

**MODELING AND PREDICTING PIPED WATER THEFT
USING MACHINE LEARNING APPROACH**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by:

Khabusi Simon Peter
Roll No: 2k18/CSE/22

Under the Supervision of

Prof. Rajni Jindal



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

JUNE, 2020

DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY DELHI COLLEGE OF ENGINEERING)

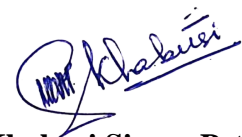
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Khabusi Simon Peter, Roll No 2K18/CSE/22, student of M.Tech in Computer Science & Engineering, hereby declare that the project Report titled "Modeling and Predicting Piped Water Theft using Machine Learning Approach" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirements for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Date: 20th June, 2020



Khabusi Simon Peter

STUDENT


DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Report titled "Modeling and Predicting Piped Water Theft using Machine Learning Approach" which is submitted by Khabusi Simon Peter, Roll No 2K18/CSE/22 Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 20th June, 2020


Prof. Rajni Jindal (HOD)

SUPERVISOR

ABSTRACT

Water theft is a prevalent problem in most countries across the globe which leads to loss of money, contamination of water, poor water supply, pipe bursts, water leakage and unbalanced flow among other problems. The existing works on this topic basically use hardware to detect water theft. Basing on the sophistication and dynamic nature of the problem, intelligent techniques are needed. Machine Learning and data analysis today form an integrative segment of the latest scientific methodology providing intelligent and automated approaches for predicting phenomenon based on past observations, discovering hidden patterns in data and giving insights about the problem. Machine Learning should however not be used as a black box tool, but as a method whose application should be formulated based on the study problem. Understanding the properties, mechanisms and limitations of the Random Forest algorithm was hence necessary prior to its use. This thesis therefore provides in-depth study of Random Forest and the water theft problem in water distribution pipelines and uses the RF algorithm to model a classifier for piped water theft prediction. Data was collected over a three hour period in 10 seconds intervals using the experimental setup of a hardware framework across a distribution pipeline interconnected with flowrate sensors at branch nodes interfaced with Arduino controller. The state of the network was recorded based on the sms alerts received on the mobile phone through GSM modem. The data was then tabulated, cleaned, explored and visualized to understand its pattern. RF model training was done using 80% of the data and likewise for the other benchmark techniques, that is; SVM, KNN and LR. Testing the models utilized 20% of the data and the four models were evaluated on the basis of accuracy, precision, recall and F-score. RF and KNN models achieved the highest accuracy of 97%. Conclusively, the proposed RF model is more advantageous compared to the other techniques in terms of reliable feature importance estimate and efficiency in test error estimation.

ACKNOWLEDGEMENT

First, I would like to extend my sincere gratitude to my mentor, Prof. Rajni Jindal for her constant encouragement and insight. I thank her for suggesting the initial idea of supervised Machine Learning, which eventually formed the core of this dissertation. I greatly appreciate her constant motivation in enabling me to understand the Machine Learning concepts which I was not familiar with. This work would not have been accomplished without her support.

The faculty members of Computer Science and Engineering played an extensive role in preparing me to pursue this research. Without, their knowledge of the various core courses they taught me, I would not have acquired the technical knowledge of undertaking such a project. On this note therefore, I sincerely thank them for preparing me for the journey that has successfully come to its completion.

I was fortunate enough to be part of a vibrant class of members who were a great pillar in my Masters studies. I am grateful for the intellectual support they rendered to me during the two years of vigorous training. Working with classmates from various backgrounds also enabled me to develop a true sense of awareness of equality, diversity and inclusion

I would like to thank my parents, Mr. Waburoko Thadeus and Mrs. Waburoko Elizabeth, for being supportive in every step of my academic journey. I will always be grateful for the advice and support they have given me over the years.

I also take this honour to sincerely appreciate the Indian Government through Indian Council for Cultural Relations (ICCR), my sponsoring organization for funding my Masters Education.

Above all, I thank the Almighty God for enabling me to accomplish this project.

CONTENTS

CANDIDATE'S DECLARATION	ii
CERTIFICATE	iii
ABSTRACT	iv
ACKNOWLEDGEMENT	v
CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	ix
LIST OF ABBREVIATIONS AND NOMENCLATURE	x
CHAPTER 1: INTRODUCTION	1
1.0 Chapter Summary	1
1.1 Introduction	1
1.2 Research Questions and Objectives	3
1.3 Problem Formulation	3
1.4 Scope	4
1.5 Limitation	4
CHAPTER 2: BACKGROUND AND MOTIVATION	5
2.0 Chapter Summary	5
2.1 Background	5
2.2 Concepts of Water Distribution	9
2.2.1 Piped Water	9
2.2.2 Flow Rate	10
2.2.3 Hydraulics	10
2.2.4 Water Theft	10
2.2.5 Water Meters	11
2.2.6 Water distribution	12
2.2.7 Water Distribution Pressure	16

