab-based (TMLB) techniques. Unlike the traditional approaches, the WSNs method can replace outdated and expensive equipment with low-cost sensors. They eliminate the need to transport data samples to the laboratory, thereby saving time and reducing costs in the process. The training of workers, collection of samples, data recording, and data analysis with the WSNs has proven cheaper when compared to the traditional WQM techniques [28], [32], [33]. Therefore, the WSNs method is currently preferred to its traditional competitors.

The WSNs can be designed to track the quality of water in freshwater sites. However, various aspects must be considered before its implementation, including; sensing abilities of the nodes, signal processing, network layout, and whether the sensors are likely to use acoustic or radio communication. The WSNs can be used to track the water quality parameters such as the pH, temperature, turbidity, and the level of dissolved oxygen in water [34], [35], [36]. For underwater tracking, acoustic communication is required for WQM. However, if the process involves surface monitoring, radio communication is the preferred mode. Exclusive examples of multi-sensor systems and embedded WSN systems include the SmartCoast [37] and the LakeNet [12]. Most of the current approaches based on WSNs relied on the Arduino UNO as the microcontroller and WiFi a mode of wireless data transmission as in the cases of Meghana et al. [38], and Chowdury et al. [39].

2.5 WSN-Based WQM Framework

This segment briefly explains the WSNs framework. The diverse components that form the framework are discussed below to expound on the WSN-based WQM structure. Figure 2.1 illustrates the WSN structure as a unit. The unit is built by four components responsible for four primary operations: data acquisition, data filtering, data transfer, and ultimately, data analysis, information storage, and presentation.

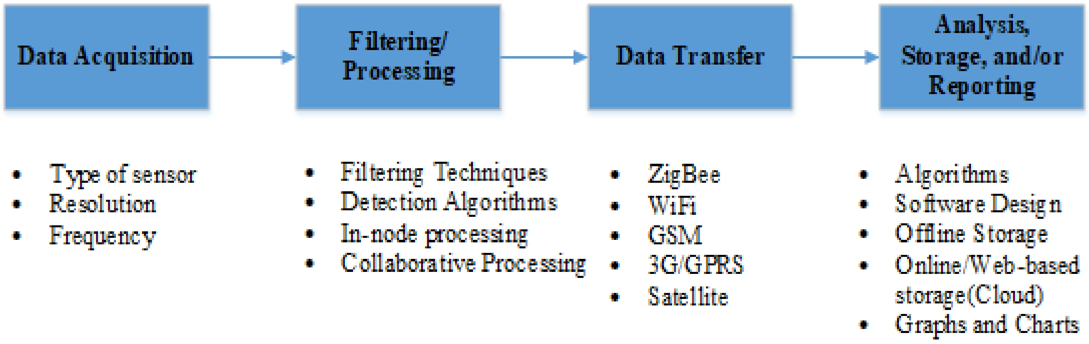


Figure 2.1: A WSN structure unit components [12]

2.5.1 Data Acquisition

Spatially distributed sensor nodes periodically collect data from different water samples. This is to increase the chances of getting accurate results from various locations [12]. The high frequency of the sensors makes it easy for the researchers to collect data for their WQM projects.

2.5.2 Filtering/Processing

In the filtering and processing component, every sample collected in the data acquisition stage is processed [12]. This phase calls for the application of particular computations and devices that are remotely capable of high-level operations. Filtering techniques are deployed to categorize different water quality parameters.

2.5.3 Network Communication

The WSN system focuses its attention on the structure of the communication network to which it applies. This framework has two network designs, which are remote and local communication. Remote communication involves transmitting data from the local station to remote stations for potential users to access it. Local communication involves data transfer from sensor nodes to base stations where it can be accessed. These local networks use ZigBee and WiFi, among other systems, to transmit data. On the other hand, remote networks can apply cellular communication such as LTE, GSM, and GPRS to facilitate the data transmission to the local monitoring stations. Sometimes the data may be transferred to the cloud instead of the local tracking stations. Many studies indicate that remote networks are preferred [12], [29], [30], [40], [41], [42].

Communication modes such as GSM, WiMax, and LTE usually offer a coverage of approximately 100 km. Such ranges are suitable for the remote tracking of water environments. ZigBee, WiFi Direct, and WiFi are ideal for local monitoring. Their coverage range is between 50m and 200m.

2.5.4 Energy Management

At the onset of the design of the WSN system for WQM, researchers focused on the power consumption of devices used in the system and how it was managed. The primary objective was to create a system that is continuously operational without the continuous replacement of batteries. There are two approaches that researchers can use to enable continuous operation. The first one involves increasing the energy supplied to the nodes by using renewable energy such as wind, hydroelectric power (HEP), and radiofrequency radiation (RF). Secondly, duty cycling, wake-up radios, power control, and standard communication procedures require low data. Lasting WSN systems require energy harvesting to be successful. Many scholars have also suggested the use of hybrid systems of power storage which combine direct solar energy and the use of secondary batteries to sustainably power the WSN systems [29], [43], [44].

2.5.5 Data Processing, Storage, and Retrieval

The WSNs model pipeline involves an analysis of data, storage, and communication. The system must involve extra computations, organization of data, and classification of data that the system has collected. The data can be stored offline, online, or in the cloud. Data presentation can be done through methods such as graphs and tables. There are several places for processing data collected using multiple sensor nodes [45]. They include local monitoring stations, wireless sensor nodes, and remote tracking stations. The algorithms deployed at every stage of data processing depend on the parameter of interest. Data is enabled, processed, and retrieved to be analyzed further. All analysis is geared towards the laid-out objectives of the water quality management system.

2.6 Emerging and Trending Issues

The preceding sections consider the various aspects of the WSN system of WQM, energy requirements, water quality measurement, network design, and the implementation of each element. Relevant current developments associated with the WSN framework have also been highlighted. Many scholars have proposed the WSN-based model as the most suitable approach for analyzing water quality. Even so, some areas of this approach require further study and trials [12]. Notably, the system has to address issues that arise from different network architectures and their implementation. This research identifies that WSN can be exploited for its gains. The obstacles related to power management, computation of data, and transmission of the data are addressed in this research.

2.6.1 Data Computation, Analysis, and Reporting

Data calculation, analyses, and reporting of the results can be conducted at either base or remote stations. All these procedures can be automated [17]. Different quality aspects can be estimated at multiple points to determine if water has been contaminated. Notably, water testing algorithms that link numerous water testing qualities from different sections of the water source are yet to be designed.

2.6.2 Data Communication and Transmission

The discussion in this research reveals techniques used in already installed wireless technologies such as WiFi Direct, LTE, and ZigBee, among others [44]. LoRa technology has not been efficiently implemented in WSN platforms. Extensively implementing these communication technologies may benefit bandwidth and network coverage. For instance, if WSNs employ WiFi Direct, they gain more bandwidth and network coverage than ZigBee. Conversely, the ZigBee consumes lower power than WiFi-Direct and hence a trade-off between power consumption and benefits such as higher bandwidth and coverage.

2.6.3 Energy Management

The WSN-based frameworks consider energy as a vital resource in the determination of water quality. It is integral to the transfer of data from local tracking stations. Solar energy has proven to contain the most excellent density (15mW/cm^2 on sunny days) compared with other energy sources [46]. Possible instances of power minimization are; the utilization of low power sensors, duty cycling, and scaling of voltages, and algorithms used to guide sleep times at specific levels [47].

2.7 The Intended Approach

The telecommunication industry has long shown interest in low-power wide-area networks (LPWANs) [48]. Many technology service providers are competing to offer LPWAN services, including the LoRa Alliance [49]. They have standardization of costs, optimization of batteries, and coverage. They incur low rates of data transmission since they have low bandwidth and low power transmissions. WSNs have gained from standard procedures by vendors who have robust features all over the world [30]. This study, therefore, employed low-power sensors in tracking the progress of water quality using the LoRa technology. The technology addressed energy consumption, transmission, data reporting, and storage challenges in the existing WQMs. Moreover, data analysis and anomaly detection is handled with a more advanced approach of machine learning anomaly detection algorithms to eradicate the challenges of the current methods of anomaly detection: complex computations, false anomalies, too much computation time, among other loopholes.

2.8 Overview of LoRa

Long-range communication can be achieved in LoRa-LPWAN through the deployment of sub-Gigahertz radio bands and limited network data rates that improve the sensitivity of receivers. This technique involves low power consumption [49]. Thus, there is the usage of devices powered by long-lasting batteries.

2.8.1 Long Range Wide Area Network (LoRaWAN)

LoRaWAN describes the communication protocol and architecture used in LoRa communication while LoRa’s physical layer establishes the communication link in LoRa. The system's protocol and design determine the battery life of node batteries, network capacity, the number of network applications, and network security.

2.8.2 LoRa Network Architecture

“A star-of-stars topology” is a typical LoRa network with the inclusion of three diverse types of devices [49], as shown in Figure 2.2. Individual end-nodes in a mesh network forward information from other nodes. The aim is to widen the range of communication and the network’s cell size. When irrelevant information is received and forwarded by the nodes, the range is increased, the network's capacity is reduced, complexity is added, and the battery life is reduced. Thus, achieving long-range connectivity in the star architecture fosters a long battery life by enhancing battery preservation. The LoRa-WAN simple design described in the following sections elaborates on how LoRa-WAN uses gateways to communicate with end devices. Using Ethernet or 3G networks, the portals forward LoRa-WAN frames to network servers from the identified devices. The gateways are bidirectional, whereas the network server accounts for decoding the packets sent to and from the devices on the LoRa-WAN.

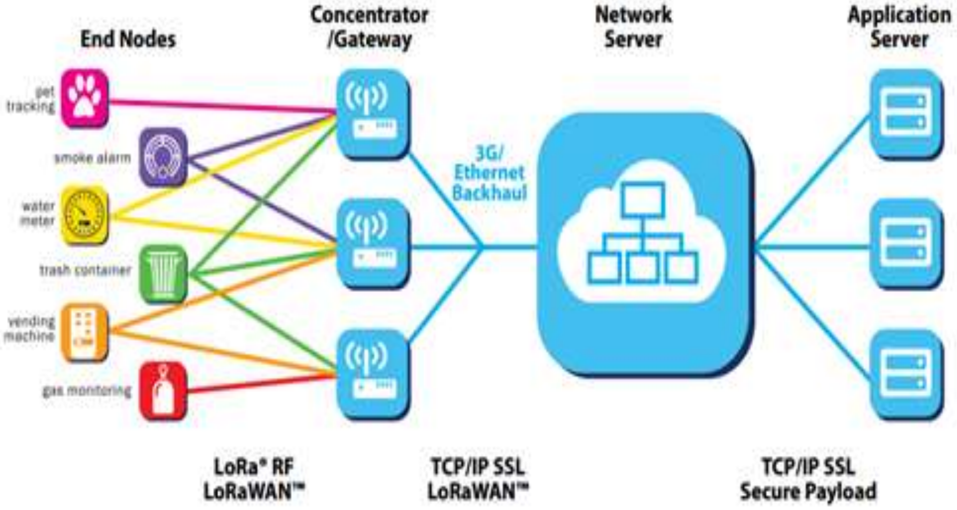


Figure 2.2: LoRa Network Architecture [45]

2.8.3 Parameters of the Physical Layer and Network Capacity

The parameters of LoRa that can be modified to include LoRa harmonization encompasses the bandwidth (BW), code rate (CR), and the spreading factor (SF) [49]. The chirp rate determines the BW (one chirp/s/Hz of BW). The BW frequency is proportional to both the symbol and bit rates at a specific spreading factor. Therefore, if the BW is doubled, the transmission rate is also doubled.

Conversely, an increase in BW reduces the sensitivity of the receivers. Similarly, an increase in the spreading factor increases the receiver's sensitivity. Table 2.1 below illustrates this.

Table 0.1: Semtech LoRa Receiver Sensitivity in dBm at Different Bandwidths and Spreading Factors [48]

SF BW	7	8	9	10	11	12
125kHz	-123	-126	-129	-132	-133	-136
250 kHz	-120	-123	-125	-128	-130	-133
500 kHz	-116	-119	-122	-125	-128	-130

2.8.4 Why LoRa Technology

The LoRa technology has several merits for the WSN-based WQM system. First, it is highly compatible with 868 MHz ISM bands that are accessible globally. Second, it offers a wide range of coverage, approximately 5 km and 15km in rural and urban areas, respectively. Third, a single LoRa gateway base station can serve thousands of other nodes and devices, contributing to the system's reliability. Given the simple design that it entails, the technology can be easily deployed. Fourth, the technology utilizes little power, thus enhancing longer battery life. Moreover, LoRa facilitates battery life prolongation since it uses adaptive data rates when varying output rates for its devices.

2.9 Anomaly Detection and Machine Learning

2.9.1 Anomalies

Data subsets that are considerably different from the rest of the data set are known as anomalies, outliers, noise, or novelties [50]. They usually happen due to measurement variability or some measurement errors often excluded from the data set. Other causes of anomalies include hardware transient malfunction experiences, data transmission errors, system behavior changes, human impacts or fraudulent behaviors, instrument errors, among many different reasons [15]. They can cause complications in data analysis and, consequently, wrong decisions. In regards to water quality management systems, ecological phenomena like rainfall or floods are expected, affecting water quality. Anomalies may also arise from human and technical errors: Sensor probes are dirty or pulled out of water for cleaning.

Anomalies can be classified into three: Global or point anomalies, an individual data point far from others in a subset. If a data instance is strange in a specific context, but not otherwise, then it is called a contextual or a conditional anomaly. And finally, collective anomalies, which are a collection of data instances in a given subset. These are illustrated in Figure 2.3 below. The process of finding out patterns in data that do not imitate the expected performance or trend is referred to as anomaly detection or outlier detection [50].

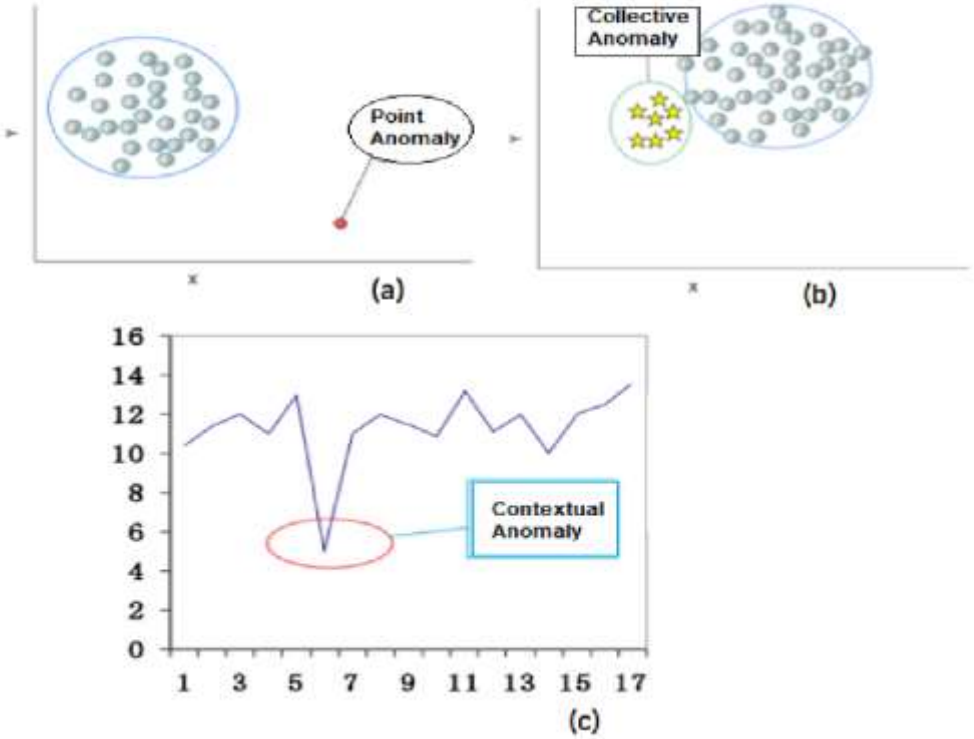


Figure 2.3: Types of anomalies [50]

2.9.2 Machine Learning Algorithms

Machine learning originated from pattern recognition and is a data analysis technique that explicitly gives computers the capability to learn minus any program [39]. Algorithms that can learn from data, identify patterns and make decisions are explored, examined, and developed. The main classes of machine learning algorithms include supervised, unsupervised or semi-supervised learning [51]. Labeled data for training is required in supervised learning, while unsupervised learning does not entail desired classified or labeled test data. Its algorithms can infer a function to describe hidden data structures from unclassified test data short of any guidance. On the other hand, semi-supervised learning lies between supervised and unsupervised learning [51]. These are illustrated in Figure 2.4 below.

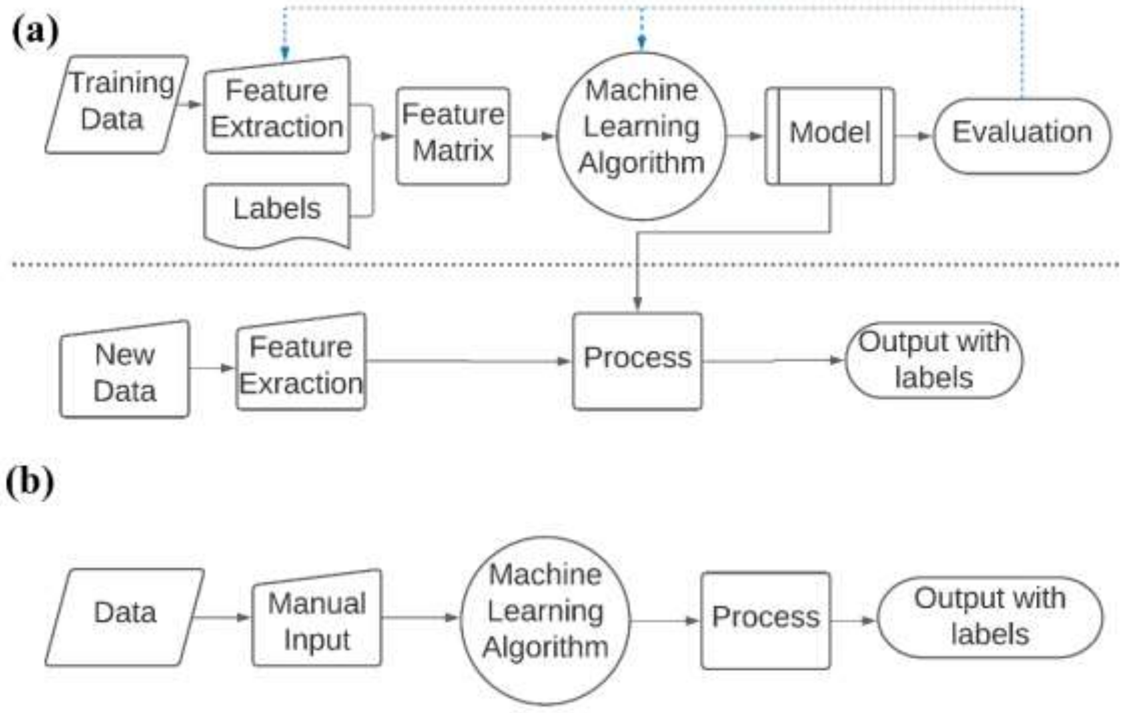
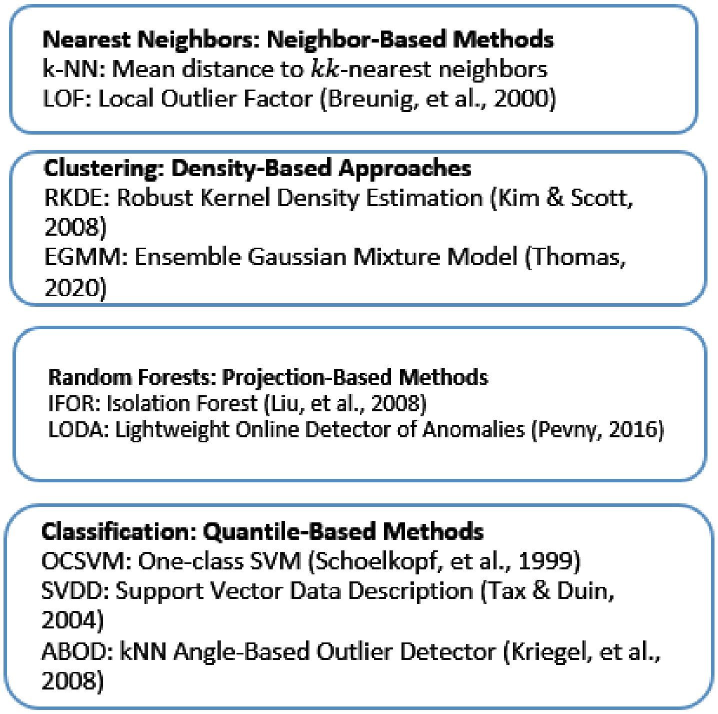


Figure 2.4: Machine learning algorithms categories: a) Supervised learning, b) Unsupervised learning [51]

2.9.3 Anomaly Detection Algorithms

The most popular anomaly detection techniques are shown in Figure 2.5. A brief overview is given in this section, followed by a deep discussion of the four methods used in this research.



Nearest Neighbors: Neighbor-Based Methods
k-NN: Mean distance to *kk*-nearest neighbors
LOF: Local Outlier Factor (Breunig, et al., 2000)

Clustering: Density-Based Approaches
RKDE: Robust Kernel Density Estimation (Kim & Scott, 2008)
EGMM: Ensemble Gaussian Mixture Model (Thomas, 2020)

Random Forests: Projection-Based Methods
IFOR: Isolation Forest (Liu, et al., 2008)
LODA: Lightweight Online Detector of Anomalies (Pevny, 2016)

Classification: Quantile-Based Methods
OCSVM: One-class SVM (Schoelkopf, et al., 1999)
SVDD: Support Vector Data Description (Tax & Duin, 2004)
ABOD: kNN Angle-Based Outlier Detector (Kriegel, et al., 2008)

Figure 2.5: Anomaly detection techniques

Local density gives a base for nearest neighbors anomaly detection techniques built on the *k*-nearest neighbors algorithm. Clustering-based anomaly detection is unsupervised learning. There is an assumption by these algorithms that similar objects tend to belong to similar groups (clusters) and that distance determines the similarity. The classification technique does the categorization of data into different classes with labels. Anomaly detection involves only two distinct classes: normal class and abnormal class. Random forests is the learning algorithm that functions by constructing decision trees' multitude at training time and class outputting (i.e., class mode-classification; prediction of mean-regression). This algorithm classifies and regresses individual trees. [50].

The Isolation Forest (IF), the Extended Isolation Forest (EIF), the Local Outlier Factor (LOF), and the Robust Random Cut Forest (RRCF) algorithms were chosen because:

- Time series data with only one variable, such as water quality data, is not suitable for clustering, supervised learning algorithms.

- There is a requirement of model training and labeled data for neural networks and support vector machine (SVM) classification-based algorithms.
- Training a generic model for classification is complex and different results for the same data point may be generated by other models.
- LOF is based on K-NN, and extensions of LOF are other nearest neighbors-based algorithms.
- And, in general, low computational complexity.