2.2.8 Pressure Loss	16
2.2.9 Uphill and downhill Pressure Maintenance	17
CHAPTER 3: LITERATURE REVIEW	18
3.0 Chapter Summary	18
3.1 Related Work	18
3.2 Existing Water theft Applications and Prevention	23
3.2.1 Main metering	23
3.2.2 Prepaid metering Systems	24
3.2.3 Physical Monitoring	24
3.3 Ensemble Methods	26
3.3.1 Bagging	28
3.3.2 Boosting	29
3.3.3 Stacking	31
3.4 Decision Trees	31
CHAPTER 4: METHODOLOGY	33
4.0 Chapter Summary	33
4.1 Materials and Methods (Description of the algorithms)	33
4.1.1 Random Forests	33
4.2 Benchmark Methods	36
4.2.1 K-Nearest Neighbour (KNN)	36
4.2.2 Support Vector Machine (SVM)	37
4.2.3 Logistic Regression (LR)	38
4.3 Project Description	38
4.4 System Design for data collection	39
4.4.1 Block Diagram	39
4.4.1 Flow chart	40
4.5 Components Description	40

4.6 Experimentation for Data Collection	42
4.6.1 Data Collection	43
4.6.2 Data Analysis and Processing	44
CHAPTER 5: RESULTS AND DISCUSSION	46
5.0 Chapter Summary	46
5.1 Modeling and Prediction	46
5.2 Model Evaluation	46
5.3 Discussion	48
CHAPTER 6: CONCLUSION AND FUTURE WORK	50
6.0 Chapter Summary	50
6.1 Conclusion	50
6.2 Future Work	51
APPENDICES	52
Appendix 1: System Prototype for Data Collection	52
REFERENCES	53
LIST OF PUBLICATIONS	60

LIST OF TABLES

Table 1: Confusion Matrix	47
Table 2: Performance Measure	48

LIST OF FIGURES

Figure 1: Conceptual framework of existing systems	8
Figure 2: Branch Distribution Network	14
Figure 3: Grid Distribution Network	15
Figure 4: In-house piped water connection	15
Figure 5: Common Ensemble architecture	27
Figure 6: Bagging Example	29
Figure 7: Decision boundaries of (a) single DT (b) Bagging (c) 10 DTs used by Bagging on the three Gaussians dataset	29
Figure 8: Decision Tree Example	32
Figure 9: SVM Plane	37
Figure 10: Block diagram	39
Figure 11: Flow chart	40
Figure 12: Arduino Uno Board	41
Figure 13: Flowrate sensor/Flow meter	41
Figure 14: GSM Modem	42
Figure 15: System Prototype positioning along the water pipeline	43
Figure 16: Water flow pattern in experimental system	44
Figure 17: Performance of prediction models	48
Figure 18: System Prototype for data collection	52

LIST OF ABBREVIATIONS AND NOMENCLATURE

RF	Random Forest
SVM	Support Vector Machine
LR	Logistic Regression
KNN	K-Nearest Neighbour
DT	Decision Tree
PLC	Programmable Logic Controller
SCADA	Supervisory Control and Data Acquisition
GSM	Global System for Mobile Communication
GPS	Global Positioning System
IoT	Internet of Things
HOG	Histogram of Oriented Gradients
OPF	Optimum Path Forest
RTU	Remote Terminal Unit
HMI	Human Machine Interface
OPC	Open Platform Communications (technology)
WAN	Wide Area Network
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
LTI	Linear Time Invariant
CUSUM	Cumulative Sum
AdaBoost	Adaptive Boosting
FTDI	Future Technology Devices International
OST	Orthogonal Structure Tree
ICSP	In-Circuit Service Programming
UART	Universal Asynchronous Receiver Transmitter
LCD	Liquid Crystal Display
EPA	United States Environmental Protection Agency Water Supply and Water Resources Division.
EPANET	Software tool used by EPA for monitoring water flow, pressure variation and chemical content of water in a simulation period.
MI Hub	Smart control center for connection of other smart devices

CHAPTER 1: INTRODUCTION

1.0 Chapter Summary

In this chapter, the introductory description of the topic of study has been presented. This is followed by the research question that guided the study and the objectives. The problem being addressed has been formulated in the next section and the scope and limitation of the work have been well elaborated.

1.1 Introduction

Machine Learning algorithms are able to be trained and hence can improve their performance over time. The recent trends in Machine learning have positively impacted computing over the years. Its direct application in data mining has greatly enabled big data analytics and predictive modeling. More specifically, Machine learning has applications in Bioinformatics, Cheminformatics, Computational advertising and finance, affective computing information retrieval, game playing and robot locomotion, Medical diagnosis and Natural Language processing among others [1]. The learning signal or rather feedback that the learning system utilizes for training purposes categorizes the Machine Learning tasks into Supervised, Unsupervised and Reinforcement learning [2].

Supervised Learning has been used more commonly because the ease with which the models can be trained. Datasets with labeled examples make perfect inputs to a supervised learning algorithm which is task driven [3]. Besides, the datasets for training purposes do need require being so large. Using the given parameters in the dataset, a cause-effect relationship between the variables is created in training. At the end of the raining, a function that maps the input to the output is then generated to form the prediction model. The Supervised learning algorithms are further categorized into classification and regression algorithms. Examples of such algorithms include; Random Forest, Bayesian Network, Support Vector Machine and K-Nearest Neighbour among others. Tasks such as sperm classification and face recognition are supervised learning problems [4].