I. Local Outlier Factor

The local outlier factor (LOF) is the unsupervised outlier detection algorithm that detects the outliers by comparing the local density of the data instance with its neighbors is called the local outlier factor (LOF). It was the first algorithm based on k-neighborhood and local density [50]. LOF is the anomaly score of each sample in the training data set, and it indicates the degree of its outlier-ness, as shown in figure 2.6.

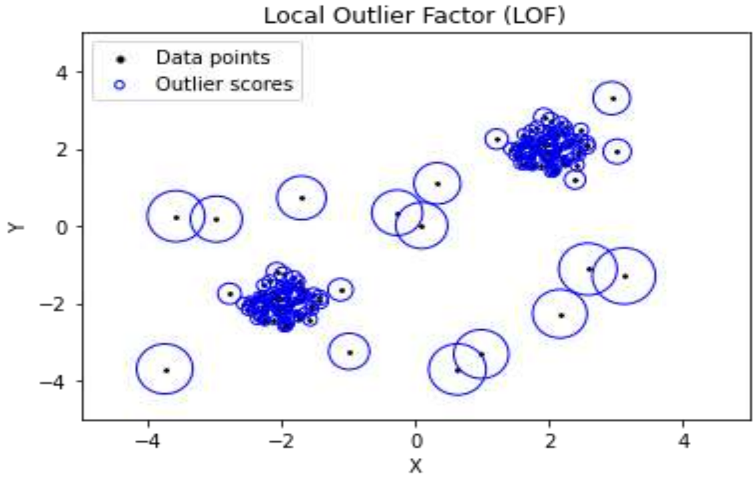


Figure 2.6: Local Outlier Factor (LOF) degree of outlier-ness

Determination of the local neighborhood of the LOF is based on the number of nearest neighbors. For the LOF to accomplish the whole process, the following definitions are used. The k-distance of instance p, denoted as k-distance (p), is defined as the distance $d\ (p, o)$ between p and an object $o \in D$ so that for at least k instances $o' \in D\backslash\{p\}$ it holds that $d(p, o') \leq d(p, o)$, and for at most $k - 1$

instance $o' \in D\backslash\{p\}$ it holds that $d(p,o') < d(p,o)$. The k-distance neighborhood of instance p is a subset with instances whose distances are not greater than the k-distance from it. With regard to instance o, the definition of reachability distance of instance p is:

$$reach - dist_k = max\{k - distance(o), d(p,o)\} \qquad 2.1$$

Figure 2.7 shows examples of reachability distance for $k = 4$. Between these two instances, the reachability distance is their actual distance when they are far away from each other actual distance (like o and $p2$); but, the reachability distance is $k - distance$ of o if they are close enough (like o and $p1$). Consequently, there can be a significant reduction of the statistical fluctuations of $d(p,o)$ for all of the $p's$ close to o. The parameter k controls the strength of this smoothing effect; therefore, the higher the value of k, the more the reachability distances similarities are within the same neighborhood [52].

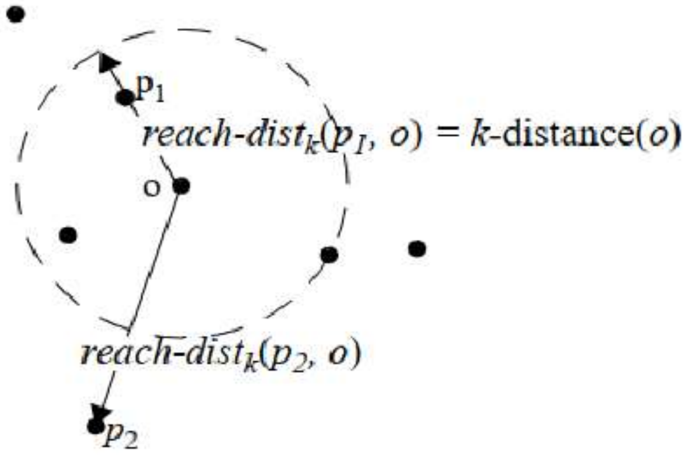


Figure 2.7: Example of reachability distance for k=4 [52]

For object $o \in N_{MinPts(p)}$, the local reachability density of point p is defined as:

$$lrd_{MinPts}(p) = 1/(\frac{\sum o \in N_{MinPts(p)} reach - dist_{MinPts(p,o)}}{|N_{MinPts(p)}|}) \qquad 2.2$$

Where;

- $MinPts$ specifies a minimum number of objects

- $reach{-}dist_{MinPts(p,o)}$ represents the reachability distance of object p with respect to object o

For object $o \in N_{MinPts(p)}$, the definition of (local) outlier factor of p is:

$$LOF_{MinPts(p)} = 1/(\frac{\sum o \in N_{MinPts(p)} \frac{lrd_{MinPts}(0)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}) \qquad 2.3$$

Where,

- $lrd_{MinPts}(p)$ is the local reachability density of p
- $lrd_{MinPts}(o)$ represents the local reachability density of p's $MinPts$ -nearest neighbors

For a given dataset, the following five calculations are obtained in the order outlined below:

i. The distances between every two instances.

ii. The distances between the k^{th} nearest neighbors to p.

iii. All the k-nearest neighbors of p.

iv. The reachability density (lrd) of p.

v. The LOFs (anomalies) of p.

II. Isolation Forest

Isolation Forest (IF) used in this research refers to an unsupervised algorithm that can identify the presence of outliers in a dataset [53]. The algorithm was designed to detect anomalies depending on their isolation. Besides, the IF analyzes mobile time series data to identify change points and outliers. The algorithm uses standard data-specific anomalies and a few anomalies in each dataset as quantitative attributes detecting outliers.

The IF algorithm begins with data training that includes tree diagrams construction [54]. First, an *N*-dimension dataset leads to a random subsample that constructs a binary tree. During the process, branching happens by selecting the dimension x_i randomly where $i \in \{1, 2, \ldots, N\}$**.** Then, the algorithm selects another random value *v* from the range of random values. If the specific data point has a smaller value than *v* for the stated dimension, the point is branched leftwards. If it exceeds *v*, the point is branched rightwards in the tree. The tree is split twice on the current node using a similar procedure. The recursive branching process continues until a single data point is isolated or a specific depth limit is achieved. The process is repeated to construct another random tree for another sub-sample. A large ensemble of trees is created to complete a process that is collectively termed as forest training. The process moves onto the scoring step, where the algorithm runs a candidate data point chosen from the trees through all the trees. An anomaly score is given to each data point depending on the depth reached by each candidate data point on the tree, as illustrated in figure 2.8. A radial line represents each tree in the model: red represents an outlier, whereas the blue radial line represents a nominal point [54].

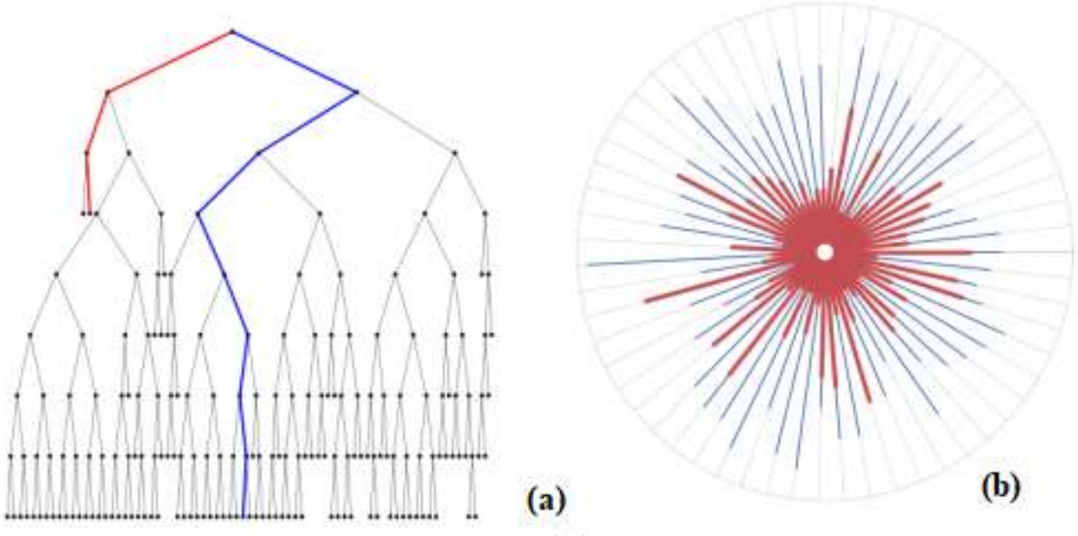


Figure 2.8: The schematic diagram of a single tree (a), and the forest (b) [54]

Every occurrence x in the anomaly detection is assigned an outlier score s useful in analysis. The outlier score s and occurrence x can be formulated as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \qquad 2.4$$

Where, $E(h(x)$ represents the depth mean-value every datapoint x reaches in every tree and $c(n)$

represents the normalizing factor (mean depth for Binary Search Tree (BST) searches that are not successful) [54].

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \qquad 2.5$$

Where $H\ (i)$, in this case, represents the harmonic number $ln(i) + 0.5772156649$ (Euler's constant) and *n* is the total number of change points used in building the trees [53].

Figure 2.9 (a) below represents the anomalous datapoint branching process where branching occurs until the point in question (red point) is isolated. Three random cuts were used to arrive at the desired isolation point. In Figure 2.9 (b), the branching process for a nominal is illustrated. The branching process requires multiple cuts to identify and isolate the point since it sits deep in the initial dataset. The tree depth limit is achieved before the point is reached. The line numbers in the figure demonstrate the order of the branching process.

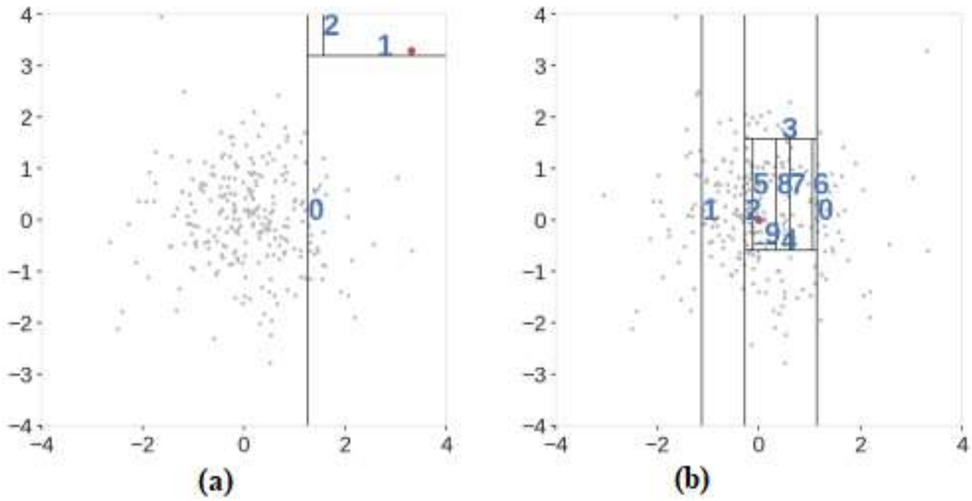


Figure 2.9: Branching process for an anomalous data (a) and a nominal point (b) [54]

The IF Algorithm utilizes two input parameters: ψ (sub-sample size) and *t* representing the number of trees. In this context, ψ determines the data size used for training. Three algorithm procedures are used to determine the anomaly score: Algorithm 1 [53], Algorithm 2 [53], and Algorithm 3 [53] (**Appendix I**).

III. Extended Isolation Forest

The Extended Isolation Forest (EIF) facilitates the improvement of anomaly score consistency and reliability. The EIF identifies various slopes for making branching cuts and then randomly assigns intercept values within the training dataset. This EIF phenomenon is different from the usual random attribute-random value method used by Isolation Forest (IF) [54], as discussed below.

Figure 2.10 (a) below demonstrates the branching process of determining an outlier. As earlier stated, branching continues till the desired point is determined, and this process took three cuts to isolate the required point. Figure 2.10 (a) shows how the branching process is used to arrive at the nominal point. The point is nearly at the center of the dataset, and therefore, several random cuts are required to isolate it. However, the depth limit is achieved before the isolation of the point for this scenario.

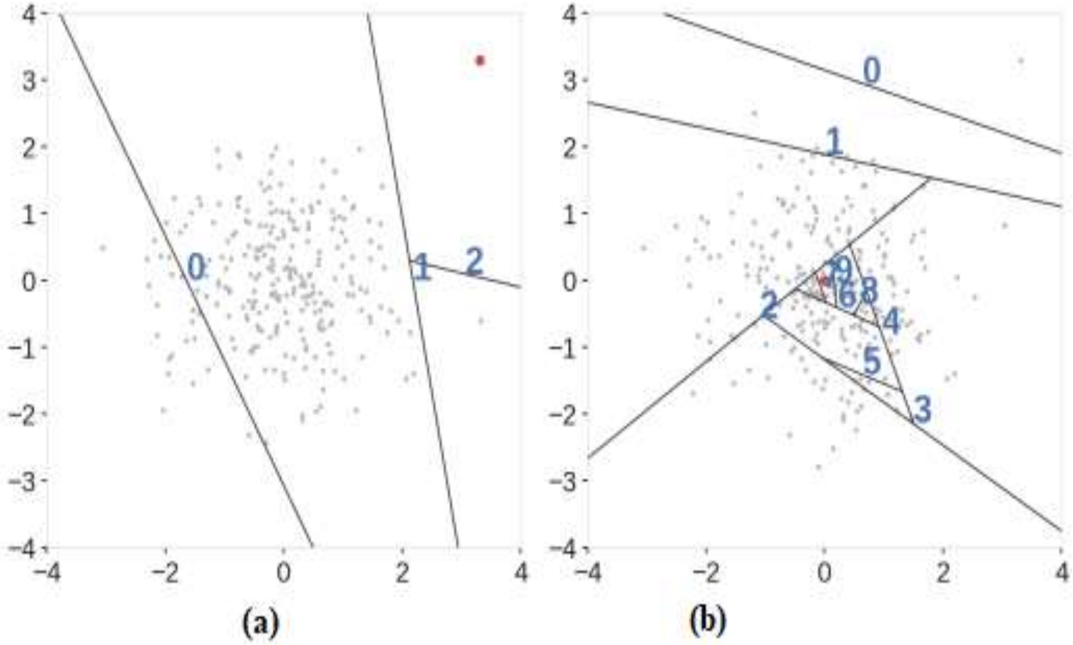


Figure 2.10: Branching in the EIF [54]

In the normal IF algorithm, two types of information are necessary for branch cut to be achieved: the coordinates and the random value from the dataset. Conversely, the EIF branch cut needs two pieces of information: the random slope-intercept from the training dataset and the branch cut slope.

Choosing a random slope from a branch cut in an *N*-dimension dataset is like selecting a normal vector $\overrightarrow{n}$ uniformly per unit of an *N*-sphere. This can be achieved through drawing random numbers for every *n*-coordinate from a normal distribution $\mathrm{N}\ (0, 1)$ and uniform *N*-sphere selection points are reached as a result. The $\overrightarrow{p}$ intercept can be obtained from a uniform dataset used at each point of branching [54]. When the two types of information are received, the branching process for splitting data for a particular point *x* proceeds as follows:

$$(\vec{x} - \vec{p})\ .\vec{n} \leq 0 \qquad\qquad 2.6$$

The data point $\vec{x}$ is passed to the left branch of the process if the condition is achieved. However, it is passed to the correct branch if that condition is not fulfilled. Then, Algorithm 2 [53] becomes Algorithm 4 [54] (**Appendix I**).

IV. Robust Random Cut Forest

This is an unsupervised algorithm for anomaly detection on streaming data proposed in 2016 [55]. Developing the machine learning model is done using current records in the stream. Neither older records nor statistics from previous executions are used by the Robust Random Cut Forest (RRCF). The standard procedure of anomaly detection using RRCF is as follows:

i. A bunch of random instances is taken by RRCF (Random).

ii. It then cuts them into the same number of instances and creates trees (Cut).

iii. Finally, all trees together are considered by determining whether a particular instance is an anomaly (Forest).

A Robust Random Cut Tree (RRCT) on point set *S* is generated as follows:

i. A random dimension proportional to $\frac{l_i}{\sum_i l_j}$ where $l_i = max_{x \in S}\ x_i - min_{x \in S}\ x_i$ is chosen.

ii. Choose $X_i \sim Uniform[min_{x \in S}\ x_i\ ,\ max_{x \in S}\ x_i]$

iii. Let $S1 = \{x | x \in S, x_i \leq X_i\}$, $S2 = S/S1$ and recurse on $S1$ and $S2$

Figure 2.8 shows how RRCF cut instance happens into pieces recursively. When each point is isolated, the cutting is stopped.

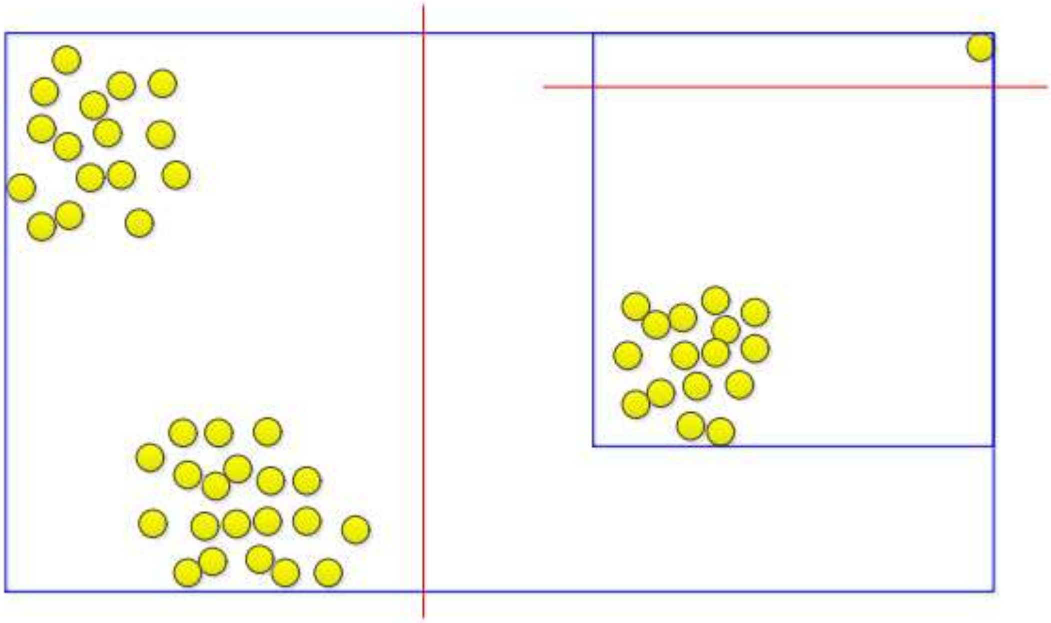


Figure 2.11: Random Cut Tree [55]

Deletion (ForgetPoint Algorithm 5) [55] and insertion (ForgetPoint Algorithm 6) [55] operations (**Appendix I**) can be used to dynamically maintain robust random cut trees when anomalies on data are detected using RRCF. For deletion: If T were drawn from the distribution $RRCF$ (S) then the *ForgetPoint* algorithm produces a tree T' which is drawn at random from the probability distribution $RRCF$ $(S - \{p\})$. On the other hand, for insertion**:** Given T drawn from distribution $RRCF$ (S) and $p \in S$ produce a T' drawn from $RRCF$ $(S \cup p)$, the *InsertPoint* algorithm is used.

2.10 The Research Gap

The extensive literature indicates that the current water quality monitoring methods are based on outdated procedures. These include traditional manual lab-based methods that are time costly and prone to data alteration or interference during packaging and transfer of the samples. Similarly, the traditional manual in-situ procedures aren’t better off. Therefore, evolving the current water quality monitoring methods is inevitable.

Wireless sensor networks (WSNs) have gained popularity in the current industry as pertains to the Internet of Things (IoT), including the application in water quality management. However, the choice of a wireless transmission technology has posed a challenge in terms of several aspects, including power consumption, the ease of use, the cost of acquisition, the distance of coverage, and the amount of data transmitted, among many other factors.

Real-time anomaly detection on time-series data has also made the utilization of advanced machine-learning (ML) algorithms possible. Many current anomaly detection algorithms exist for various applications, and a lot has not been done regarding water quality parameters. Some of them have false anomaly detection, very high computation complexity, and too much processing time. Therefore, there is a need to determine an effective algorithm regarding water quality contamination event detection.

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter describes the design of a WQM system using WSNs. The system design followed four major phases outlined below.

i. LoRa connectivity and range evaluation.

ii. Sensors calibration and integration with the system.

iii. General system designing, fabrication, testing, and deployment.

iv. Anomaly detection algorithm determination.

3.1 LoRa Connectivity and Range Evaluation

3.1.1 Statement of Purpose

An experimental evaluation of the proprietary parts of LoRa Technology was conducted to ascertain if LoRa performs as advertised. This procedure's aims were two-fold: To conduct performance experiments on LoRa connectivity and range evaluation for wireless sensor networks and present and discuss the results obtained for purposes of the developed WQM system. The experiments relied on the received signal strength indication (RSSI); which refers to the signal power received in dBm. How clear a receiver can "hear" from a sender can be measured using this value. The value range of typical LoRa RSSI is -120 dBm to -30dBm.

3.1.2 Measurement Setup

The outlined parameters were measured at the Dedan Kimathi University of Technology, Kenya, at different times throughout the day over several days. The university is located in a rural area, and the highest residential buildings are four (4) floors high. The site has an irregular terrain, with notable differences in geographical elevation. The base station remained stationary all through the measurements. End devices that sent payloads periodically to the base station were deployed at different locations away from the base station. These locations were 100m apart, at a 1km path range along a line of sight (LoS) from a 2.5m stand node, as shown in Figure 3.1. For every transmitted payload, there was a measure of the received signal strength (RSSI) used in the connectivity and range of evaluation studies herein.