With unsupervised learning, unlabeled data is used in training the algorithm. In other words, the output values are unknown in this type of learning. Clustering such as K-

means clustering and unsupervised deep learning are perfect examples of unsupervised learning. In recommender systems and user login grouping, unsupervised learning algorithms are used to perform the clustering [5].

On the other hand, Reinforcement learning can be viewed as basing on mistakes to learn [1]. The algorithm placed in a real environment tries to perform with many errors at the start. However, the output improves over time if the algorithm is provided with a signal to correlate positive signal to good behavior and negative signal to bad behavior. In this way, the algorithm can be reinforced to choose good behavior and ignore the bad. Once this kind of reinforcement is done over time, the number of mistakes made eventually reduces [6]. Reinforcement learning can take different forms among which include; game playing and Industrial simulation to enable machines learn how to execute their tasks with efficiency over time.

However, to-date, no study has been undertaken on water theft detection, leakage control and water quality using Machine Learning. A number of research initiatives in this area emphasized the use of IoT, Embedded systems and other technologies for detecting cases of theft in water distribution pipelines [47] [48] [30]. The methods however face a number of draw backs. Firstly, the ineffectiveness of the software programs that run on the hardware to deduce the existence of anomaly flow. Secondly, the hardcoded rules are fixed which cannot adjust with the fluctuations that may occur in the water distribution processes. Thirdly, the problem of water theft is sophisticated and often times may not conform to the known stipulated methodologies.

This work therefore formulates a new direction of research on piped water theft detection by employing Random forest classifier to predict water theft on labeled data collected from the hardware prototype that forms a greater part of the existing work. Random Forest is one of the supervised learning algorithms that present a better classification rate by using a number of decision trees [1]. A vast number of decision trees can be constructed to form a random forest and this collection is what is referred to as an ensemble of trees. In classification, the model value of the classes is output whereas in regression, the mean prediction of these classes is given as the output. Random forests are therefore a learning method that uses an ensemble of trees for classification purposes and also regression. Through ensemble, the problem of overfitting to the training sets which often occurs in decision trees is addressed. The algorithm for Random Forest was developed by Adele Cutler [7] and Leo Breiman [8]. Geman [9], Amit and Ho [10] [11]

introduced the idea of selecting features randomly and in combination with this, the bagging concept designed by Breiman form the basis for the random forest classifier which aims at using controlled variance in the constructed multitude of decision trees.

In this thesis therefore, we propose a random forest classification model for piped water theft detection. In the binary classification, the labeled data collected by experimentation was used to train and test the classifier. Other Supervised learning algorithms such as Support Vector Machine, K-Nearest Neighbour and Logistic regression were used to compare the results of the classifier with those obtained by the benchmark methods.

1.2 Research Questions and Objectives

Research Questions

- (1) Can we use hardware (network of flowrate sensors interfaced with Arduino) to collect flowrate data?
- (2) Can we use data collected from a water distribution pipeline to detect changes in the pipeline?

Objectives

- (1) To design and Implement a hardware prototype of flowrate sensors interfaced with Arduino for data collection
- (2) To train and test a random forest classifier for piped water theft prediction
- (3) To evaluate the model on the basis of the confusion matrix and perform the training and testing on other benchmark techniques, that is; Support Vector Machine, K-Nearest Neighbour and Logistic Regression.

1.3 Problem Formulation

Piped water theft is a widely spread practice across various communities. The various reasons for the growing trend are; lowering the water bills, need for higher volumes of water that what is supplied and the financial inability to afford the service among other reasons. The water theft problem takes various forms which may include meter tampering, meter bypass, meter reversals, illegal connections and vandalism. There have been various initiatives approaches that have been put in place in trying to address the problem though its prevalence affirms the need for more intelligent solutions. The research initiatives so far undertaken emphasize the use of hardware with software applications that are rule based to detect and report theft. This has various drawbacks

since the basis for theft detection is dynamic pressure fluctuation and audit. Though the water distribution companies try to maintain a fixed range of distribution pressure within a given pressure, this is often not attainable. Additionally, there are no clearly defined mechanisms of how piped water theft is practiced which makes it more sophisticated. Therefore this study presents an intelligent approach to piped water theft prediction using Random Forest prediction model that has been trained and tested on water flow data collected over a period of time. This can help in countering the problem in equal measure by adapting to the state of the pipeline whenever changes occur. Continued training with more data collected from a water distribution pipeline can improve the model accuracy considerably.

1.4 Scope

The work uses experimental data, that is; water flowrate data collected over a 3 hour period in 10 seconds intervals to train and test a random forest classifier for piped water theft prediction. The trained model can predict theft only along the distribution pipeline used in the experiment since the flowrate and volume flow captured and used are those of the experimental network. For wider application, the network on which such a model is to be applied should be instead be used in data collection such as the network state can be predicted by the trained model whose training set is data from such network.

1.5 Limitation

A distinction cannot be made between water loss due to theft and water loss due to leaks resulting from pipe bursts and vandalism.