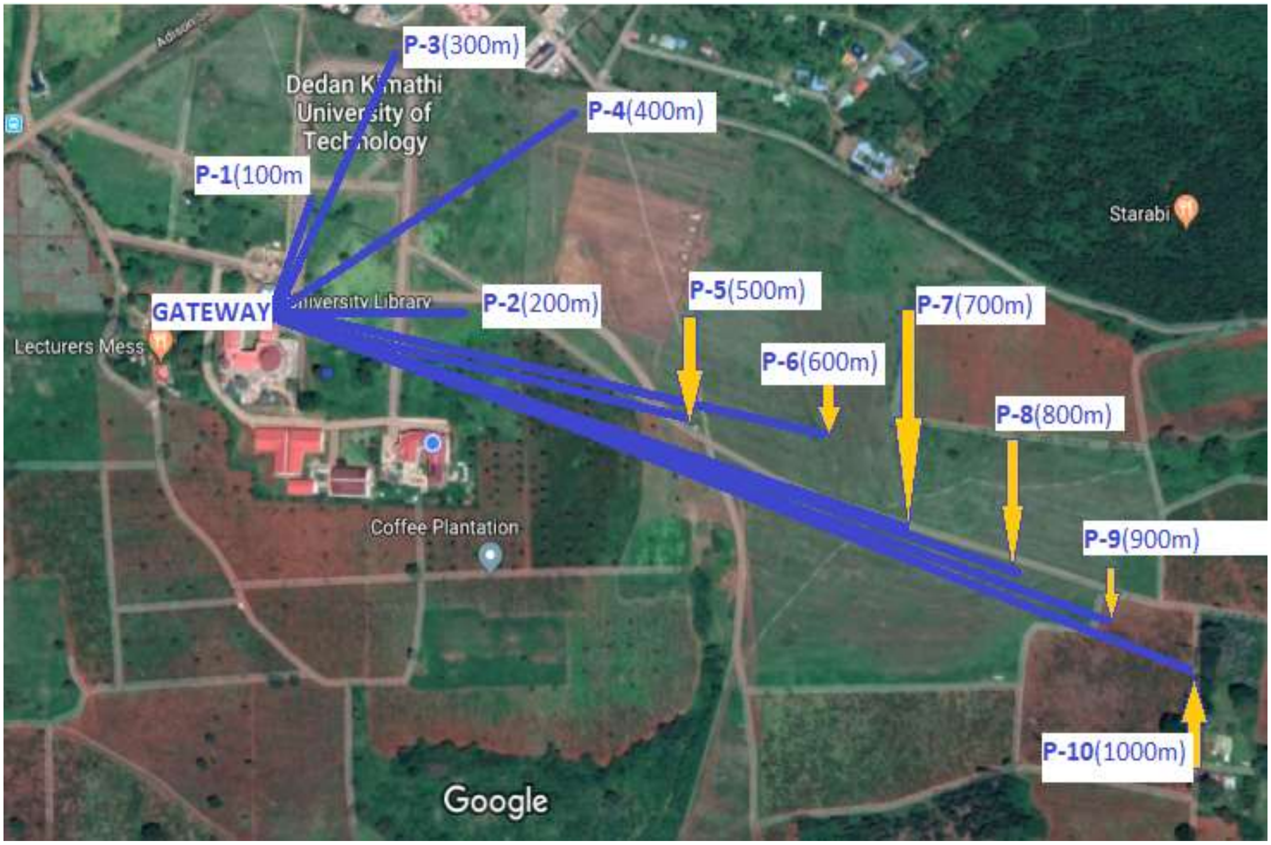


Figure 3.1: Test points geographical locations. [Extracted from Google Maps]

I. Base Station

The configured and installed LoRaWAN industry gateway at the Dedan Kimathi University of Technology was used (Figure 3.2). The gateway's location is approximately 25 meters above the ground, on the roof of a centrally situated building (*The Resource Center*) at DeKUT. It is based on the MultiTech Conduit, a quickly deployable and programmable gateway. It is also intended for the internet of things (IoT). Moreover, it is suitable for both public and private LoRaWAN projects. Tables 3.1 and 3.2 summarize the specifications and operations of the gateway.

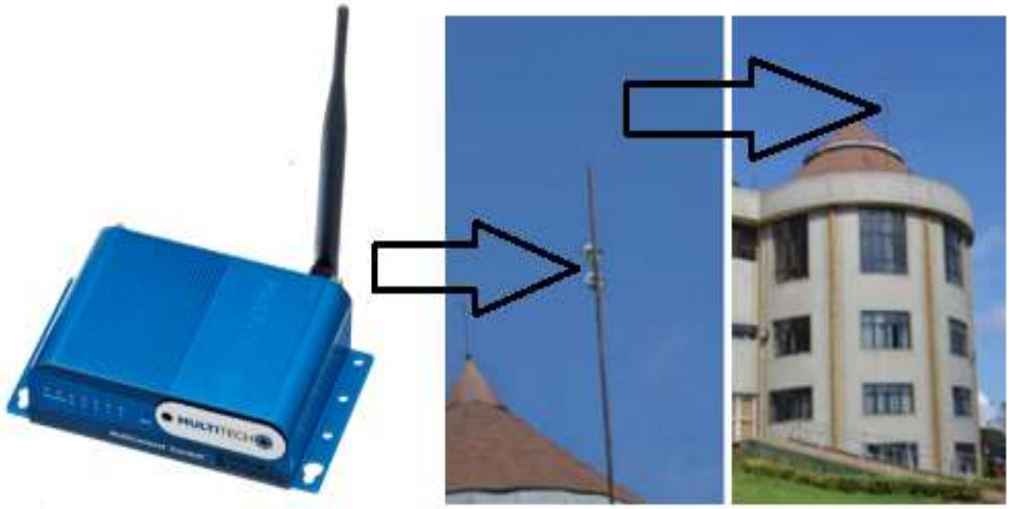


Figure 3.2: The LoRaWAN industry gateway (based on the MultiTech Conduit)

Table 3.1: The LoRaWAN Industry Gateway Specifications (Subject To Environmental Factors and Placement of Nodes/Sensors and Gateways)

Antenna	LoRa Female SMA, Cell 2dBi	27dBm max output
Connectivity	Ethernet (RJ_{45})	Optimal $_{3}$FF Micro SIM
Enclosure	Size (161 mm by 107mm by 42mm)	Weight 1.45kg

Table 3.2: The LoRaWAN Industry Gateway Operation (Subject To Environmental Factors and Placement of Nodes/Sensors and Gateways)

Operating Temperature	Min: -30 °C	Max: +70 ℃
Communicating Range	Line of sight(*Antenna): 20kms	Urban: up to 3kms
Installation	Wall or Desktop mount	Power 9V UK/EU

II. End Device

The end device was an STM32 Nucleo board (Figure 3.3) equipped with a LoRaWAN Transceiver

Shield. While taking the measurements, the nodes were powered by 3V batteries. The transmit power was +14 dBm at a frequency of 868MHz. The node was attached to a stand, approximately 2.5 m high from the ground level for on-ground measurements. Each device was registered on The Things Network (TTN) platform, which forwarded data to the database for storage after retrieving it from the device. The security of data transmission was ensured by TTN, which provides credentials for device authentication. Each node periodically transmitted a payload during measurements, including the received signal strength (RSSI) to the base station. Payloads were sent every 60 seconds for one hour in each test location.

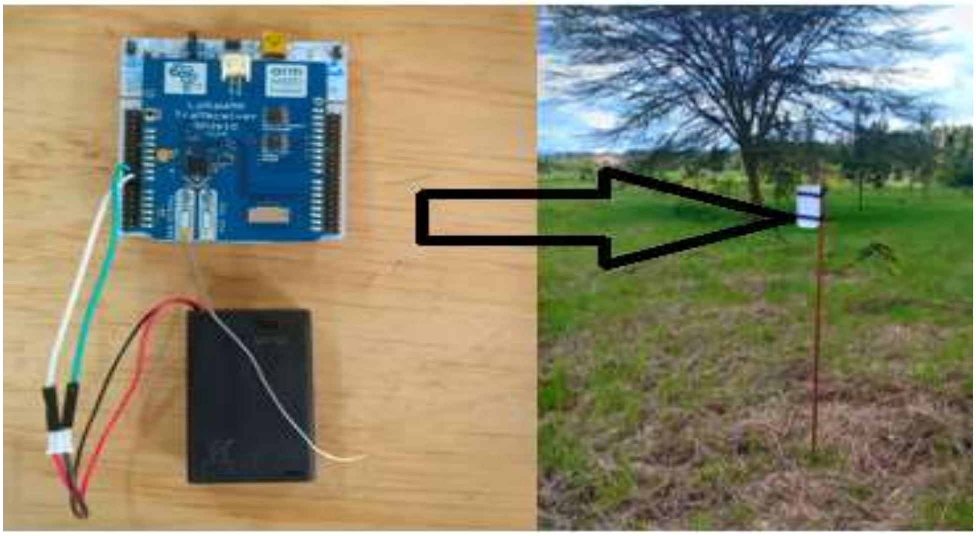


Figure 3.3: STM32 Nucleo board, equipped with a LoRaWAN transceiver shield

3.2 Water Quality Parameters

The parameters of turbidity and pH and their significance to water quality are described in the next section. Calibration techniques for the pH and turbidity sensors are also described. Significant consideration was given to obtaining linear responses, mitigating noise, and achieving high accuracy and quality resolution. A lab assessment (utilizing standard buffer solutions and reference instruments) was conducted at the NYEWASCO water treatment plant laboratory at Kamakwa, Nyeri.

3.2.1 Turbidity

The DFRobot Gravity Arduino turbidity sensor (Figure 3.4) was used to detect the opaqueness levels of water. It utilized light to sense the suspended solid particles that affect the transmission and scattering of light. The sensor provided both analog and digital signal modes (with adjustable threshold), and it operates at a voltage of 5V DC and a maximum of 40 mA. A temperature range of between 5°C and 90°C proves ideal for this sensor. Its response time is 500ms and has a resolution of 0.01V analog output voltage.

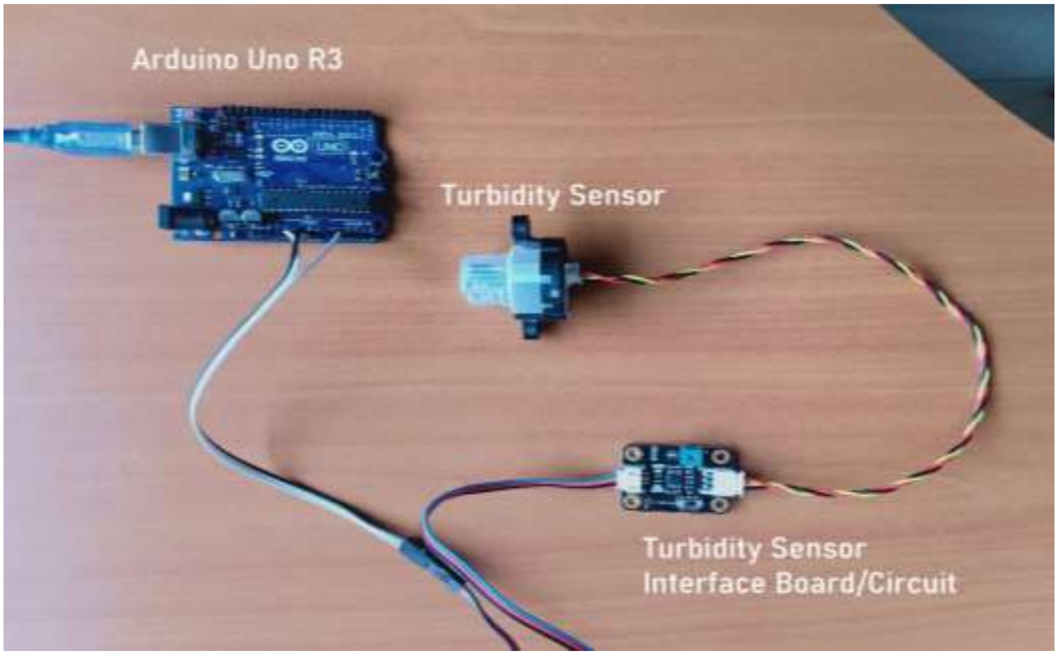


Figure 3.4: The DFRobot Gravity Arduino turbidity sensor interfaced with Arduino UNO R3 Board

For its calibration, the primary aim was to get a voltage value from the sensor and transform it into turbidity information. The module was connected to the Arduino Uno R3 board using three pins only: VCC (data), GND (ground), and SIGNAL (data). The sensor has both a light transmitter and a receiver. Where the waters are clear, the dispersion of light is recorded as the least while the recipient of light receives the most light. With more turbidity, the light receiver gets less light progressively.

There is a switch (producing a hybrid of analog and digital signals) on the interface board, switching between analog and digital modes. The official DFRobot's wiki notes that the sensor yields diminished values in analog mode. At the same time, its output goes high if the sensor is in digital mode, reaching the threshold that the onboard potentiometer has established. The analog mode is preferred to the digital mode in the measurement of the turbidity levels. The turbidity and voltage follow the following relationship; where TU is the turbidity and V is the voltage.

$$TU = 1120.4 * V^2 + 5742.3 * V - 4352.9 \qquad 3.1$$

Nevertheless, the equation described above is only suitable for the sensor if it produces 4.2 volts at zero turbidity (0 NTU). This sensor did not produce a voltage of 4.2 V, and thus, the sensor probe was opened, and the trimmer (Figure 3.5) tuned to obtain 4.2 V with the sensor afloat.

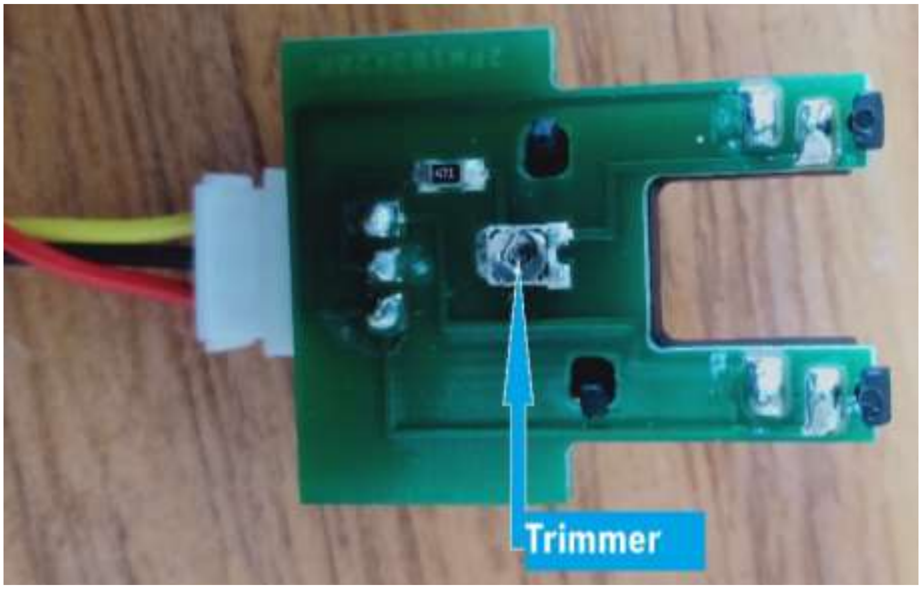


Figure 3.5: The DFRobot Gravity Arduino turbidity sensor Trimmer

After that, the voltage was converted into NTU. Despite that, the equation from the shown graph only applies when the voltage ranges between 2.5V and 4.2 V. To ascertain the provided equation, standard solutions (*0, 20, 40, 100, 200, 1000, and 4000 NTUs*) available at NYEWASCO water quality laboratory were used to carry out verification. Fifteen (15) trials were done for each standard solution. Consequently, there was a need to set limits as 1000 NTU for voltages below 2.5 V to be the sensor's highest possible NTU value attainable. This exercise was conducted on Wednesday, August 2020. At the time of the study, the room thermometer showed a temperature of 22° C. Using a turbidimeter (Figure 3.6), that is calibrated by the Kenya Bureau of Standards (KEBS), as a primary instrument, the turbidity of these standard solutions was measured and the experiments repeated with this probe sensor.

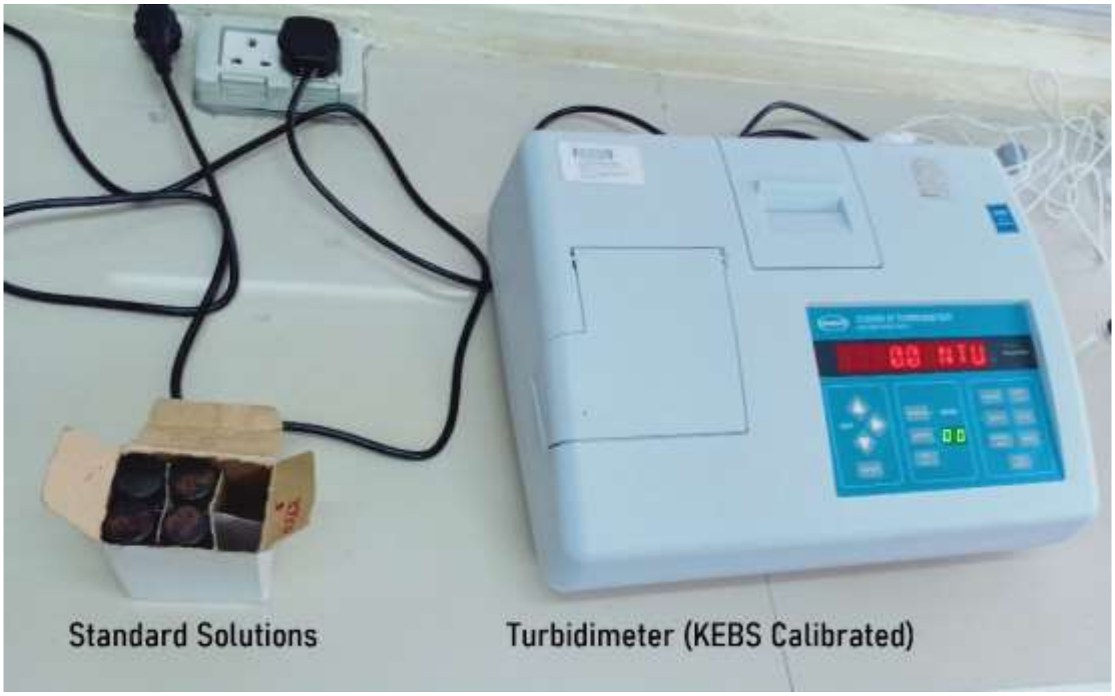


Figure 3.6: The KEBS calibrated turbidimeter

3.2.2 pH

The DFRobot’s Gravity Analog pH sensor to gauge the solution pH and mirror its acidic or basic was deployed for this framework (Figure 3.7). Its activity voltage ranges between 3.3 to 5.5V, with an accuracy of ±0.1 at 22℃; recognition scope of 0 to 14, and activity temperature range between 5 and 60°C. Its response time is stipulated to be one minute and a resolution of 0.01.

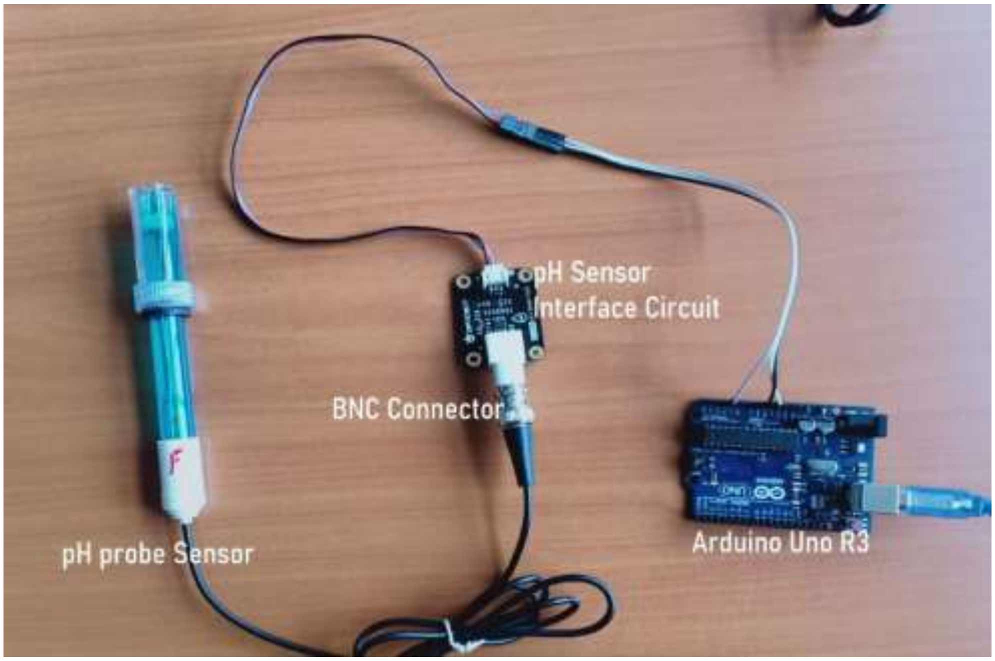


Figure 3.7: DFRobot's Gravity Analog pH sensor interfaced with Arduino UNO R3 Module

The manufacturer directs that there must be a two-point calibration, and as such, there should be two buffer solutions (4.0 and 7.0) to be used as the standard solutions. Consequently, the two-point calibration followed the following guidelines;

i. A calibration code was uploaded to the Arduino UNO R3 board, and the serial monitor opened on a laptop computer.

ii. The probe was washed using distilled water, and allowed to dry. Then, the probe used to measure pH was put in a 7.0 buffer solution, stirred thoroughly until stable values were obtained.

iii. Once the stable values were attained, the first point was marked using the *ENTER* instruction in the serial monitor on the computer to set it to calibration.

iv. The calibration (CAL) instruction was used to calibrate more inputs in the serial monitor. The program then identified the buffer solution 7.0 to be present.

v. The EXIT command was used to move out of the calibration mode. This command also allowed the data entered in the serial monitor to be saved.

The *EXIT* instruction marked the end of the calibration steps, and the first point calibration was over. The second-point calibration followed a similar pattern to the first-point calibration, but it used the 4.0 buffer solution. After that, the pH of three standard solutions (4, 7, and 9) was obtained using the KEBS calibrated pH meter alongside this sensor probe. Fifteen (15) trials were done for each standard solution. This exercise was conducted at the NYEWASCO water quality laboratory on Wednesday, August 2020, at a room temperature of 21℃, as depicted in Figure 3.8 for result validation purposes.

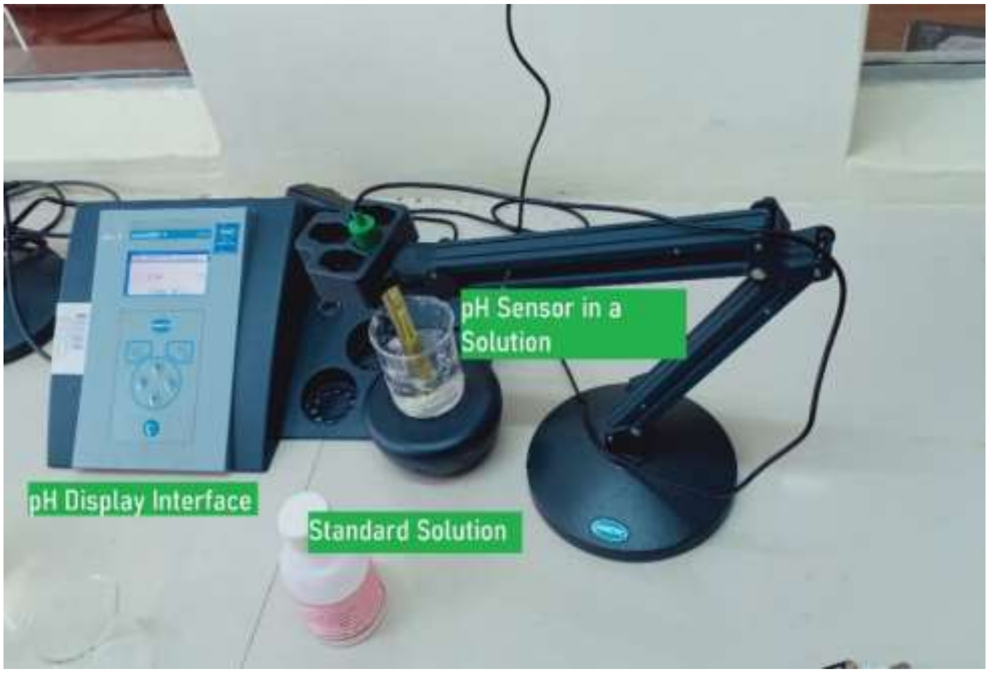


Figure 3.8: KEBS Calibrated pH meter at NYEWASCO Water Quality Lab

3.3 Sensor Node Designing and Fabrication

Printed circuit boards (PCBs) play a very crucial role in the development of microcontroller systems. They easily allow circuits to be realized with the minimum number of connectors resulting in the optimization of the occupied space on a fabricated PCB. For this prototype, the STM32 Nucleo F466RE microcontroller was used. The schematic and layout of the STM32 Nucleo F466RE microcontroller board were prepared using the KiCad 4.0.7 software, a suite for electronic design automation (EDA). For electronic circuits and their conversion to PCB designs, facilitation of schematics design is easy with this software. An integrated environment for schematic capture and

PCB layout design is a common feature with it. It also has tools that enable the creation of a bill of materials, sketch designs, artworks, and 3D views of the PCB and its components, among others. Figure 3.9 shows the schematic design of the sensor node board. The circuit comprises the following functional blocks; the microcontroller, the L293D IC, a power supply circuit, and the sensor connectors (Power, Ground, and Data). Figure 3.10 shows the PCB outcome design in 2D view, putting into the actual fabrication result.

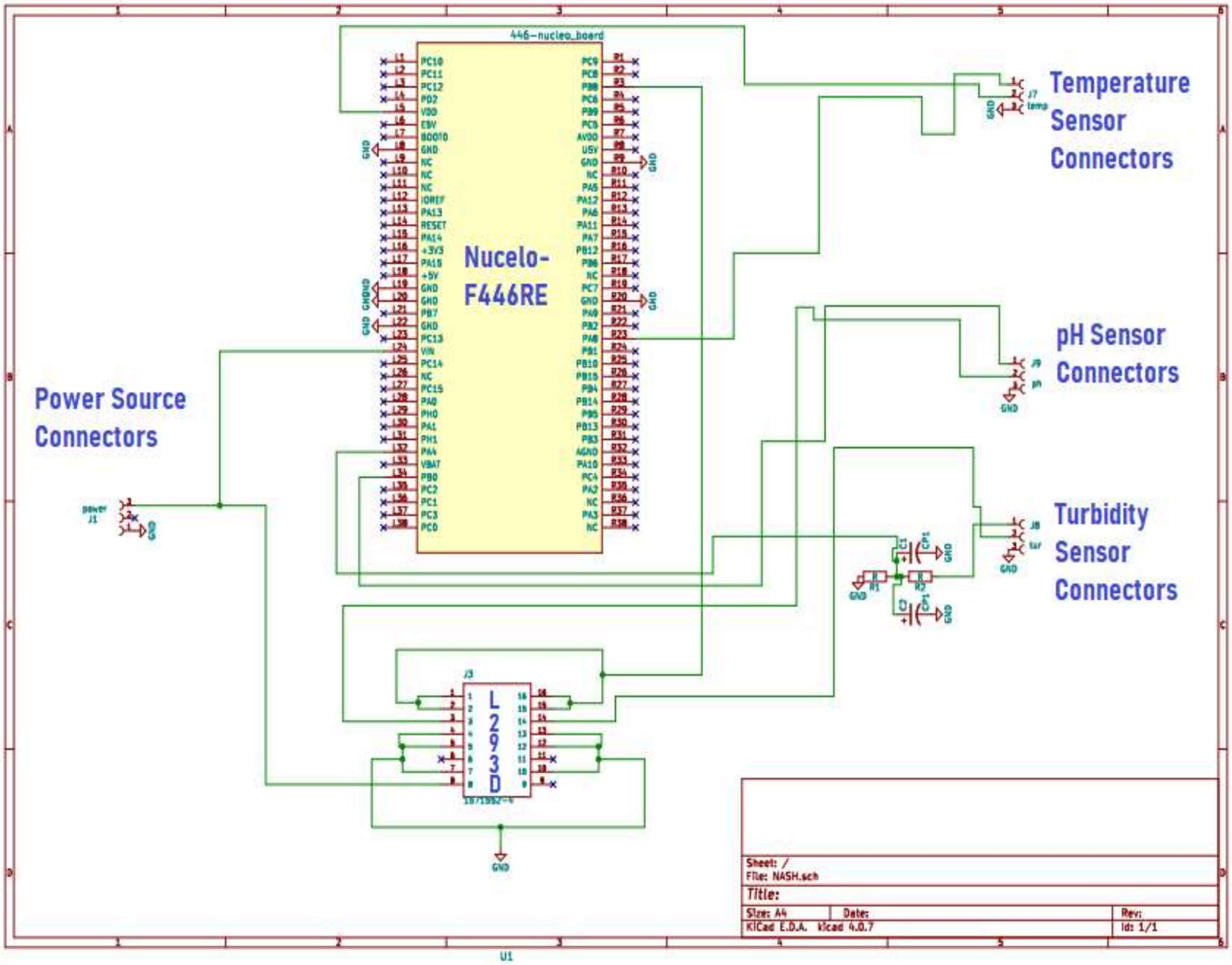


Figure 3.9: The schematic design of one sensor node board

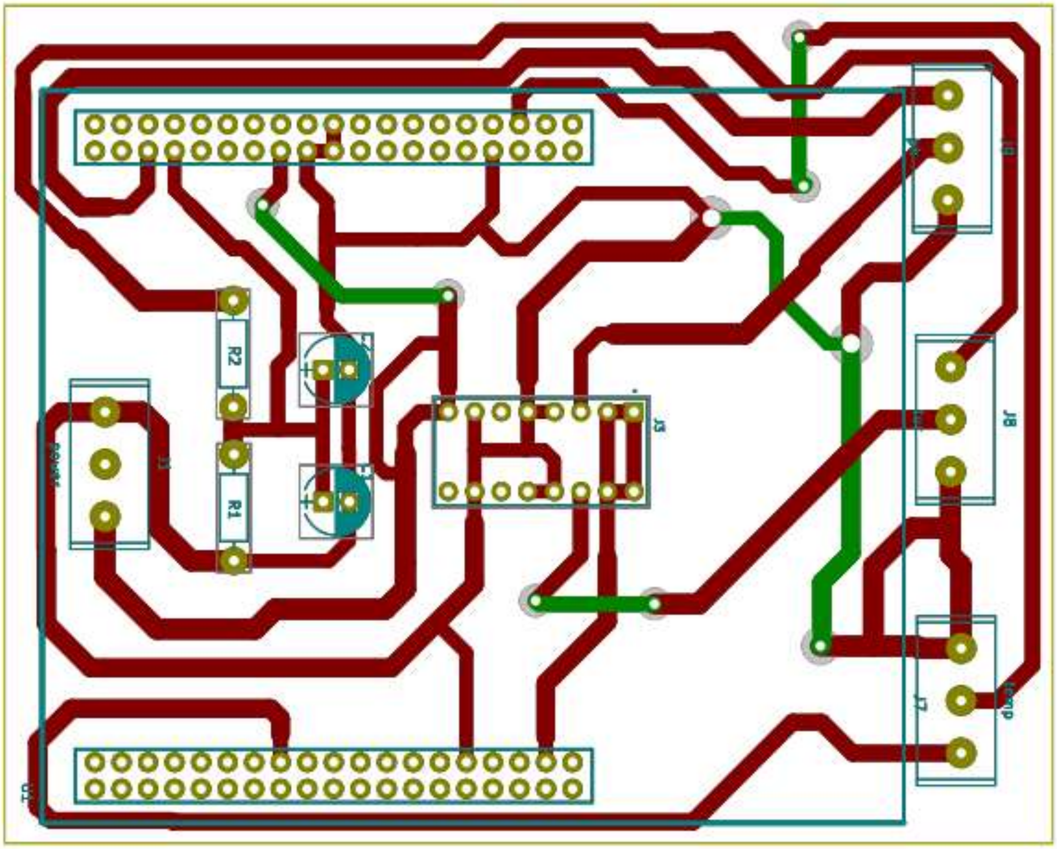


Figure 3.10: The PCB Design of the sensor node board

3.4 System Energy Management

The general power utilization includes the focal measurement sensor hub and the LoRaWAN transceiver module that transmits water quality data. The node operates at about 50mA at 5V working voltage per minute. These sensors expend large amounts of power. Many coordinated circuits, including the STM32 Nucleo board, do not adequately supply the power of such intensities. Interfacing these sensors and the Nucleo board power pins directly and constantly may harm it. An H-Bridge motor control circuit utilizing the L293D Motor Driver IC (Figure 3.11) to connect the sensors and the Nucleo board was deployed [56]. This also enables shutting down the sensor modules when the microcontroller has no data reading, which is equally set to sleep mode for 58 minutes within an hour of operation.

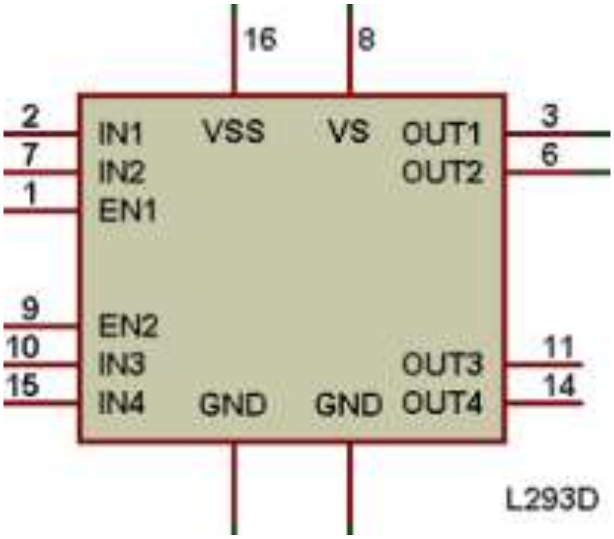


Figure 3.11: the L293D Motor Driver IC Pin-out Diagram

3.5 Data and the Analytical Tool Determination

In this section, the techniques discussed in the previous chapter were evaluated thoroughly using the Jupyter Notebook. In this open-source web application, one can create and share documents containing computer code (Python) and text elements (figures, equations, and other files). Since this thesis is based on the concept of reproducibility, links to the resources used in this section are as follows:

- Anaconda was employed to set up the environment for all analytical experiments in this research.

https://www.anaconda.com

- Jupyter, installed in Anaconda. Documents shared for experiments with embedded Python code, visualizations, and explanatory markdown were created here.

https://jupyter.org

- Python 3.7, the programming language used to build the codes used.

https://www.python.org

- Pandas, which was used for data retrieving and handling in Jupyter Notebook, is a data analysis library in Python.

https://pandas.pydata.org

- Matplotlib is used to make figures in Jupyter Notebook, a Python plotting library.

https://matplotlib.org

- Scikit-learn Python machine learning library was used in evaluation experiments. It is a free open source library, and it provides simple and effective tools for data analysis.

https://scikit-learn.org

- Package eif, for the extended isolation forest algorithm Python library.

https://github.com/sahandha/eif

- Package rrcf is a Python execution of the robust random cut forest algorithm for anomaly detection.

https://github.com/kLabUM/rrcf

First and foremost, a web-based application linked to a Google Cloud Protocol (GCP) Console-based InfluxdB database Virtual Machine (VM) was developed using the Dash *Plotly* framework to analyze and visualize the data collected in real-time. Properties of the dataset are shown in table 3.3.

Table 3.3: Properties of the turbidity and water pH dataset

File Name	Data_Raw_Water.csv
Location	Nyeri-Kenya, Kamakwa, NYEWASCO Treatment Plant
Interval	Every 30 Minutes
Number of Records	2568
Columns	time, turbidity, pH
Data Range	2020-11-04 11:00:31.822439+00:00 to 2021-01-04 09:54:25.214766+00:00

A subset of the dataset, with 291 records, was extracted considering a region with graphically notable anomalies and used as the ground truth. A section of the dataset between 11th and 18th November 2020 (7 days) was used. It was then manually examined, and all the outlier instances were identified. Using this subset, four anomaly detection evaluation experiments were performed for each parameter (turbidity and pH), first using the LOF algorithm, the IF and the EIF algorithms, and the RRCF algorithm. The experimental procedures are summarized in the flowchart of Figure 3.12 below.

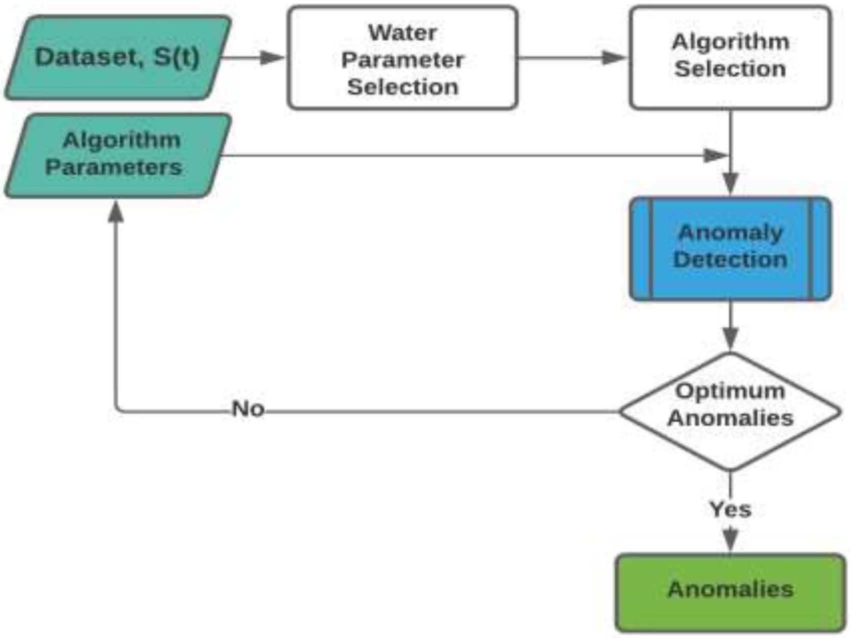


Figure 3.12: Experimental evaluation for the anomaly detection algorithms

3.6 General System Design Overview

There was the employment of a holistic, modular approach to creating this system. The organization of the developed and the deployed system is shown in Figure 3.13. The system is composed of the major sections of the sensors, the STM32 Nucleoboard microcontroller, and the LoRa Transceiver.

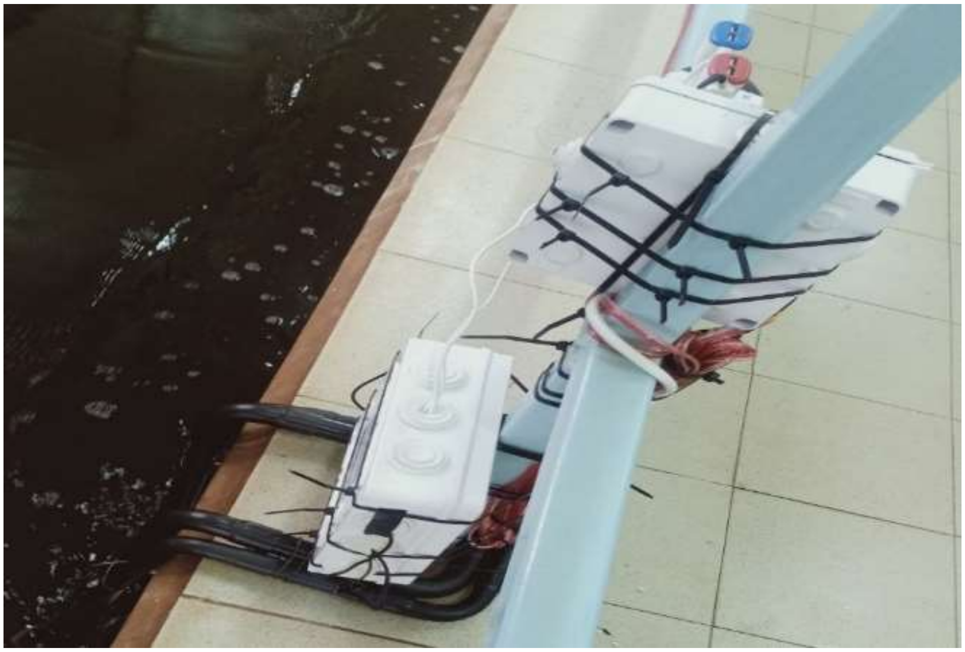


Figure 3.13: A Picture of the sensor node of the WQM System deployed at NYEWASCO

The general operation of the system starts with reading sensor values and terminates with anomaly detection, as shown in Figure 3.14 below.

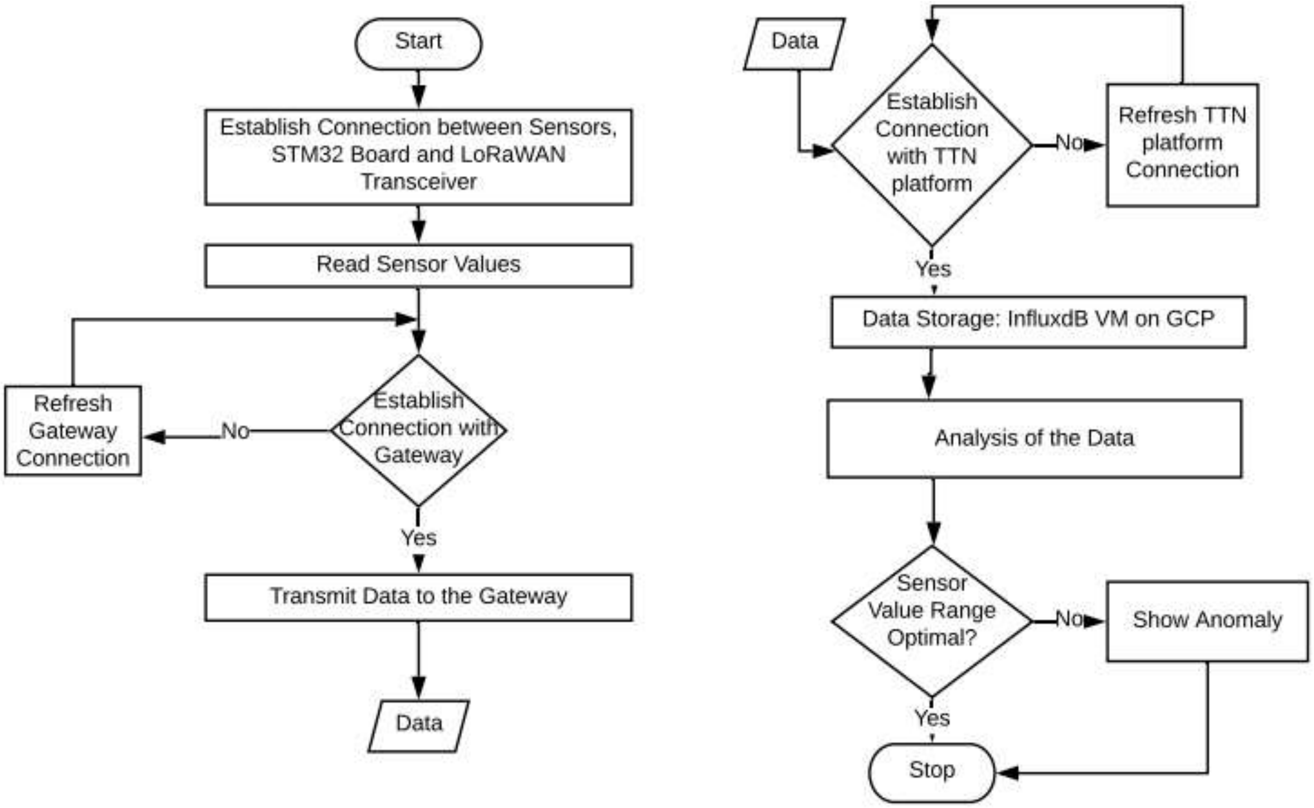

Start
Establish Connection between Sensors, STM32 Board and LoRaWAN Transceiver
Read Sensor Values
Establish Connection with Gateway
No
Refresh Gateway Connection
Yes
Transmit Data to the Gateway
Data
Data
Establish Connection with TTN platform
No
Refresh TTN platform Connection
Yes
Data Storage: InfluxdB VM on GCP
Analysis of the Data
Sensor Value Range Optimal?
No
Show Anomaly
Yes
Stop

Figure 3.14: General system operation flow chart

CHAPTER FOUR

RESULTS AND DISCUSSION

In this chapter, the results and the discussion of the developed system are presented as follows.

i. Performance experiments on LoRa connectivity and range evaluation for wireless sensor networks in the rural area around the DeKUT.

ii. The WQM sensor node was developed to collect values of two water parameters: pH and turbidity.

iii. The obtained data was analyzed based on the selected machine learning algorithms: IF, EIF, LOF, and RRCF.

4.1 LoRa Connectivity and Range Evaluation

For this research, the mean RSSI was computed for each of the 10 test locations used. At 100m away from the gateway, a mean strength of -102.7 dBm was recorded, while at 200m, the mean signal strength was -106.5 dBm. A complete record of the computed mean RSSI values for every test location is shown in Table 4.1. The best strength was realized at test location 3 (300m), while it is notable that the RSSI decreased (worsened) as the distance from the gateway increased.

Table 4.1: The Mean RSSI (in dBm) for the Ten (10) Test Locations

Test Location (m)	100	200	300	400	500	600	700	800	900	1000
Mean RSSI (dBm)	-102.7	-106.5	-96.3	-100.5	-109.6	-108.2	-111.9	-112.2	-112.2	-113.1

The box plot in Figure 4.1 provides a quick graphical examination of the RSSI for each of the ten (10) data sets. Outlier RSSIs were realized in test locations 3, 4, 8, and 9, and they are plotted as individual points. The highest notable degree of dispersion (spread) and skewness in the RSSI is observed with test locations 1, 2, 4, and 5, whereas test location 8 depicts the contrary.