CHAPTER 2: BACKGROUND AND MOTIVATION

2.0 Chapter Summary

This chapter is divided into two main sections. The first section gives a thorough background of the study and the motivating factors behind undertaking this topic, the second section discusses the various concepts of water distribution.

2.1 Background

Water is a basic need in the life of every human being hence its scarcity can lead to detrimental effects in life [12]. The problem of water scarcity cuts across all continents as indicated in the report on world water developed released by United Nations in 2015 [12]. The issue of limited water that is safe and clean has been highlighted in the report as a prevalent issue all over the world and more predominant in Africa and other underdeveloped countries [12]. Water crisis can be viewed in two aspects, that is; natural and human made phenomenon [13]. Natural phenomenon generally leads to physical scarcity and this comprises of global warming and climate change. On the other hand, water sources may be adequately available but economic scarcity leads to inability in utilization of these sources. This is more common in low income communities in under developed economies around the world, most especially Africa [14].

Clean and safe drinking water access in rural communities in Africa is very limited. The world health organization report, 2015 shows that less than 50% of the African population living in rural areas have access to clean piped water for drinking [15]. Sanitation is also challenging in most of such communities. The need to improve such basic services is the motivation behind different research initiatives on piped water distribution and quality control. The well-being of mankind hinges on a number of issues that make up their day-to-day lives. These include agriculture, health, education, industrial production and economic growth among others which are directly or indirectly connected to the water supply quality and improved sanitation [14]. The increasing population and urbanization has also not only led to water scarcity but also poor quality water supply [16]. This is also partly contributed by the increasing inefficiency of government to fulfil her mandate of providing social services which include clean piped water supply to the citizens [17].

Vanda Felbab in one of the lectures held in 2015 defines the water theft problem as a non-discriminative act that is practiced by both the rich, the poor and those who are simply deprived of accessing clean and safe water. She illustrates with examples that water theft is a challenge across various countries and different continents such as Europe, California, Kenya, Nigeria, South Asia and the Middle East [18]. Piped water smuggling around the world using sophisticated approaches mafias is an emergency to ensure sustainability of water supply and preservation of its quality.

Felbab further acknowledges the difficulty to address the various shortcomings associated with water theft [18]. For the big consuming clients such as commercial agriculture and processing industry, controlling illegal water usage is difficult as they often beat the measures put in place by the supply companies, and water loss prevention units and regulators. Most poor communities resort to water theft as the only way of accessing clean drinking water since they cannot afford to get connected on the pipeline. This is very common across many slum areas around urban centers. Most citizens are against the water prices which are normally high and unaffordable by the poor, which leads to increased piped water crimes. The continued perpetration of water theft is a threat to sustainability of water distribution, water quality and equitable water use [18].

Amidst all these limitations, many African states are making tremendous strides in extending clean piped water to the communities [17]. A considerable improvement in piped water connectivity has been registered so far. For example, in 2015 the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) published a report on Investments and Sector reforms across Africa on the state of piped water distribution and connectivity network in 5 countries with Burkina Faso at 76% of piped water connectivity, followed by Kenya, Zambia and Tanzania at 70%, 66% and 59% respectively and last Uganda which was at 56% [17]. This data indicates a promising improvement in water supply to the citizens and there is still hope for further development.

The financial limitations and insufficient water supply in most urban centers in Uganda have resulted into increased illegal access of water from piped water distribution networks [13]. Water is a necessity of life and that justifies for the researchers' interest in its supply and distribution. A constant water supply is necessary in almost every aspect of society from residential consumption to public offices, hospitals, water intensive manufacturing companies such as Coca Cola and Nile Breweries to construction. Any disruption in water supply or scarcity in general could result into bad economic impact or

even loss of lives. For most water dependent entities, the default water supply rate is generally not sufficient to meet their needs despite the fact that a big chunk of their budget is allocated for water bills [19]. Such limitations trigger different water theft practices such as meter tampering, meter bypass, meter reversals, illegal connections and vandalism. Such practices aim at ensuring higher water supply at lower rates [20].

The use of suction pumps as a means of drawing more water from service lines is more common. This leads to poor water supply to other clients and even the quality of the water is also compromised. The high pressure suction pumps also lead to pipe bursts and mixing debris and algae that grows on the pipe surfaces with water hence polluting it. Such extreme pressures are impractical unachievable in the piped water supply systems [21]. A pressure zone is a region within which a constant water pressure is used for water supply. Pressure in these regions is maintained at a certain nominal value hence use of suction pressures interfere with water supply patterns, and leads to leaks and pipe bursts [21] [22] [23]. The wide spread of piped water theft, perpetrated by both the rich and the poor affirms the intensity of the problem [18].

Reports by National Water and Sewerage Corporation (NWSC) indicate that about 30 – 35% of revenue is lost through water theft cases [25]. The government body assigned with the sole role of extending and supplying clean and safe piped water to Ugandan Nationals make great losses through cases of piped water theft [25]. The corporation has made tremendous efforts in trying to address the problem through use of police, regular supervision and community sensitization. However, but there seems to be no considerable positive results registered so far. The value of Non-Revenue Water (NRW) is so high that it calls for urgent solutions. Different researchers have therefore taken on various research initiatives to devise a mechanism for automation in piped water distribution, monitoring and reporting.

Newumbe Vivien, the Kampala water public relations officer reports on 15th September 2015 in the New Vision National Newspaper, the challenges faced by National Water and Sewerage Corporation, a government body mandated to supply clean and safe piped water to the citizens. The statement highlights increased theft of water and vandalism of the water pipeline as major problem to the operations of the Corporation [26]. The increased theft and vandalism is attributed to lack of effective monitoring which greatly affects those at higher altitude settlements. Pipe bursts and illegal connections lower the distribution pressure hence making in difficult to pump water to the extreme end points

of the water pipeline. The resulting low pressures do not only affect the water supply chain but also makes water unsafe. Water contamination happens when there is a pipe burst and the surrounding people use various containers to collect the water. Agnes Kyotalengerire also explains the impact of such activities on the operations of NWSC. About 20% of the supplied water is unaccounted for and therefore such losses of income and contamination of water makes the task of the corporation is even harder. Water contamination also necessitates sub-water treatment plants which are expensive to maintain

The current work on piped water automation and theft identification uses flow rate sensors distributed along distribution networks to capture values of water flow rate and the data is accessed remotely through IoT component. The different communication technologies have also been applied, these are; SCADA (Supervisory Control and data Acquisition), Ethernet Shield, GSM, and other wireless communication mechanisms. The conceptual framework of these systems is shown in figure 1 below.

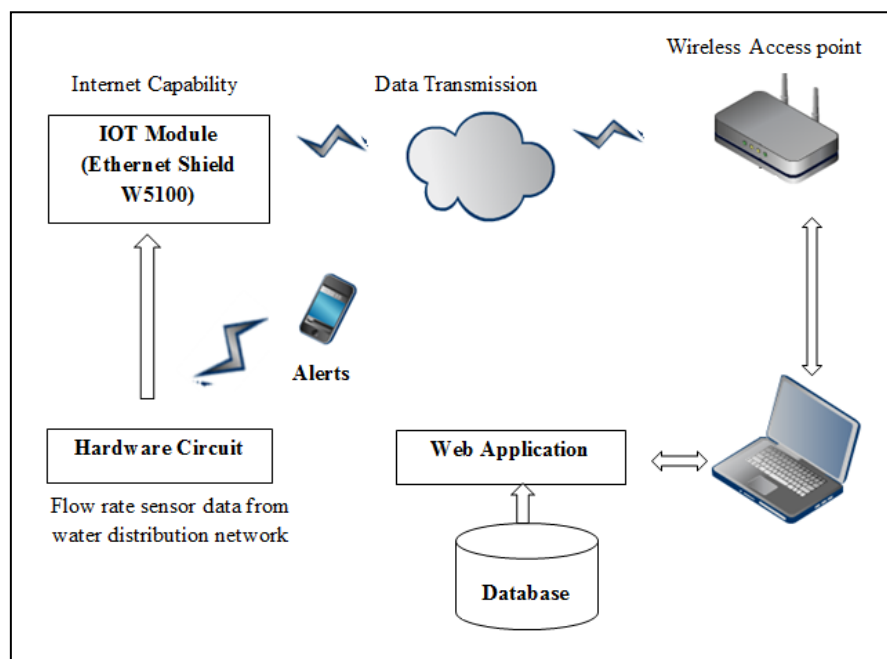


Figure 1: Conceptual framework of existing systems

An array of water flowrate sensors that are interfaced with microcontroller are used in capturing flow rate trends across the distribution system or network. A normal range of water flow rate value is hard coded in the microcontroller memory and variations of the captured flow rate data is designated as water theft. The GSM modem/ communication is mostly used to convey a sms alert to the authorities with an indication of the exact location of the theft; enabled using the GPS. Other technologies such as Ethernet shield

and SCADA are used for alerting and remote sensor data relay. The data is transmitted to a remote database that serves a specially designed web application [27]. The remote server enables the analysis of the inner workings of the distribution system as often as possible. Taking into consideration the sophistication of the water theft problem, these hardcoded rules cannot efficiently detect the problem. Therefore, intelligent mechanisms are required to isolate and detect the problem in real-time. Hence using the Random Forest Classifier, the water theft problem has been predicted with a higher accuracy. Such machine learning techniques can be developed further to use in complete systems.

The use of prediction models in various disciplines has registered considerable success over the years as Machine Learning field advances. One of the principle components that lead to their usage is performance that is focused on related aspects which include; (i) use of appropriate data sets with positive data labels and negative data which labeled (ii) use of various performance metrics to evaluate their performance (iii) Statistical tests are also used to evaluate the results on the basis of reliability (iv) using different datasets to train and test and validate the models [28].

Therefore, this thesis aimed at modeling and predicting water theft in piped water distribution networks using Machine Learning approach which employed random Forest Classifier and compared the results with other benchmark supervised learning methods. The data used in the study was collected using a hardware system prototype designed to detect water theft based on the hardcoded rules. With such a model trained and tested on real world data, water theft can be efficiently detected.

A discussion of the related work, materials and methods, project description, modeling and prediction, results and discussion, conclusion and future work have been presented in the proceeding sections.