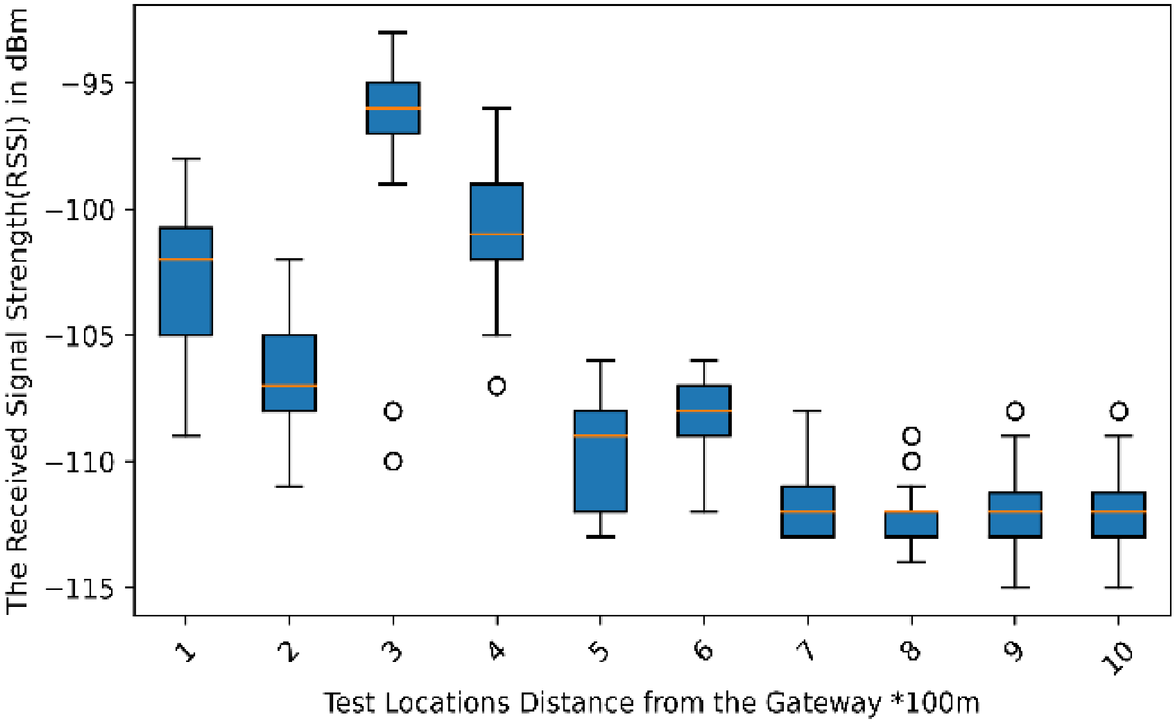


Figure 4.1: The Received Strength Whisker and Box Plots for the 10 Test Locations

Therefore, the use of LoRa technology that provided a low power long-range connectivity and good connectivity is proved to be an appropriate method compared to the current systems that mainly rely on WiFi technology. While LoRa technology covers up to a distance of up to 1km as demonstrated herein, WiFi coverage usually is up to 200km only.

4.2 The WQM Sensor Network System Development

This section presents the results of the WQM sensor network system developed in the vital area; sensor calibration.

4.2.1 The pH Sensor

For the three standard solutions, mean values of both the pH meter and our sensor probes are tabulated in Table 4.2. There were notable differences with less significant margins.

Table 4.2: pH validation results for the three standard solutions

Standard Solution	**4.0**	**7.0**	**9.0**
pHmeter Values(*KEBS Calibrated*)	4.11	7.12	9.14
Probe Sensor Mean Values (n=15)	4.08	7.25	9.15

The probe was ascertained to have been well-calibrated from a plot of Figure 4.2 that gave a correlation coefficient approximately equal to one.

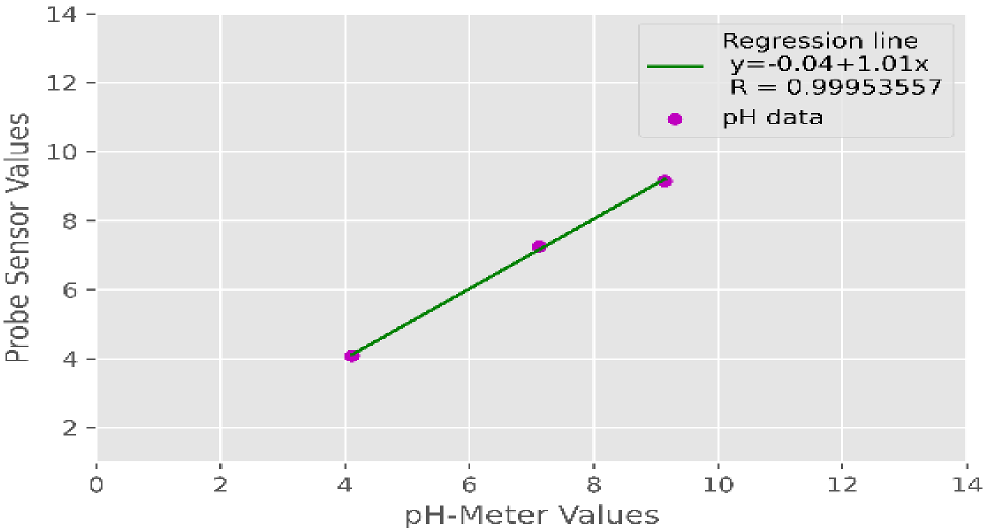


Figure 4.2: A plot of probe sensor mean values against the pH-meter mean values

4.2.2 Turbidity Sensor

During validation of the Gravity DFRobot Arduino sensor using a turbidimeter as a primary instrument, the results obtained are tabulated in table 4.3.

Table 4.3: Turbidity Validation Results using the 7 Standard Solutions

Standard Solution (NTU)	**0**	**20**	**50**	**100**	**200**	**1000**	**4000**
Turbidimeter Values (NTU)	1.1	20.3	49.1	102.3	205.2	1008.7	4022.8
Probe Sensor Mean Values (n=15) (NTU)	0.9	20.2	48.8	101.5	202.5	1004.2	3000

The probe was ascertained to have been well-calibrated from a plot of the results (figure 4.3) that gave a correlation coefficient approximately equal to one.

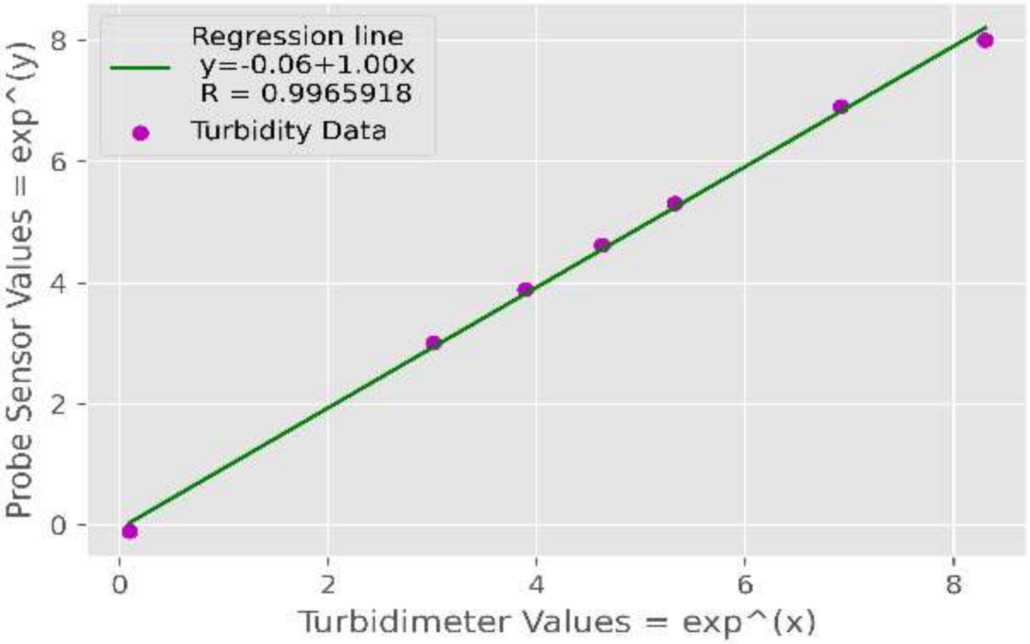


Figure 4.3: A plot of probe sensor mean values against the turbidimeter mean values

These sensors: The DFRobot Gravity Arduino pH and Turbidity probes indeed provided precise measurements with desirable resolutions and accuracy. Being power-hungry was their drawback that was considerably addressed by the 'sleep time' methods using the motor drive IC. Their quick response time was also crucial in the achievement of this compensation and power use minimization procedures.

4.3 Anomaly Detection and Machine Learning

This section presents the results of parameter analysis towards the determination of an efficient contamination event detection algorithm.

4.3.1 Performance Evaluation based on a Subset

A. Turbidity Dataset

Table 4.4 below shows a subsection of the 2,658 records of the water turbidity data in NTUs collected in 60 days.

Table 4.4: Turbidity Dataset Subsection

	time	turbidity
0	2020-11-04 11:00:31.822439+00:00	21.063435
1	2020-11-04 11:01:22.124333+00:00	20.868153
2	2020-11-04 11:01:51.663062+00:00	20.584553
3	2020-11-04 11:02:29.373718+00:00	21.185328
4	2020-11-04 11:03:45.517010+00:00	21.063435
...	...	...
2653	2021-01-04 07:53:20.987423+00:00	10.611506
2654	2021-01-04 08:23:37.035804+00:00	17.975997
2655	2021-01-04 08:53:53.104009+00:00	17.734662
2656	2021-01-04 09:24:09.578901+00:00	15.094176
2657	2021-01-04 09:54:25.214766+00:00	14.611506
2658 rows × two columns		

A plot of this turbidity data against time in figure 4.4 below shows several contextual anomalies. The subset under evaluation is indicated in the round corner rectangle.

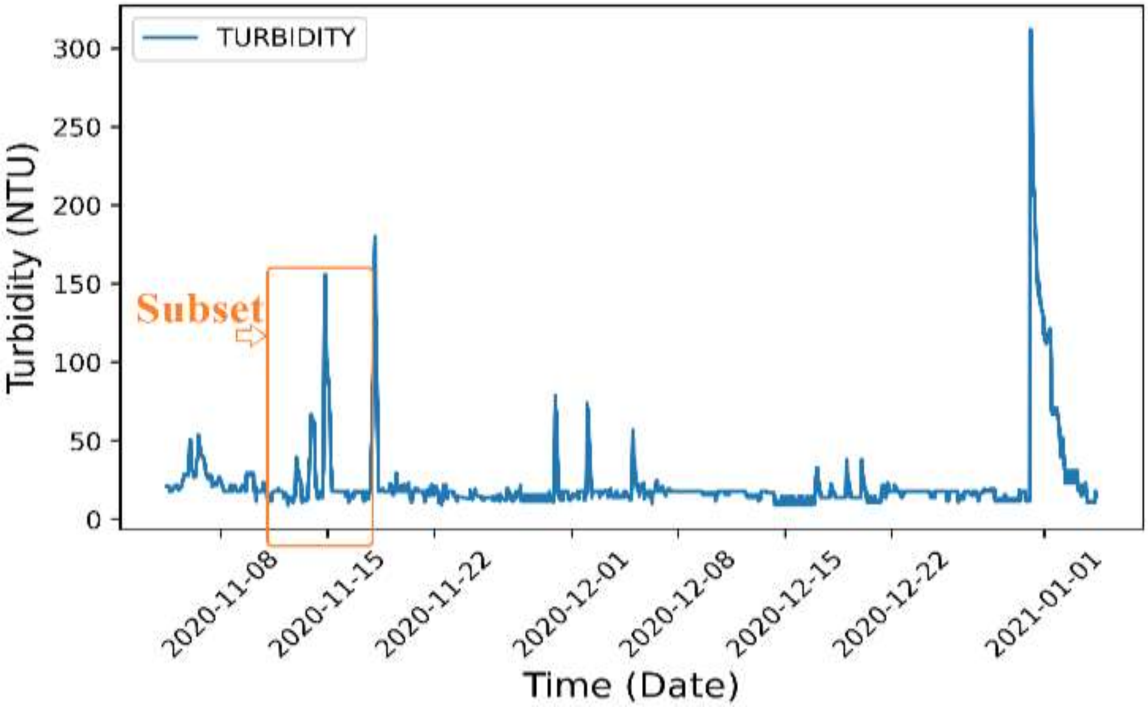


Figure 4.4: Turbidity Dataset for the Sixty (60) days

I. The Local Outlier Factor Algorithm

For the turbidity outliers, the algorithm detected a total of 75 outliers highlighted in Table 4.5. These anomalies are plotted as shown in Figure 4.4. The red stars are the 75 instances seen as anomalies in the turbidity data with the number of neighbors $k = 100$. It took the LOF algorithm 38.9 seconds to complete this process. Finding an optimal value of k was essential for detection performance. There were no false alarms as well as undetected outliers.

Table 4.5: Turbidity outliers for the subset data as detected by the LOF algorithm

time	**turbidity**
2020-11-12 21:57:18.752276+00:00	33.856159
2020-11-12 22:27:34.814333+00:00	35.975997
2020-11-12 22:57:50.858115+00:00	39.486692
2020-11-12 23:28:06.931813+00:00	38.856159
2020-11-12 23:58:22.991919+00:00	37.611506
... ...	
2020-11-17 21:57:09.848484+00:00	50.094176
2020-11-17 22:27:25.894124+00:00	59.856159
2020-11-17 22:57:41.959856+00:00	77.975997
2020-11-17 23:27:58.035168+00:00	88.856159
2020-11-17 23:58:14.065228+00:00	100.975997
75 rows × two columns	

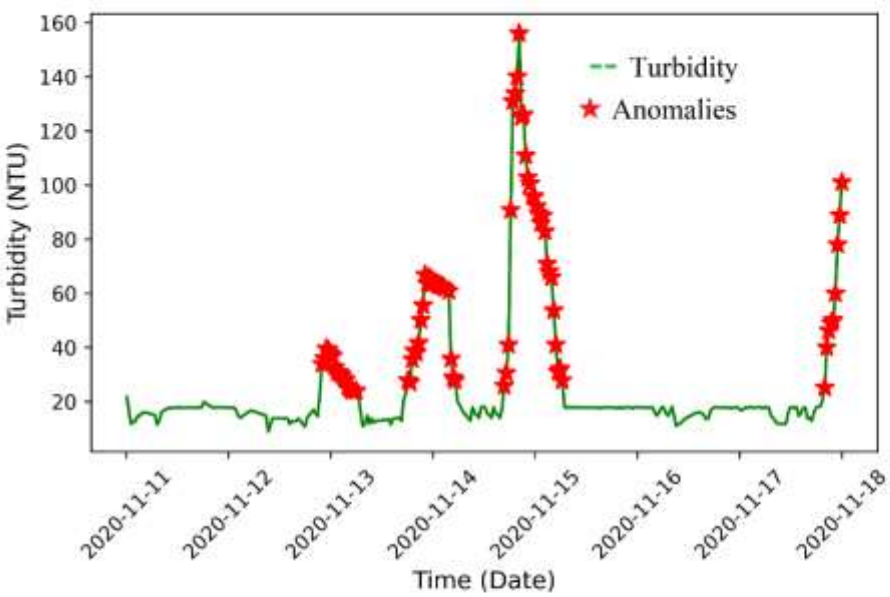


Figure 4.5: A plot of LOF turbidity outliers for the subset data

II. Isolation Forest and the Extended Isolation Forest Algorithms

In these algorithms, the sub-sampling size ψ controlled the size of the training data. It was determined that the *iForest* demonstrated high precision and reliability when the ψ was gradually increased to the desired value.

After the required value of $\psi = 200$ was achieved, there was no need to vary the ψ since it would unnecessarily increase memory consumption and time to process data. Moreover, it was observed that the number of trees *t* directly controlled the ensemble size. It was also found that the ideal paths converged at $t = 50$. When the tree training process was completed, several trees were returned, ready for the next evaluation stage.

Anomalies are always assigned scores by the IF and the EIF algorithms. Over 120 points for turbidity data were marked as anomalies above 0.6, and in this case, it wasn't easy to find a feasible threshold for improvement. For instance, most anomalies were considered inliers if 0.7 was set as the threshold score. However, several normal points were still considered anomalies when 0.65 was established as a threshold.

However, a plot of the top 60 instances (Figure 4.6) based on the score shows that standard EIF worked better than IF for turbidity data and found more anomalies with fewer false anomalies. This process took 3.66s for the IF algorithm and 3.99s for the EIF algorithm.

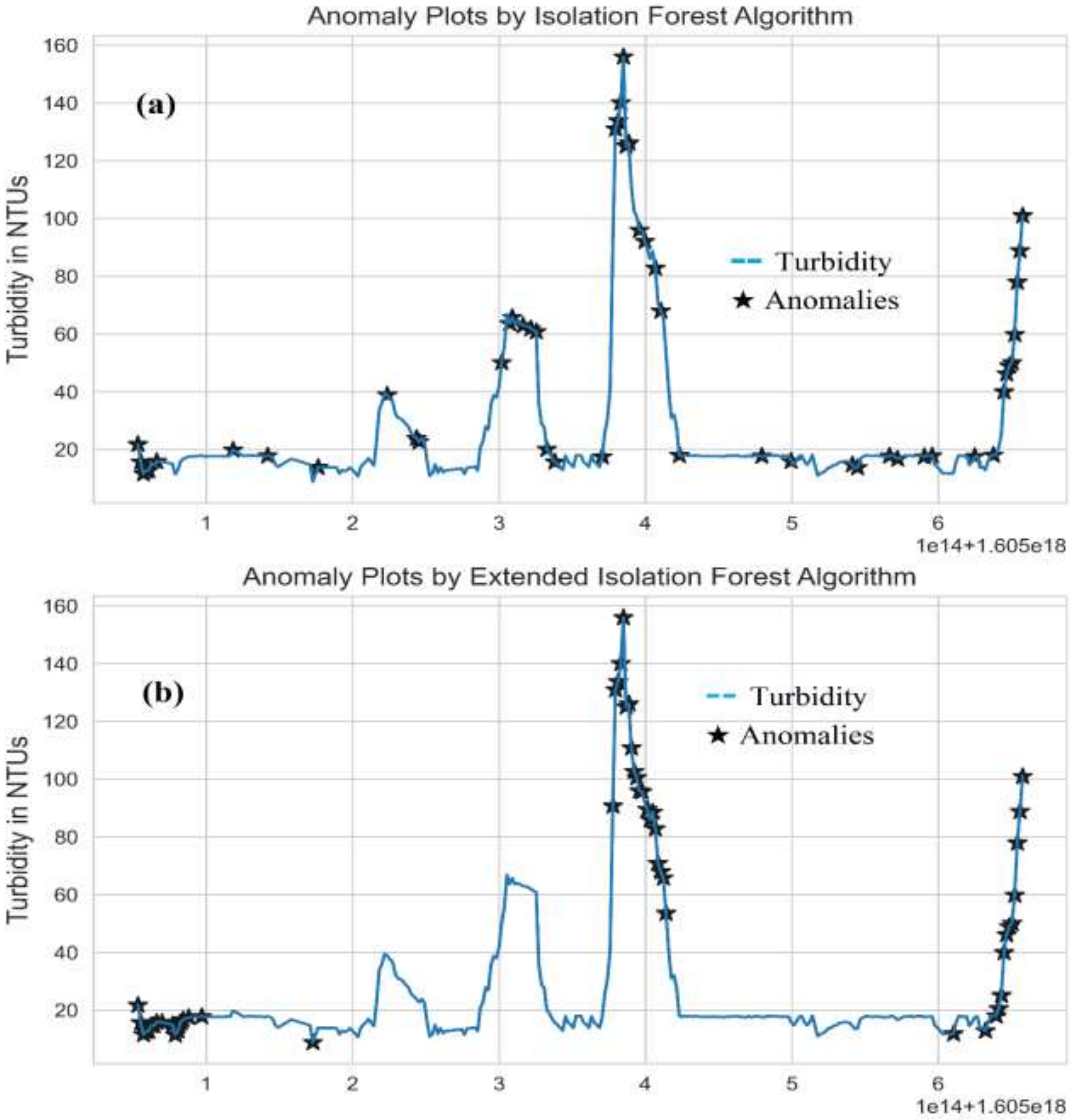


Figure 4.6: A plot of the IF and EIF turbidity outliers for the subset data

III. The Robust Random Cut Forest Algorithm

It was not easy to find a feasible threshold to split the outliers since some outliers were marked with low anomaly scores (for example, instances at the beginning of the dataset). In contrast, some regular points are marked with high anomaly scores. Therefore the top 67 outlier records having the highest scores were listed in Table 4.6, taking 7.1 seconds. These results were plotted as shown in Figure 4.7.

This algorithm did not detect a significant number of point outliers detected by the LOF algorithm. False alarms are also contained in this list: For example, for turbidity record 19.856159 NTU [2020-11-11 18:12:35.713854+00:00], whose value is almost equal to the next value (only 30 minutes apart) is marked as an outlier.

Table 4.6: Turbidity outliers as detected by the RRCF algorithm for the subset data

time	**turbidity**
2020-11-11 18:12:35.713854+00:00	19.856159
2020-11-12 09:20:37.353014+00:00	8.856159
2020-11-12 17:55:10.311827+00:00	10.734662
2020-11-12 21:27:02.699108+00:00	22.975997
2020-11-12 21:57:18.752276+00:00	33.856159
... ...	
2020-11-17 19:56:05.608199+00:00	25.094176
2020-11-17 20:26:21.676559+00:00	39.975997
2020-11-17 20:56:37.735037+00:00	46.210696
2020-11-17 22:57:41.959856+00:00	77.975997
2020-11-17 23:58:14.065228+00:00	100.975997
67 rows × three columns	

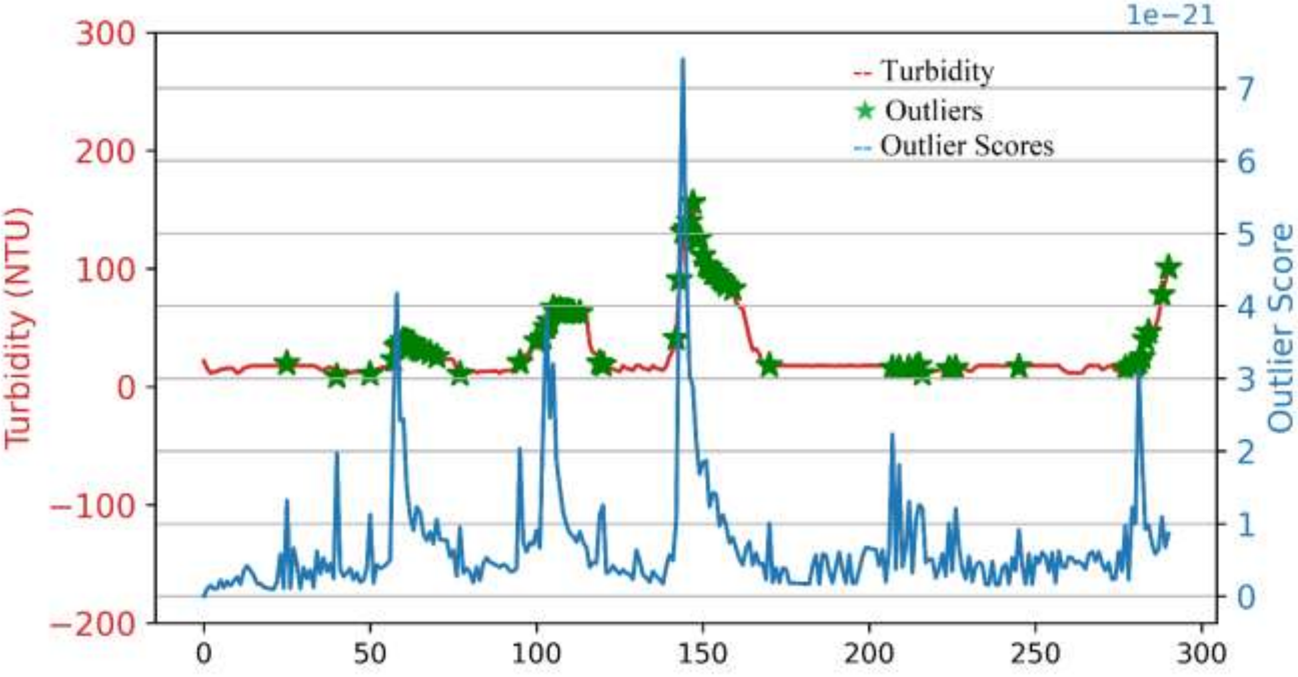


Figure 4.7: A plot of RRCF turbidity outliers for the subset data

B. pH Dataset

Table 4.7 below shows a subsection of the 2,658 records of the pH dataset that were collected in 60 days.