2.2 Concepts of Water Distribution

2.2.1 Piped Water

Water is an oxygen and hydrogen compound with a boiling point of 100°C and freezing at 0°C [29]. Water which is a transparent liquid is characterized by lack of taste and odor. Being chemical compound, it's PH which is neutral can be altered due to various alterations caused by contamination. Water distribution tries to maintain a fixed range distribution pressures or flow rate [30]. Water theft and leakage in these piped water systems can also result into alteration of the chemical composition of the water hence

compromising with water quality. The pipes used for water distribution could also alter the chemical composition of water. In cases where lead pipes are used, it could result into lead poisoning as the surfaces of the pipe corrode and mixes with the water. However, this is not a common material used. uPVC is the major material used in these pipes and occasionally, plastic, ductile iron and steel are used [31].

2.2.2 Flow Rate

The amount of water or fluid that flows through a channel within a stipulated time is referred to as flow rate [29]. Given A as the cross-sectional area of a pipe and v as the average velocity of the fluid, the flow rate, Q of the fluid is calculated as [32];

$$Q = A \times v \quad (2.1)$$

Q can therefore be referred to as the volume per unit time, which is denoted at times as the volumetric flow rate. As an example, for a fluid flowing at 1 meter per second (average velocity) through a pipe of cross-sectional area 1 square meter, its flow rate is 1 cubic meter per second.

2.2.3 Hydraulics

The fluid behaviour and properties vary tremendously with the state of the fluid, that is when stationary and when in flow. This is the major concern of hydraulics [33]. A number of issues affect water flow in pipes, and these include;

- a) The pipe's cross-sectional area
- b) The texture of the inner surface of the pipe (roughness or smoothness)
- c) Pipe blockages
- d) Pipe bursts and leaks
- e) Water diversions and exertion of high suction pressures on the distribution network.

2.2.4 Water Theft

This refers to the use of piped water without consent of the authorities mandated to distribute it. In other words, the perpetrators access the water from the distribution network illegally [34]. Such activities could take the form of meter by-pass, illegal connection, meter tampering and meter reversals among others.

A brief description of these forms of theft is given as follows.

2.2.4.1 Meter Tampering

This involves meter alteration or removal or generally corrupting the meter to render it useless [35]. The capability of taking water consumption readings is compromised. Such acts may include but not limited to; melting meters' internal components by applying some heat, distortion of the internal mechanical gears that affect the readings and impurity addition among others [34].

2.2.4.2 Meter reversals:

This involves reversing the operation of the meter by installing it in the reverse pattern. The meter readings are then recorded in the decreasing manner. The readings can also be reversed manually.

2.2.4.3 Illegal Connections

Consumers connect diversion pipes on the supply network. This may occur under a number of cases; when the client has been disconnected due to non-payment, when the perpetrator is not a client but has the service pipe passing by their premises, the client simply wants to minimize the total bill taken by their service meters.

2.2.4.4 Meter Bypass.

In cases where the client is legally connected but would want to minimize the meter readings, they may connect a pipe in-between the meter connectors so as to limit the meter readings. With the use of suction pumps, more water is drawn from the network rendering limited supply or no supply to other clients on the network [36]. Meter bypass could also involve removing the water meter and fixing a bypass pipe also called a "T-Junction pipe.

2.2.5 Water Meters

Water meter also referred to as flow meter is a device or instrument used for the measurement of the liquid or gas's linear, nonlinear or volumetric flowrate. The choice of meters to be used in a water distribution company is dependent on the technicality and expertise of the staff, the ease to calibrate and maintain the meters, the availability of the spare parts in case replacement is required and the user reviews on the performance [37]. The sole purpose of the meters is billing. However, with the current issues of water theft, researchers have formulated various platforms for remote billing such as the cayenne application [30] [36], and the use of prepaid meters. The main qualities of a physical

meter that can directly influence it's the user's choice to select it are; the operating principles of the meters, the accuracy levels of the meters, the technicalities in its installation and disconnection, tolerance to particles and debris, the pipe diameter and flow rate of the water, the flow type which could be turbulent or laminar flow, the environmental temperatures, meter durability and ability to calibrate and maintain it, among others [38].

The existing water meters can be categorized into Velocity and positive displacement meters as discussed below.

2.2.5.1 Positive Displacement Meters

The flow of water induces movement of a tiny compartment with a specified volume of a fluid. The compartments are filled and emptied periodically and they are the basis for the operation of positive displacement meters. The number of rounds of filling and emptying the compartments is a basis for the calculation of the flowrate. The volume of water passing through the meter is registered by the gears rotated by the piston which is in motion. The piston and Nutating disc are the variants of water meters [39]. Due to their high sensitivity to lower flowrates, positive displacement meters have considerably high accuracies for a wide flowrate range. These meters are commonly used in residential and small business facilities [39].

2.2.5.2 Velocity Meters

The operation of velocity meters hinges on of principles that states that a fluid flowing through a specified cross-sectional area with a known velocity can be equated into volume of flow. Applications that involve high flowrates require this type of meters. The various types of velocity meters are; ultrasonic, turbine, propeller, multi-jet, venture, and orifice meters [40].

2.2.6 Water distribution

Water distribution entails the pipeline and the various connections that ensure water supply to the clients [35]. This process involves drawing water from the water source, treating the water in the treatment plant, storing it in the reservoirs and allowing it to flow into the supply system either through gravitational flow or by pumping. The ingress in the pipe surfaces may affect the water through contamination if the pressure of flow is not regulated within certain limits. This therefore ensures that the right pressure is maintained for distribution [30] [33].

2.2.6.1 Water Reservoirs

After water treatment, water is stored in reservoirs which could be overhead tanks used in gravitation flow or underground tanks which are used in cases of pumping. Reservoirs play a big role in the water distribution cycle by; ensuring a steady continued and interrupted supply to meet the demands of the clients, keeping the water levels high enough to uniform water pressure distribution, ensuring water availability even in cases of vandalism of the supply pipes to the reservoir from the water source in cases of road construction and other works and lastly allowing the operation of pumps at uniform rates to output the necessary pressure required to drive the water through the network [33].

Some of the terminologies used in reservoir management include; *Minimum water level* which refers to the minimum level of water in the reservoir that is sufficient enough to provide the threshold residual pressure necessary to drive water to the extreme end of the distribution system, *Maximum level of water* which is the maximum possible level the water can assume in the reservoir, *Working pressure* which refers to the least possible pressure at which the distribution network can perform, and lastly *Safe working pressure* which is the product of the safety factor and the actual working pressure.

Reservoir classification is based on functionality, model of operation, construction material and position in relation to the ground. These include;

2.2.6.1.1 Elevated Reservoirs

These are reservoirs that are placed at an elevation from the ground. This could be overhead tanks placed on raised surfaces/towers and raised ground such as hills. When the height of the reservoir is much greater than the diameter reservoir, we refer to such reservoirs as stand pipes.

2.2.6.1.2 Ground Level Reservoirs

These may be made of plastic or concrete and placed underground. Water is pumped into the distribution system using pumps.

2.2.6.2 Reservoir Operation

The operation of the reservoirs varies basing on the mode of filling and distribution of water into the supply network, as discussed [33]:

Floating on the line Reservoir - In this kind of operation, filling the tank and distributing water is done simultaneously. Therefore, the reservoir inlet and

distribution system are all open hence supply to the system and reservoir are done concurrently bearing in mind that under low water demands, then more water flows to the reservoir to as opposed to the distribution network and the during peak hours of high water consumption, inflow to the reservoir is cut off. Instead, more water from the reservoir flows to the distribution system at the same. With this method, continued pumping is required hence the risk of limited flow is so high.

Fill first and draw Reservoir – The reservoir is filled first but distribution. After filling up the reservoir, the outlets are opened and water flows freely under gravity into the distribution network. The tank is normally situated closer to the source of the water to avoid losses due to friction. The pumping capacity is high in this system therefore the reservoirs are filled at faster rates compared to the former [33].

2.2.6.3 Types of water distribution systems

The distribution systems are categorized into Branch and loop.

Branch: The branch networks basically find their applications in water distribution to low capacity consumers who are mostly residential and small business apartments and generally have few connections. This kind of network makes it possible to connect new clients.

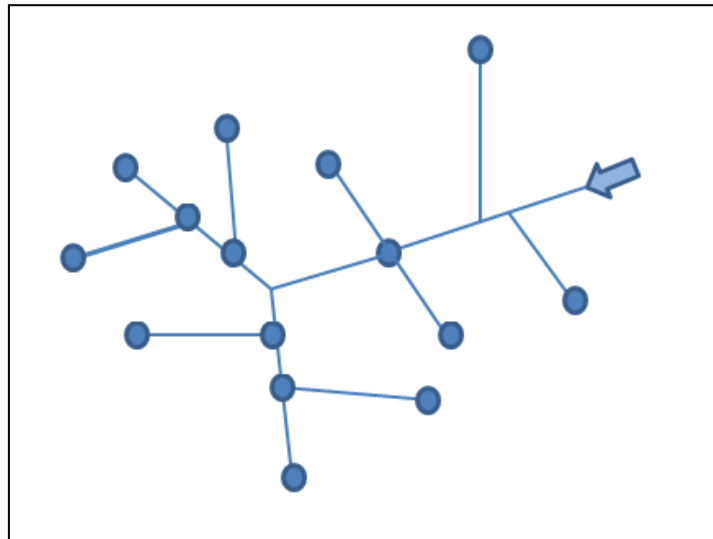


Figure 2: Branch Distribution Network

Loop/grid: The loop network normally comprises of some secondary connection distribution lines which may form branched architecture, single loop or multiple loops. Water from these connection pipelines is supplied to the distribution/service pipes and

finally to the consumers [42].