Table 4.7 pH Dataset Subsection

	time	pH
0	2020-11-04 11:00:31.822439+00:00	7.34
1	2020-11-04 11:01:22.124333+00:00	7.33
2	2020-11-04 11:01:51.663062+00:00	7.32
3	2020-11-04 11:02:29.373718+00:00	7.33
4	2020-11-04 11:03:45.517010+00:00	7.32
...	...	...
2653	2021-01-04 07:53:20.987423+00:00	7.35
2654	2021-01-04 08:23:37.035804+00:00	7.35
2655	2021-01-04 08:53:53.104009+00:00	7.34
2656	2021-01-04 09:24:09.578901+00:00	7.36
2657	2021-01-04 09:54:25.214766+00:00	7.36
2658 rows × two columns		

Similarly, the pH data in figure 4.8 below clearly shows that several instances are far from the other, which indicates point anomalies in the collected turbidity data. However, the range of variation does not portray a significant gap. The subset that was considered as the test data is shown in the round corner box.

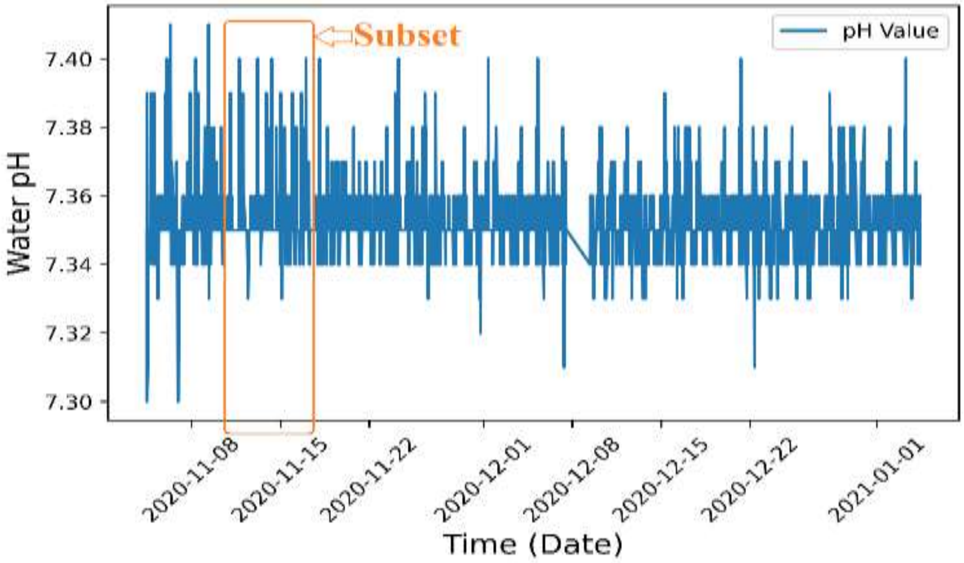


Figure 4.8: pH Dataset for the Sixty (60) days

I. The Local Outlier Factor Algorithm

The local outlier factor algorithm was used to detect the water pH outliers of the selected subset considered. In Figure 4.9, the red stars diagram shows the 63 instances detected as anomalies in the pH data using the number of neighbors, $k = 100$. The algorithm took 21 milliseconds to determine these anomalies. There were no false alarms as well as undetected outliers. Choosing an optimal *k* remained to be an essential factor for detection performance. For values of *k* too small or huge, the errors were prominent due to under-fitting.

Table 4.8: pH outliers as detected by the LOF algorithm on the subset data

time	**pH**
2020-11-11 01:11:34.389403+00:00	7.39
2020-11-11 03:42:54.674932+00:00	7.36
2020-11-11 18:12:35.713854+00:00	7.40
2020-11-11 23:15:16.261228+00:00	7.36
2020-11-12 00:15:48.373825+00:00	7.36
... ...	
2020-11-17 05:17:04.412038+00:00	7.37
2020-11-17 06:47:52.593118+00:00	7.34
2020-11-17 14:53:25.015789+00:00	7.36
2020-11-17 18:25:17.425846+00:00	7.36
2020-11-17 22:57:41.959856+00:00	7.34
63 rows × two columns	

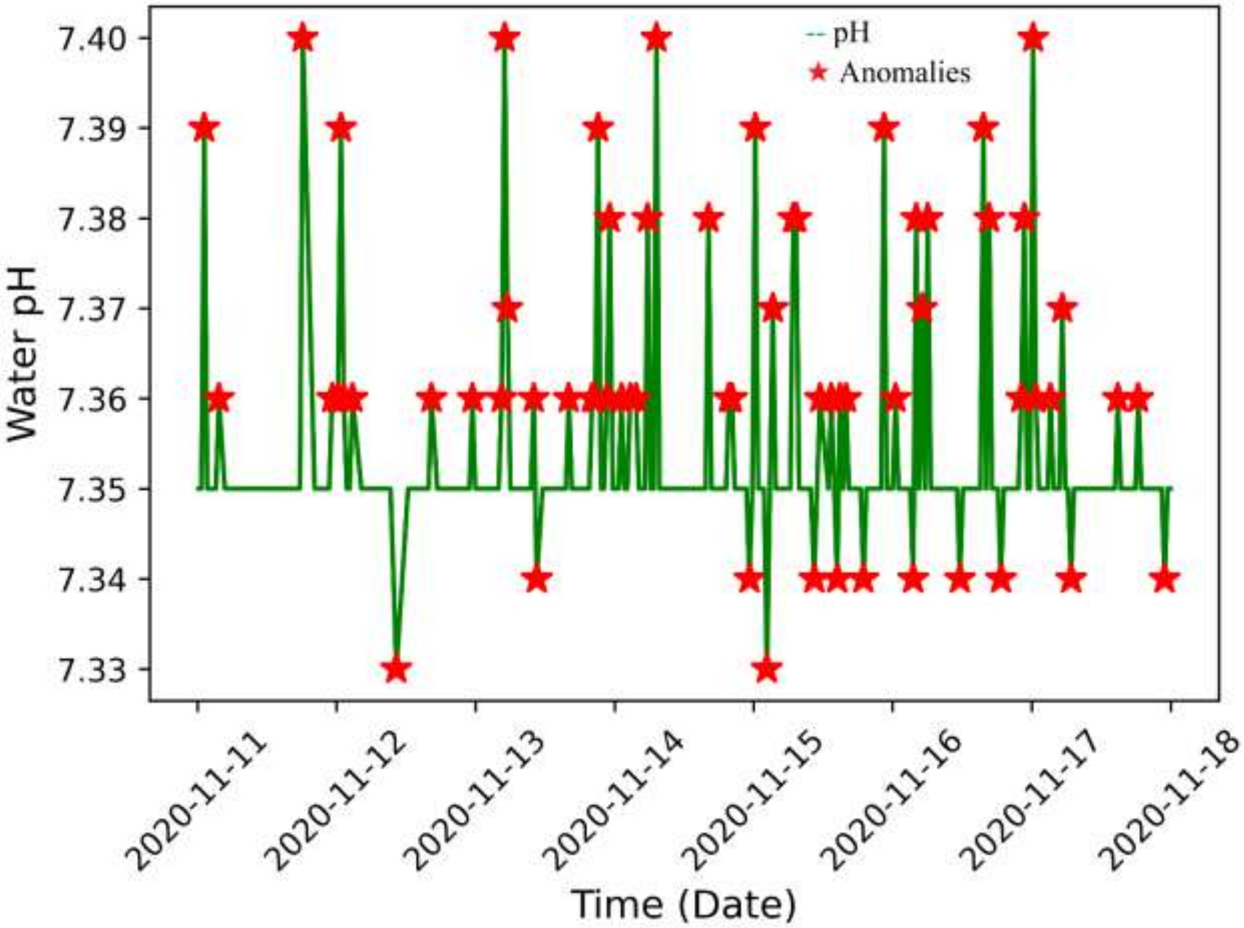


Figure 4.9: A plot of LOF pH outliers on the subset data

II. Isolation Forest and the Extended Isolation Forest Algorithms

Similar procedures as those in the turbidity data were adopted. It was also tough to find a feasible threshold for the improvement of anomaly detection. However, a plot of top 60 instances based on the score in (Figure 4.10) shows that standard IF worked better than EIF for turbidity data and found more anomalies with fewer false anomalies, a twist of what was achieved for the turbidity data. This process took 1.69s for the IF algorithm and 1.89s for the EIF algorithm.

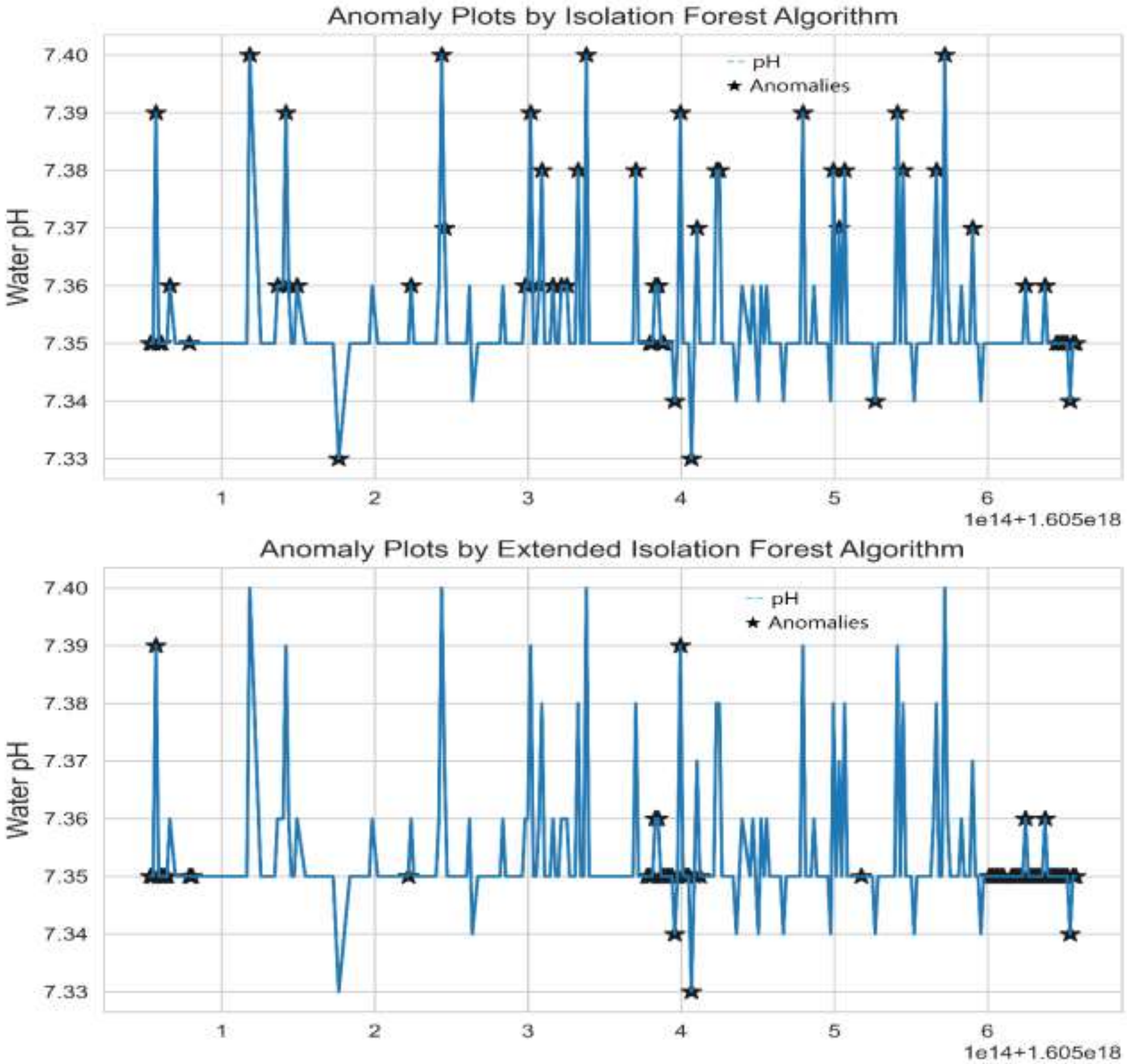


Figure 4.10: A plot of IF and EIF pH outliers on the subset data

III. The Robust Random Cut Forest Algorithm

The pH anomalies detected by the RRCF are shown in Table 4.10 and plotted in Figure 4.11. A feasible threshold to split the outliers was hard to determine and therefore the top 61 records with the highest outlier scores were identified. A significant number of point outliers from the beginning of the subset detected by the LOF algorithm were not detected, such as the pH record 7.39 [2020-11-11 01:11:34.389403+00:00]. The anomaly detection process took 2.51 seconds.

Table 4.9: pH outliers as detected by the RRCF algorithm on the subset data

time	pH
2020-11-11 18:12:35.713854+00:00	7.40
2020-11-11 23:15:16.261228+00:00	7.36
2020-11-12 00:15:48.373825+00:00	7.36
2020-11-12 00:46:04.433087+00:00	7.39
2020-11-12 01:16:20.473824+00:00	7.36
... ...	
2020-11-17 05:17:04.412038+00:00	7.37
2020-11-17 06:47:52.593118+00:00	7.34
2020-11-17 14:53:25.015789+00:00	7.36
2020-11-17 18:25:17.425846+00:00	7.36
2020-11-17 22:57:41.959856+00:00	7.34
61 rows × three columns	

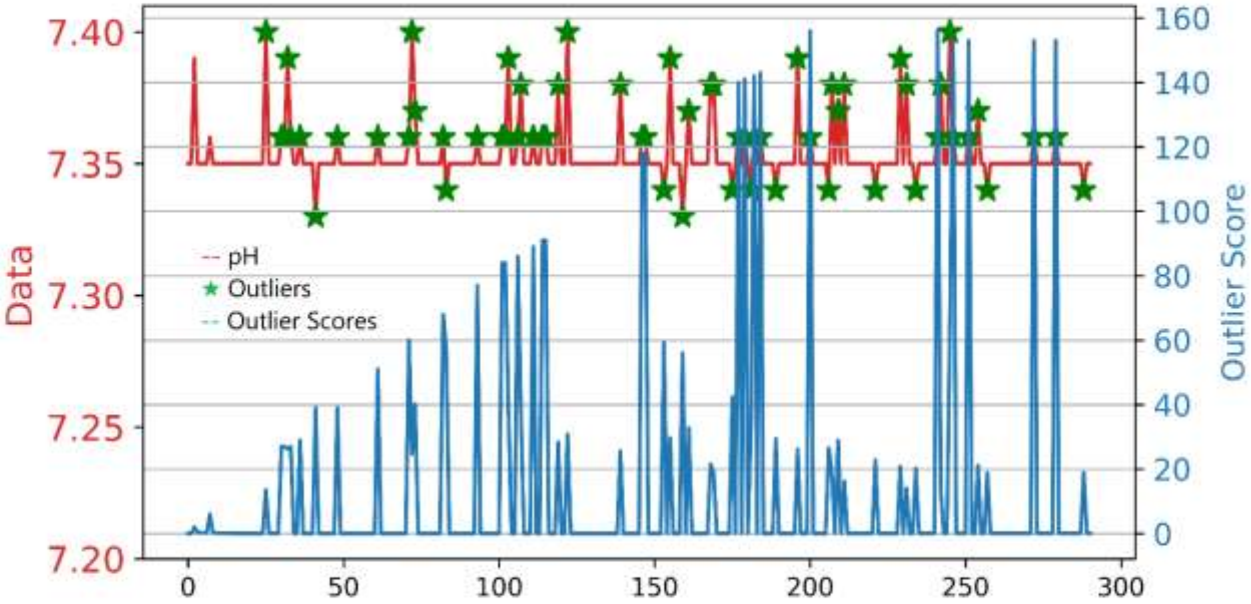


Figure 4.11: A plot of RRCF pH anomalies and their outlier scores on the subset data

Based on the ground truth subset, the LOF algorithm successfully detects all the 63 anomalies in the time series water pH subset data and all the 75 anomalies in the time series turbidity data, as summarized in table 4.10 and table 4.11. The RRCF algorithm suffers from 19 false abnormalities as well as missing 27 outliers in the turbidity subset. The case is similar for two undetected point anomalies in the pH subset data. Finding a score threshold for the IF and the EIF algorithms was complicated to determine a feasible number of anomalies. However, from the plots obtained, they suffer false anomalies and undetected anomalies. Additionally, the LOF algorithm was the fastest in detecting anomalies for both turbidity and pH data compared to the IF, EIF, and RRCF algorithms. While the LOF algorithm took only milliseconds, the IF, EIF, and RRCF algorithms consumed a second and more.

Table 4.10: pH subset data algorithms performance evaluation

Algorithm	Anomalies	False Anomalies	Undetected Anomalies	Execution Time
LOF	63	0	0	21 ms
IF	-	-	-	1.88s
EIF	-	-	-	1.25s
RRCF	61	0	2	2.51s

Table 4.11: Turbidity subset data algorithms performance evaluation

Algorithm	Anomalies	False Anomalies	Undetected Anomalies	Execution Time
LOF	75	0	0	38.9 ms
IF	-	-	-	3.66s
EIF	-	-	-	3.91s
RRCF	67	19	27	7.1s

4.3.2 Performance Evaluation Based on the overall Datasets

In this section, similar procedures as those employed in ***Section 4.4.1*** were adopted for the whole dataset except for input parameters as elaborated below.

A. Turbidity Dataset

I. The Local Outlier Factor Algorithm

The turbidity dataset was subjected to the LOF algorithm with the nearest number of neighbors parameter value $k = 800$, found to be appropriate. The algorithm took 1.55 seconds to determine a total of 278 anomalies plotted (red asterisks) in figure 4.12 below. Physical observation guarantees that the LOF algorithm successfully identifies a feasible number of contextual anomalies in the dataset.

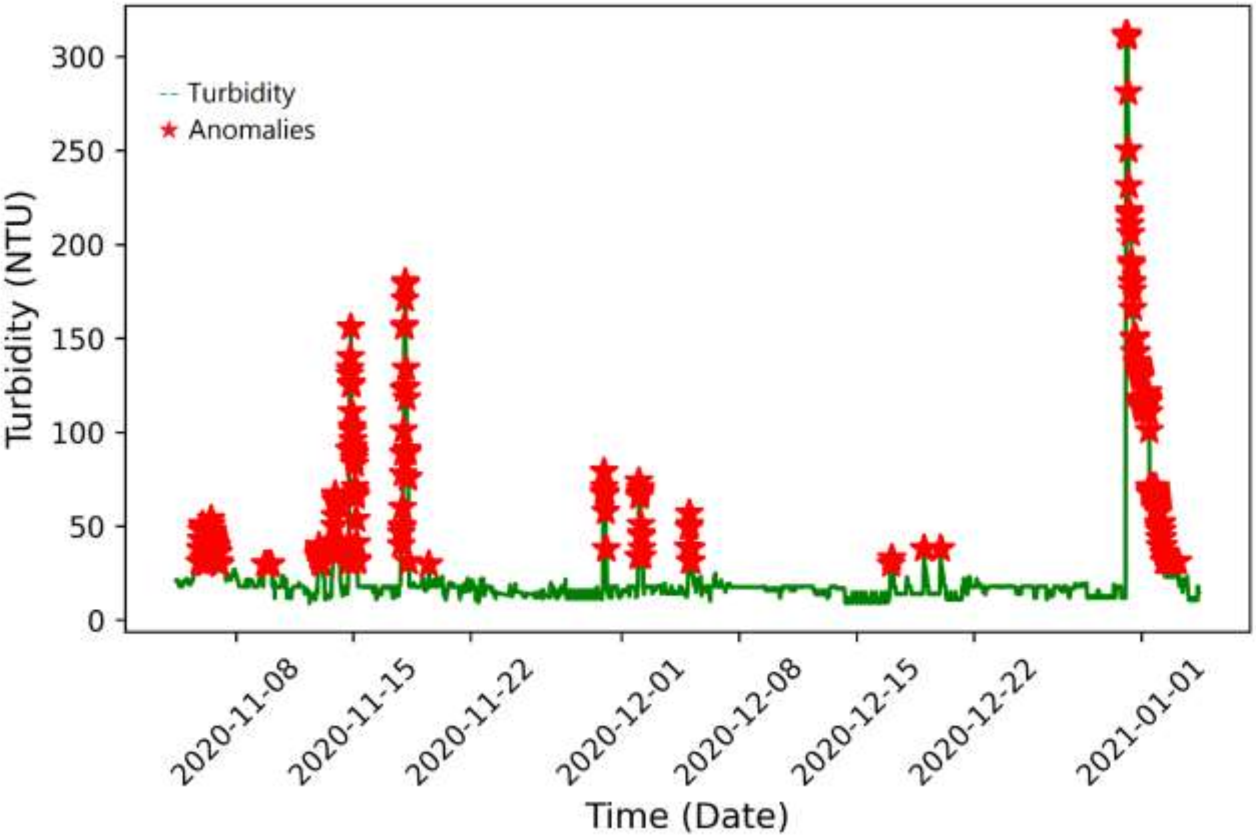


Figure 4.12: A plot of LOF turbidity anomalies on the whole dataset

II. The Isolation Forest and the Standard Isolation Forest Algorithms

In this procedure, the sub-sampling size of ψ=200 was also used for both algorithms, and there was no need to vary it more to unnecessarily increase memory consumption and the time taken to process data. Moreover, it was observed that the number of trees t directly controlled the ensemble size and the ideal paths converged at $t = 50$ as well. However, determining top anomalies was impossible as well. A plot of top 270 instances based on the score (Figure 4.13) shows that standard EIF worked better than IF for turbidity data and found more anomalies with fewer false anomalies. This process took 17.6s for the IF algorithm and 22.7s for the EIF algorithm.

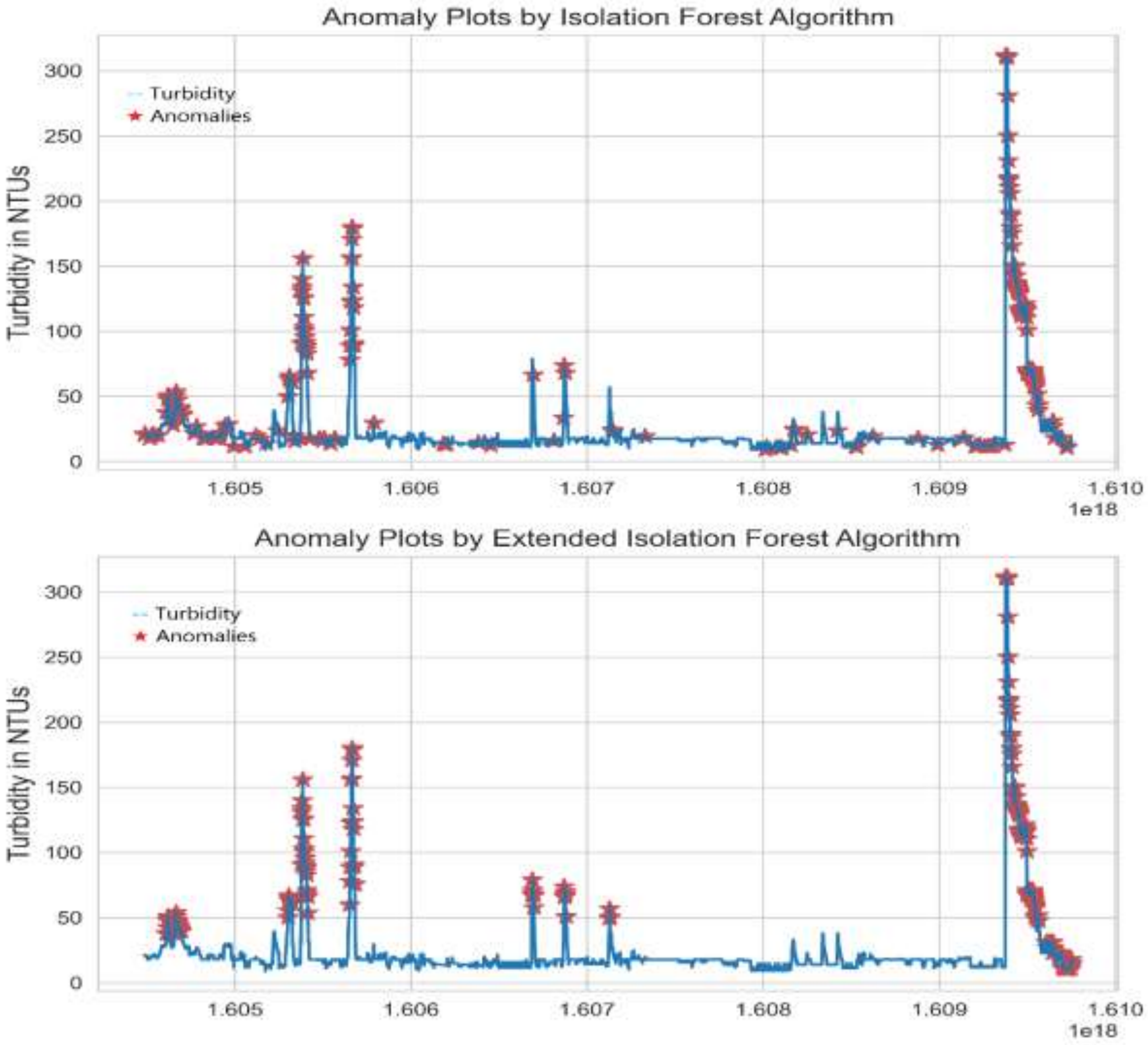


Figure 4.13: A plot of IF and EIF turbidity anomalies on the whole dataset

III. The Robust Random Cut Forest Algorithm

It was challenging to find a reasonable threshold to filter out anomalies since some outliers are marked with low anomaly scores (for example, instances at the beginning of the dataset). In contrast, some expected points are marked with high anomaly scores. The top 271 outlier records having the highest scores were found, consuming 3mins and 54seconds. These results were plotted as shown in Figure 4.14. A significant number of false outliers can be physically observed.

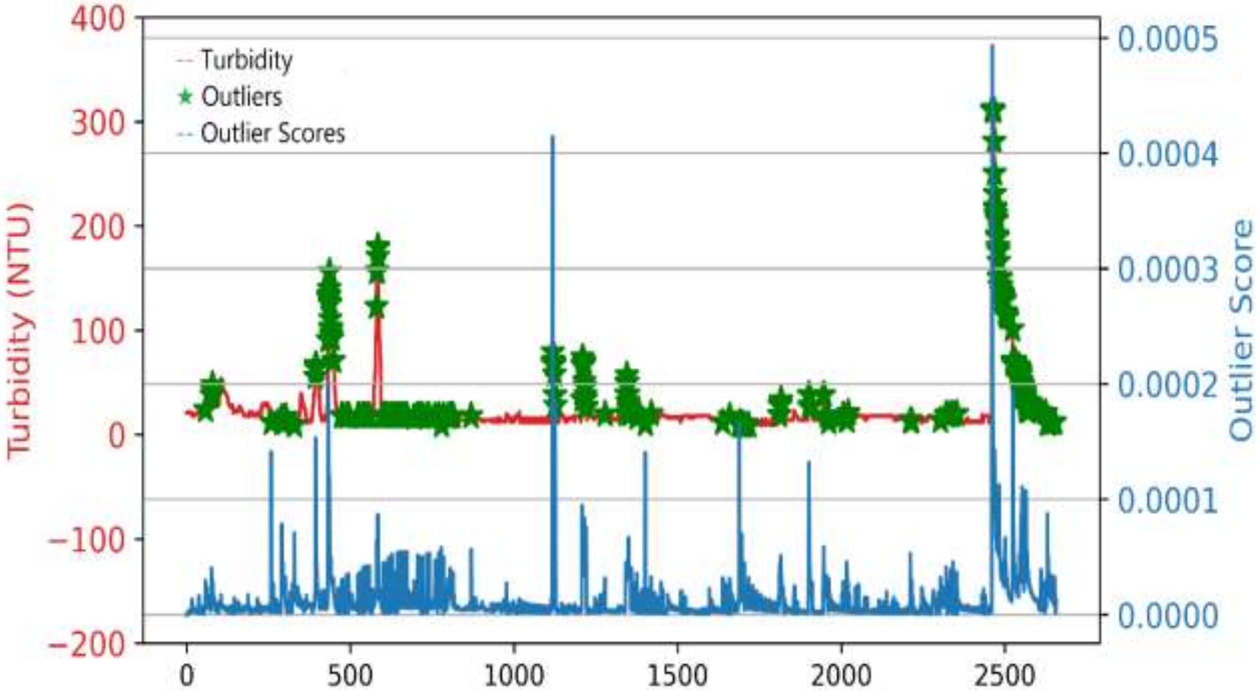


Figure 4.14: A plot of RRCF turbidity anomalies and their outlier scores on the whole dataset

B. pH Dataset

I. The Local Outlier Factor Algorithm

The pH dataset was subjected to the LOF algorithm with the nearest number of the neighbors parameter value, $k = 800$ found to be appropriate. The algorithm took 2.95 s to determine a total of 927 anomalies plotted (red asterisks) in figure 4.15 below. Physical observation guarantees that the LOF algorithm successfully identifies a feasible number of contextual anomalies in the dataset.

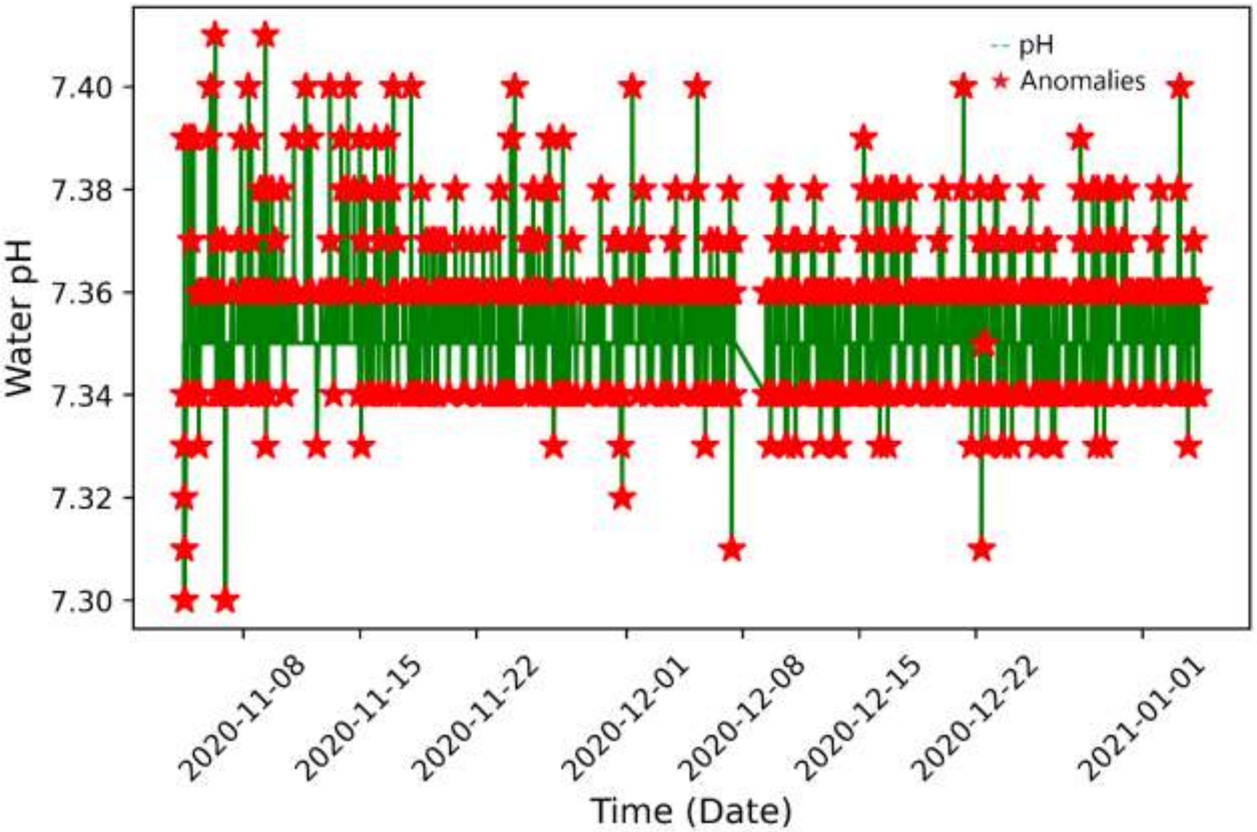


Figure 4.15: A plot of LOF pH anomalies on the whole dataset

II. The Isolation Forest and the Extended Isolation Forest

A plot of top 600 instances based on the score (Figure 4.16) shows that standard EIF worked better than IF for turbidity data and found more anomalies with fewer false anomalies. This process took 18.2s for the IF algorithm and 17.2s for the EIF algorithm.

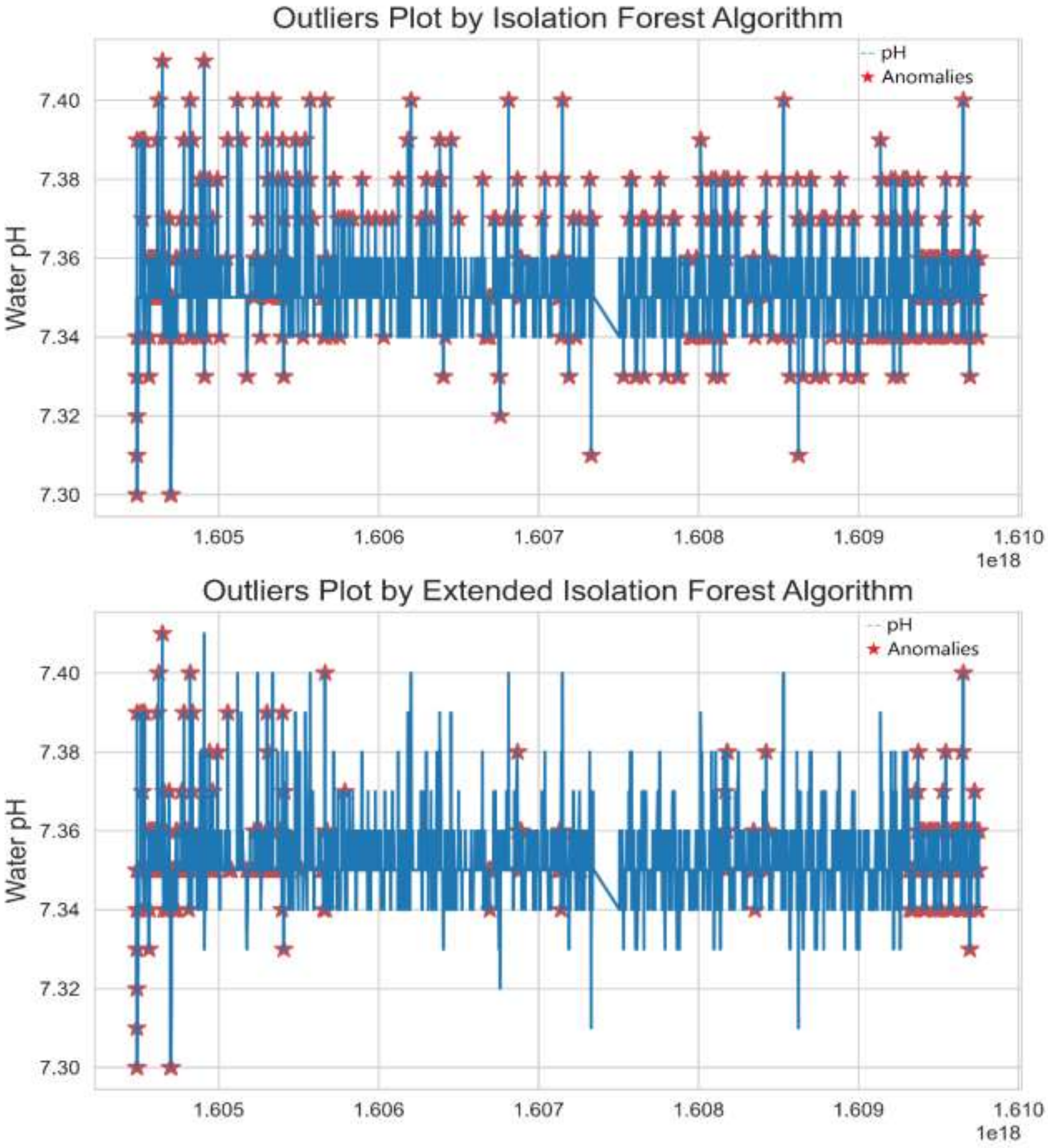


Figure 4.16: A plot of IF and EIF pH anomalies on the whole dataset

III. The Robust Random Cut Forest

The top 923 outlier records having the highest scores were found, consuming 4mins and 19seconds. These results were plotted as shown in Figure 4.17. A physical examination of its performance is recommendable.

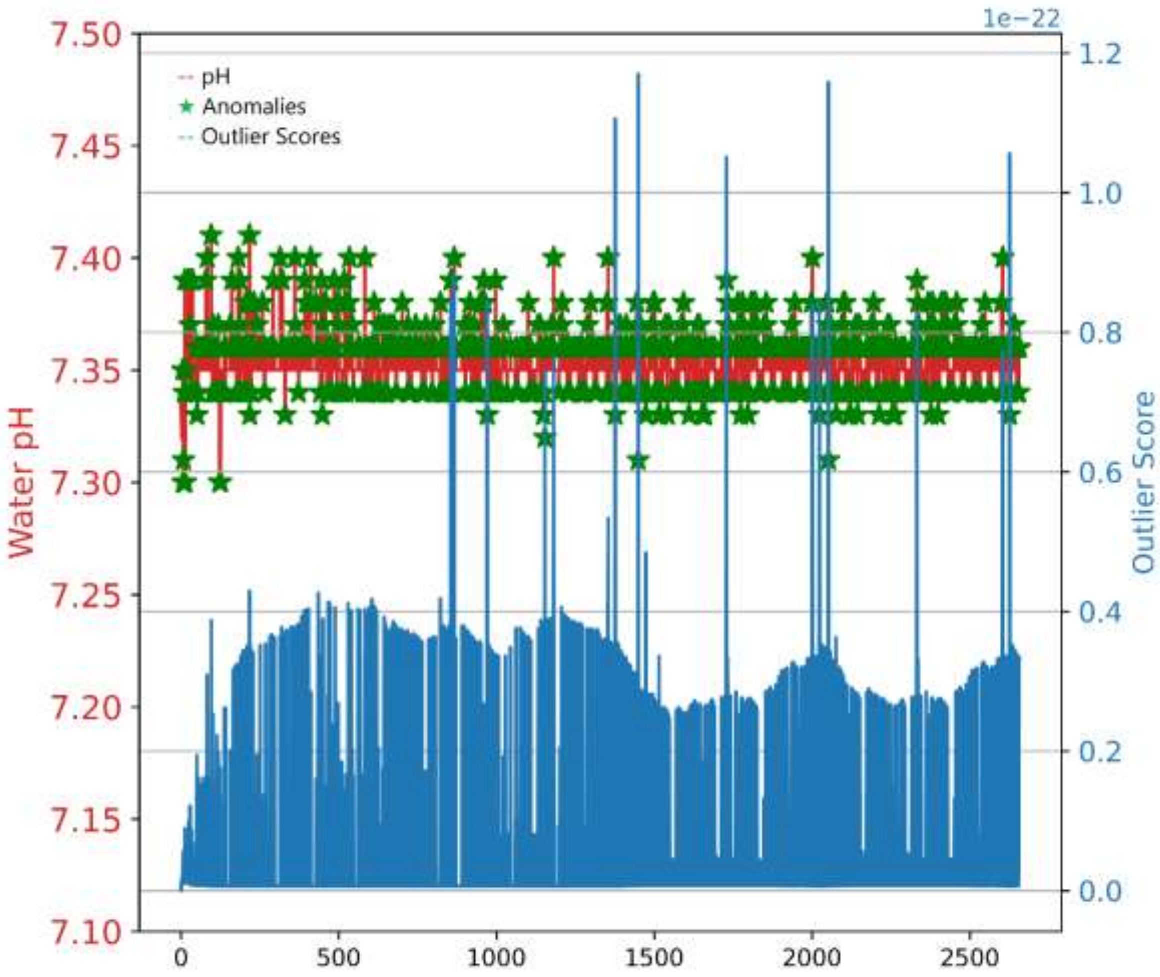


Figure 4.17: A plot of RRCF pH anomalies and their outlier scores on the whole dataset

Table 4.13 below highlights a summary of some of the evaluational factors for the four anomaly detection algorithms based on the whole dataset. The LOF algorithm emerged to be the fastest to process both datasets, while the RRCF took the longest time. It was easier to use and understand the LOF algorithm than the IF, EIF, and RRCF algorithms.

Table 4.12: pH and Turbidity whole datasets algorithms evaluation

Algorithm	**Anomalies**		**Execution Time**	
	pH	*Turbidity*	pH	*Turbidity*
LOF	927	*278*	728 ms	*1.55 s*
IF	-	-	18.2 s	*17.6s*
EIF	-	-	17.2 s	*22.7s*
RRCF	923	*271*	4min 19s	*3min 54s*

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This research presents the development of a low-cost sensor node that can be used to perform automated raw water quality monitoring in a treatment plant.

First and foremost, there were LoRa technology experiments on the range of coverage and connectivity using the RSSI parameter of the transceiver signals in the DeKUT main campus rural area. The best RSSI was realized in places near the gateway (100m away), a mean strength of -102.7 dBm, while the least RSSI was recorded at the furthest point of testing (1 km), whose mean signal strength was -113.7 dBm. Within a range of 1km, LoRa technology can satisfactorily be employed for data collection with WSNs due to good connectivity.

The developed sensor node contained two water quality sensor probes used to monitor water quality. This included the DFRobot Gravity Arduino turbidity sensor and the DFRobot's Gravity Analog pH sensor. The developed system is power-saving lightweight, and it can comfortably transmit data remotely, using LoRa technology.

Besides, this research presented a comprehensive evaluation of four different machine learning anomaly detection algorithms on two parameters from a water sensor node deployed at the NYEWASCO water treatment plant at the raw water section. A subset of 291 records extracted from the primary dataset of 2658 records was analyzed for both parameters. The LOF algorithm emerged superior to the IF, the EIF, and the RRCF algorithms in contamination event detection and hence a practical water contamination detection algorithm that can trigger alarms to alert the users when contamination is detected.

The framework is more suitable for large-scale implementation to collect and analyze raw water quality data in water supply firms and water authorities.

5.2 Recommendations

Further work can be done on the LoRa technology connectivity and range evaluation studies to develop a wireless propagation model in a rural setup of DeKUT. This can incorporate factors like free space attenuation, shadowing, reflection and transmission, and diffraction.

The developed water quality management system can be installed in multiple locations in water distribution networks to gather water quality data and classify sensor responses in practical deployments. Water is a vast network of related bodies such as rivers, lakes, swamps, dams, and other sources. If these linked parts contain different levels of pollution, assessing water quality may be a complicated endeavor.

Moreover, more water quality parameters can be incorporated into the developed system, such as temperature, making it a robust water quality parameter monitoring. Besides turbidity and water pH, other water quality parameters include total dissolved solids, oxygen reduction potential, electrical conductivity, dissolved oxygen, free residual chlorine, nitrates, to mention just but a few.

Additionally, further studies on the productivity of anomaly detection algorithms given several types of contaminants present in water can be done. Marginal risk assessment and the ability of algorithms for anomaly detections to accurately identify contaminants can be examined.

REFERENCES

[1] T. P. Lambrou and C. G. Panayiotou, "Collaborative Area Monitoring Using Wireless Sensor Networks with Stationary and Mobile Nodes," *EURASIP Journal on Advances in Signal Processing,* vol. 2009, no. 1, 2009.

[2] K. K. Khedo, R. Perseedoss and A. Mungur, "A Wireless Sensor Network Air Pollution Monitoring System," *International Journal of Wireless & Mobile Networks,* vol. 2, no. 2, pp. 31- 45, 2020.

[3] H. Luo, W. Li and X. Wu, "Design of indoor air quality monitoring system based on wireless sensor network," *IOP Conference Series: Earth and Environmental Science,* vol. 208, no. NA, p. 012070, 2018.

[4] J. Barica and D. Chapman, "An Indexed Bibliography of Toxic Contaminants in Water, Sediments and Biota of the Great Lakes Part II," *Water Quality Research Journal,* vol. 25, no. 4, pp. 506-636, 1990.

[5] J. P. Field, K. L. Farrell-Poe and J. L. Walworth, "Comparative Treatment Effectiveness of Conventional Trench and Seepage Pit Systems," *Water Environment Research,* vol. 79, no. 3, pp. 310-319, 2007.

[6] R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring and D. Estrin, "Habitat monitoring with sensor networks," *Communications of the ACM,* vol. 47, no. 6, p. 34, 2004.

[7] P. Jiang, X. L. Pang and L. Dong, "Survey on Mobile Target Localization in Wireless Sensor Networks," *Applied Mechanics and Materials,* Vols. 738-739, pp. 133-139, 2015.

[8] N. Tadayon, L. Xing, A. E. Zonouz and S. Aïssa, "Cost-effective design and evaluation of wireless sensor networks using topology-planning methods in small-world context," *IET*

Wireless Sensor Systems, vol. 4, no. 2, pp. 43-53, 2014.

[9] I. F. Akyildiz and E. P. Stuntebeck, "Wireless underground sensor networks: Research challenges," *Ad Hoc Networks,* vol. 4, no. 6, pp. 669-686, 2006.

[10] "Water: A Shared Responsibility – The United Nations World Water Development Report 2," *Development in Practice,* vol. 17, no. 2, pp. 309-311, 2007.

[11] "Millennium development goals: time to reassess strategies," *BMJ,* vol. 331, no. 7525, 2005.

[12] K. S. Adu-Manu, C. Tapparello, W. Heinzelman, F. A. Katsriku and J.-D. Abdulai, "Water Quality Monitoring Using Wireless Sensor Networks," *ACM Transactions on Sensor Networks,* vol. 13, no. 1, pp. 1-41, 2017.

[13] J. Bhardwaj, K. K. Gupta and R. Gupta, "Towards a cyber-physical era: soft computing framework based multi-sensor array for water quality monitoring," *Drinking Water Engineering and Science,* vol. 11, no. 1, pp. 9-17, 2018.

[14] K. E. Sawaya, L. G. Olmanson, N. J. Heinert, P. L. Brezonik and M. E. Bauer, "Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery," *Remote Sensing of Environment,* vol. 88, no. 1-2, pp. 144-156, 2003.

[15] J. Hall, A. D. Zaffiro, R. B. Marx, P. C. Kefauver, E. R. Krishnan, R. C. Haught and J. G. Herrmann, "On-Line water quality parameters as indicators of distribution system contamination," *Journal - American Water Works Association,* vol. 99, no. 1, pp. 66-77, 2007.

[16] H. B. Glasgow, J. M. Burkholder, R. E. Reed, A. J. Lewitus and J. E. Kleinman, "Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies," *Journal of Experimental Marine Biology and Ecology,* vol. 300, no. 1-2, pp. 409-448, 2004.

[17] R. T. Noble and S. B. Weisberg, "A review of technologies for rapid detection of bacteria in recreational waters," *Journal of Water and Health,* vol. 3, no. 4, pp. 381-392, 2005.

[18] R. Ritabrata, "An Introduction to Water Quality Analysis," *International Research Journal of Engineering and Technology (IRJET),* vol. 6, no. 01, pp. 201-205, 2019.

[19] K. Randolph, W. Jeff, T. Lenore, L. Lin, L. P. D and S. Emmanuel, "Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll a and phycocyanin," *Remote Sensing of Environment 112,* vol. 11, pp. 4009-4019, 2008.

[20] Alobaidy, H. M. Abdul, Jawad, K. M. Bahram and J. K. Abass, "Evaluating raw and treated water quality of Tigris River within Baghdad by index analysis," *Journal of Water Resource and Protection* , vol. 7, no. 2, p. 649, 2010.

[21] O. Bin, F. Ahmad and B. M. Mohd Zubir, "Turbidimeter design and analysis: a review on optical fiber sensors for the measurement of water turbidity," *Sensors 9,* no. 10, pp. 8311-8335, 2009.

[22] V. W. Peter and S. John, "Water quality requirements and management," *Farming marine shrimp in recirculating freshwater systems,* vol. 13, pp. 128-138, 1999.

[23] C. James, K. T. Vi, E. Aaron, G. Sheeana, C. Samuel, R. Piumie, L. Kay, J. C. Russell and C. Daniel, "Combining Chemometrics and Sensors: Toward New Applications in Monitoring and Environmental Analysis," *Chemical Reviews* , vol. 120, no. 13, pp. 6048-6069, 2020.

[24] S. N. Zulkifli, A. R. Herlina and L. Woei-Jye, "Detection of contaminants in water supply: A review on state-of-the-art monitoring technologies and their applications," *Sensors and Actuators B: Chemical* , vol. 255, pp. 2657-2689, 2018.

[25] R. F. Olanrewaju, S. Muyibi, T. O. Salawudeen and A. M. Aibinu, "An intelligent modeling

of coagulant dosing system for water treatment plants based on artificial neural network," *Australian Journal of Basic and Applied Sciences,* vol. 6, no. 1, pp. 93-99, 2012.

[26] S. L. Zhan, Y. Sim, W. ChinaJun, W. Thongthai and K. Chin, "Treatment technologies of palm oil mill effluent (POME) and olive mill wastewater (OMW): A brief review," *Environmental Technology and Innovation,* vol. 15, no. 100377, 2019.

[27] G. Tuna, B. Nefzi, O. Arkoc and S. M. Potirakis, "Wireless Sensor Network-Based Water Quality Monitoring System," *Key Engineering Materials,* vol. 605, pp. 47-50, 2014.

[28] J. Liu, X. Wei, S. Bai, X. Bai and X. Wang, "Autonomous underwater vehicles localisation in mobile underwater networks," *International Journal of Sensor Networks,* vol. 23, no. 1, p. 61, 2017.

[29] F. Ge and Y. Wang, "Energy Efficient Networks for Monitoring Water Quality in Subterranean Rivers," *Sustainability,* vol. 8, no. 6, p. 526, 2016.

[30] A. Alkandari, M. alnasheet, Y. Alabduljader and S. M. Moein, "Water monitoring system using Wireless Sensor Network (WSN): Case study of Kuwait beaches," *2012 Second International Conference on Digital Information Processing and Communications (ICDIPC),* 2012.

[31] M. Allegretti, "Concept for Floating and Submersible Wireless Sensor Network for Water Basin Monitoring," *Wireless Sensor Network,* vol. 06, no. 06, pp. 104-108, 2014.

[32] J. T. Heinen, "GEO Yearbook 2003. United Nations Environment Programme (UNEP). 2004. UNEP, Nairobi, Kenya. 208 pp. $20 paperback.," *Environmental Practice,* vol. 7, no. 1, pp. 58-59, 2005.

[33] F. Gui and X. Q. Liu, "Design for Multi-Parameter Wireless Sensor Network Monitoring System Based on Zigbee," *Key Engineering Materials,* vol. 464, pp. 90-94, 2011.

[34] Z. Rasin and M. R. Abdullah, "Water Quality Monitoring System Using Zigbee Based Wireless Sensor Network," *Int. J. Eng. Technol.,* vol. 9, no. 10, pp. 14-18, 2009.

[35] N. Chaamwe, "Wireless Sensor Networks for Water Quality Monitoring: A Case of Zambia," *2010 4th International Conference on Bioinformatics and Biomedical Engineering,* pp. 1-6, 2010.

[36] N. Nasser, A. Ali, L. Karim and S. Belhaouari, "An efficient Wireless Sensor Network-based water quality monitoring system," *2013 ACS International Conference on Computer Systems and Applications (AICCSA),* 2013.

[37] J. Hayes, K. Lau and D. Diamond, "A Wireless Sensor Network for Monitoring Water Treatment," *2007 International Conference on Sensor Technologies and Applications (SENSORCOMM 2007),* 2007.

[38] M, Meghana; Kiran, Kumar B M; Ravikant, Verma; Divya, Kiran, "Design and Development of Real-Time Water Quality Monitoring System," in *In 2019 Global Conference for Advancement in Technology (GCAT), IEEE*, 2019.

[39] Chowdury, S. U. Mohammad, B. E. Talha, G. Subhasish, P. Abhijit, M. A. Mohd, A. Nurul, A. Karl and S. ,. Mohammad, "IoT based real-time river water quality monitoring system," *Procedia Computer Science,* vol. 155, pp. 161-168, 2019.

[40] P. Mehne, F. Lickert, E. Bäumker, M. Kroener and P. Woias, "Energy-autonomous wireless sensor nodes for automotive applications, powered by thermoelectric energy harvesting," *Journal of Physics: Conference Series,* vol. 773, p. 012041, 2016.

[41] J. V. Capella, A. Bonastre, R. Ors and M. Peris, "A step forward in the in-line river monitoring of nitrate by means of a wireless sensor network," *Sensors and Actuators B: Chemical*, vol. 195, pp. 396-403, 2014.

[42] J. Peng, "The Design of Wetland Water Environmental Monitoring System Using Digital Video Based on Wireless Sensor Networks," *2009 WRI International Conference on Communications and Mobile Computing,* 2009.

[43] M. Zhang, D. Li, L. Wang, D. Ma and Q. Ding, "Design and Development of Water Quality Monitoring System Based on Wireless Sensor Network in Aquaculture," *Computer and Computing Technologies in Agriculture IV,* pp. 629-641, 2011.

[44] M. K. Amruta and M. T. Satish, "Solar powered water quality monitoring system using wireless sensor network," *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s),* 2013.

[45] M. Pule, A. Yahya and J. Chuma, "Wireless sensor networks: A survey on monitoring water quality," *Journal of Applied Research and Technology,* vol. 15, no. 6, pp. 562-570, 2017.

[46] R. Morais, S. G. Matos, M. A. Fernandes, A. L. Valente, S. F. Soares, P. Ferreira and M. Reis, "Sun, wind and water flow as energy supply for small stationary data acquisition platforms," *Computers and Electronics in Agriculture,* vol. 64, no. 2, pp. 120-132, 2008.

[47] F. Regan, A. Lawlor, B. O. Flynn, J. Torres, R. Martinez-Catala, C. O'Mathuna and J. Wallace, "A demonstration of wireless sensing for long term monitoring of water quality," *2009 IEEE 34th Conference on Local Computer Networks,* 2009.

[48] J. Petajajarvi, K. Mikhaylov, M. Hamalainen and J. Iinatti, "Evaluation of LoRa LPWAN technology for remote health and wellbeing monitoring," *2016 10th International Symposium on Medical Information and Communication Technology (ISMICT),* 2016.

[49] A. Augustin, J. Yi, T. Clausen and W. Townsley, "A Study of LoRa: Long Range & Low Power Networks for the Internet of Things," *Sensors,* vol. 16, no. 9, p. 1466, 2016.

[50] L. Jie, W. Peng, J. Dexun, N. Jun and Z. Weiyu, "An integrated data-driven framework for

surface water quality anomaly detection and early warning," *Journal of Cleaner Production,* vol. 251, 2020.

[51] Q. Elena, C. Francesco, G. Giulio Di and P. Riccardo, "Machine learning for anomaly detection and process phase classification to improve safety and maintenance activities," *Journal of Manufacturing Systems,* vol. 56, pp. 117-132, 2020.

[52] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying Density-Based Local Outliers," *Association for Computing Machinery,* p. 93–104, 2000.

[53] T. L. Fei, M. T. Kai and Z. Zhi-Hua, "Isolation-based Anomaly Detection," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 6, pp. 1-39, 2012.

[54] H. Sahand, C. K. Matias and R. Brunner, "Extended Isolation Forest," *IEEE Transactions on Knowledge and Data Engineering,* pp. 1 - 1, 2019.

[55] S. Guha, N. Mishra, G. Roy and O. Schrijvers, "Robust random cut forest based anomaly detection on streams," *International conference on machine learning,* pp. 2712-2721, 2016.

[56] M. Jared, K. Ngetich and M. Ciira, "Long Range Low Power Sensor Networks for Agricultural Monitoring - A Case Study in Kenya," in *2019 IST-Africa Week Conference (IST-Africa),* Nairobi, 2019.

APPENDICES

Appendix I – Anomaly Detection Algorithms

***Algorithm 1**: iForest*(X, t, ψ) [53]

Inputs: X - input data, t - number of trees, ψ - subsampling size

Output: a set of *t iTrees*

```
Initialize $Forest$
set height limit $l = ceiling(log2\ \psi)$
**for** $i = 1$ to $t$ **do**
    $X' \leftarrow sample(X, \psi)$
    $Forest \leftarrow Forest \cup iTree(X', 0, l)$
**end for**
**return** $Forest$
```

***Algorithm 2**: iTree*(X, e, l) [53]

Inputs: X - input data, e - current tree height, l – height limit

Output: an iTree

```
**if** $e \geq l$ or $|X| \leq 1$ **then**
**return** $exNode\{Size \leftarrow |X|\}$
**else**
let $Q$ be a list of attributes in $X$
randomly select an attribute $q \in Q$
randomly select a split point $p$ from *max* and *min* values of attribute $q$ in $X$
$X_l \leftarrow filter(X, q < p)$
$X_r \leftarrow filter(X, q \geq p)$
**return** $inNode\{Left \leftarrow iT ree(Xl, e + 1, l),$
                 $Right \leftarrow iT ree(Xr, e + 1, l),$
                 $SplitAtt \leftarrow q,$
                 $SplitV Value \leftarrow p\}$
**end if**
```

***Algorithm 3**: PathLength*(x, T, e) [53]

Inputs: x - an instance, T - an iTree, e - current path length; initialized to zero when first called

```
Output: path length of $x$
if $T$ is an external node then
    return $e + c(T.size)$ {$c(.)$ is defined in Equation 1}
end if
$a \leftarrow T.splitAtt$
if $xa < T.splitV\ alue$ then
    return $PathLength(x, T.left, e + 1)$
else {$xa \geq T.splitValue$}
    return $PathLength(x, T.right, e + 1)$
end if
```

Algorithm 4: $iTree(X, e, l)$ [54]

```
Input: X - input data, e - current tree height, l -height limit
Output: an iTree
if e ≥ l or |X| ≤ 1 then
return exNode {Size ← |X|}
else
randomly select a normal vector $\vec{n} \in IR|X|$ by drawing each coordinate of $\vec{n}$ from a standard Gaussian distribution.
randomly select an intercept point $\vec{p} \in IR|X|$ in the range of $X$
set coordinates of $\vec{n}$ to zero according to extension level
    $X_l \leftarrow filter(X, (X - \sim p) \cdot \sim n \leq 0)$
    $X_r \leftarrow filter(X, (X - \sim p) \cdot \sim n > 0)$
    return inNode { $Left \leftarrow iT\ ree(Xl, e + 1, l)$,
                    $Right \leftarrow iTree(Xr, e + 1, l)$,
                    $Normal \leftarrow \vec{n}$,
                    $Intercept \leftarrow \vec{p}$}
end if
```

Algorithm 5: *ForgetPoint* [55]

1: Node υ in the tree where p is isolated in T is found.

2: u is let to be the sibling of υ. The parent of υ (and of u) is deleted and replaced with u (i.e., the path from u to the root is short-circuited).

3: All bounding boxes starting from u's (new) parent upwards are updated. However, this state is not necessary for deletions.

4: The modified tree T is returned.

Algorithm 6*: *InsertPoint [55]

1: For a set of points S' and a tree$T(S')$, a new point p and produce tree $T'(S' \cup \{p\})$ is inserted.

2: If $S' = \emptyset$ then a node containing the single node p is returned.

3: Otherwise S' has a bounding box $B(S') = [x_1^l, x_1^h] \times [x_2^l, x_2^h] \times \ldots [x_d^l, x_d^h]$. Let $x_i^l \leq x_i^h$ for all i.

4: For all i let $\hat{x}_i^l = \min\{p_i, x_i^l\}$ and $\hat{x}_i^h = \max\{x_i^h, p_i\}$

5: A random number $r \in [0, \sum_i (x_i^h - x_i^l)]$ is chosen.

6: This r corresponds to a specific choice of a cut in the construction of $RRCF\ (S' \cup \{p\})$.

For instance, compute arg $\min\{j | \sum_i^j (x_i^h - x_i^l) \geq r\}$ and the cut corresponds to choosing $\hat{x}_j^l + \sum_{i=1}^j \left(x_i^h - x_i^l\right) - r\}$ in dimension j.

7: If this cut separates S' and p (i.e., is not in the interval $[x_j^l\ ,\ x_j^h\ ]$) then this is used as the first cut for $T'(S' \cup \{p\})$. A node is created whereby one side of the cut is p and the other side of the node is the tree $T(S')$.

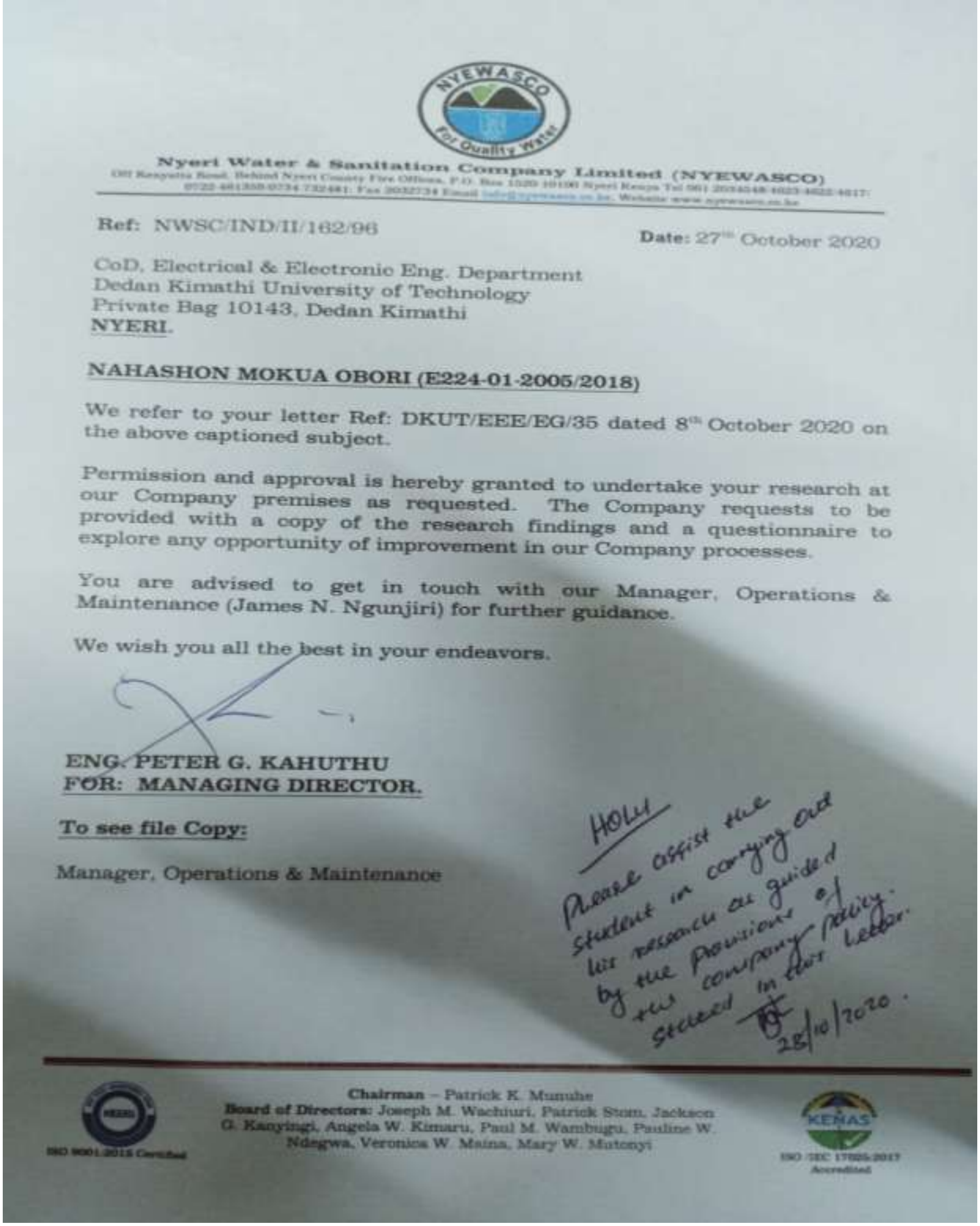
8: If this cut does not separate S' and p then the cut thrown away! The same dimension is chosen as $T(S')$ in $T'(S' \cup \{p\})$ and the exact same value of the cut chosen by $T(S')$ and a split performed. The point p goes to one of the sides with subset S''. This procedure is repeated with a smaller bounding box $B(S'')\ of\ S''$. For the other side, the same subtree as in $T(S')$ is used.

9: The bounding box of T' is updated in either cases

Appendix II – List of Conference Papers & Journal Publications

1. N. Mokua and C. Maina, "**A Study on the Performance of LoRa in a Rural Environment: Connectivity and Range Evaluation**", *DeKUT International Conference on Science, Technology, Innovation and Entrepreneurship*, vol. 5, 2019.
2. M. Nahshon, W. M. Ciira and K. Henry, "**A Raw Water Quality Monitoring System using Wireless Sensor Networks**," *International Journal of Computer Applications,* vol. 174, no. 21, pp. 35-42, 2021.
3. M. Nahshon, W. M. Ciira and K. Henry, "**Anomaly Detection for Raw Water Quality – A Comparative Analysis of the Local Outlier Factor Algorithm and the Random Forest Algorithms**" *International Journal of Computer Applications,* vol. 174, no. 26, pp. 47-54, 2021.

Appendix III – Permission and Approval Letter for Sensor Node Deployment at NYEWASCO Treatment Plant



NYEWASCO
For Quality Water

Nyeri Water & Sanitation Company Limited (NYEWASCO)
Off Kenyatta Road, Behind Nyeri County Fire Offices, P.O. Box 1520-10100 Nyeri Kenya Tel 061 2034048/4023/4622/4617; 0722 481358/0734 732481; Fax 2032734 Email info@nyewasco.co.ke, Website www.nyewasco.co.ke

Ref: NWSC/IND/II/162/96

Date: 27th October 2020

CoD, Electrical & Electronic Eng. Department
Dedan Kimathi University of Technology
Private Bag 10143, Dedan Kimathi
NYERI.

NAHASHON MOKUA OBORI (E224-01-2005/2018)

We refer to your letter Ref: DKUT/EEE/EG/35 dated 8th October 2020 on the above captioned subject.

Permission and approval is hereby granted to undertake your research at our Company premises as requested. The Company requests to be provided with a copy of the research findings and a questionnaire to explore any opportunity of improvement in our Company processes.

You are advised to get in touch with our Manager, Operations & Maintenance (James N. Ngunjiri) for further guidance.

We wish you all the best in your endeavors.

ENG. PETER G. KAHUTHU
FOR: MANAGING DIRECTOR.

To see file Copy:

Manager, Operations & Maintenance

HOLU
Please assist the student in carrying out his research as guided by the provisions of the company policy stated in this letter.
28/10/2020.

Chairman – Patrick K. Mumuhe
Board of Directors: Joseph M. Wachiuri, Patrick Stum, Jackson G. Kanyingi, Angela W. Kimaru, Paul M. Wambugu, Pauline W. Ndegwa, Veronica W. Maina, Mary W. Mutonyi

ISO 9001:2015 Certified

KENAS
ISO /IEC 17025:2017 Accredited