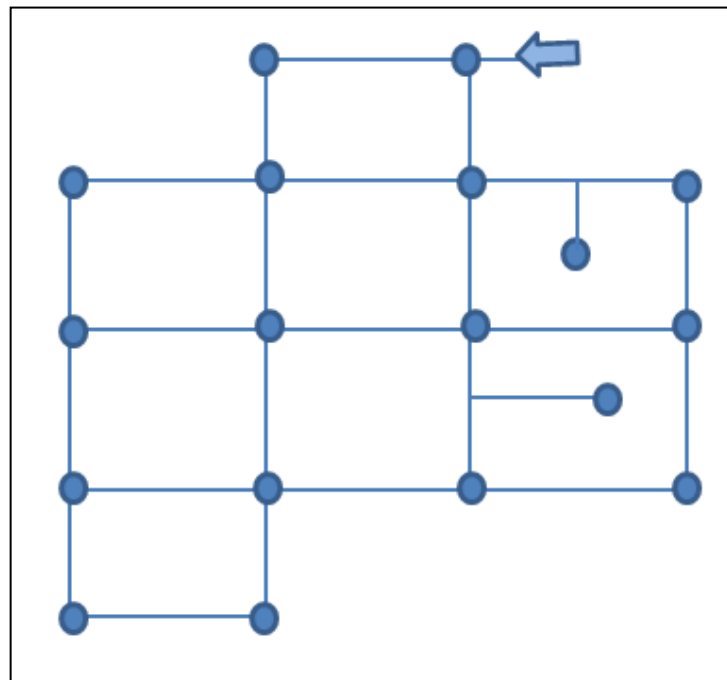


Figure 3: Grid Distribution Network

In a typical domestic connection, the service water pipe involves connecting the indoor water system to various taps which could be in the washroom, kitchen, dining room, etc. The commonly used taps are those of 9mm and 12mm. Such a layout is shown in figure 4 below.

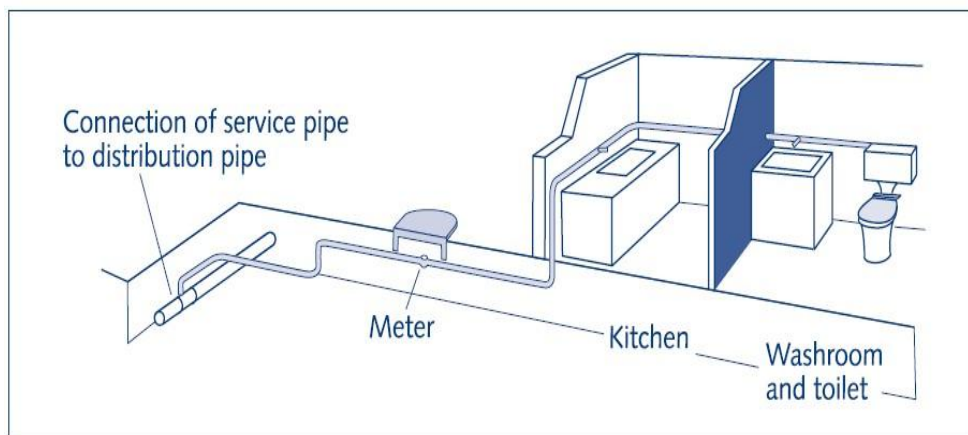


Figure 4: In-house piped water connection

The pipe that supplies water to the home (also called service piped) is connected to water supply mains using small diameter pipes called T-piece and a *ferrule* which is a small piece used for inserting the pipe or occasionally a larger sized pipe called saddle is used. These connectors are normally necessary when using ductile and cast iron pipes [41].

2.2.7 Water Distribution Pressure

In water distribution systems, pressure is what drives water from one area to another so that it can reach the various destinations of the distribution chain. Pressure can be understood in mathematical terms as the force that is exerted for every unit of the area of exertion. Denoting force as F and unit area as A , pressure, P is given by the equation;

$$P = \frac{F}{A} \quad (2.2)$$

Two kinds of pressure are used in water distribution systems, that is; dynamic and static water pressure [33].

Static Water Pressure

Stationery water exerts a pressure in the distribution system which is the potential pressure indicator that the system possesses. This kind of pressure exerted in the system when there is no flow is referred to as static water pressure [33]. To produce the static water pressure, place the water tank on an elevated surface above the surface of water use. This kind of pressure is used in reservoirs placed at higher levels from the ground. Pumping can also be used to supply water with some energy. And lastly, use of air pressure.

Dynamic Water Pressure

When a known quantity of water flows past a given point in the pipe, the pressure at any point along that pipe is what is referred to as the dynamic pressure of water. The difference between dynamic pressure and static pressure lies in the fact that dynamic pressure of the within a distribution system varies from one point to another. This is brought about by friction during flow [33].

2.2.8 Pressure Loss

There are moments in the distribution system when there is a significant pressure drop to below the threshold value [42]. Such scenarios can be defined as pressure loss events whose occurrence often happens without planning though at times, it may be a planned task which occur at any time. A scenario of planned pressure loss events could be during installation, replacement or repairs in the distribution network. On the other hand, unplanned occurrences that could lead to such events may include power outages, leakages by reservoirs, broken pipes, and failure in the pumps, high consumption rates

which could be legal or illegal for example and when suction pumps are connected to the system among other reasons [42].

The major factors that determine the pressure available in the distribution network are; the topography of the pressure zone, water availability in the reservoirs and at the source and the state of the supply channels (pipes) [41].

To ensure that a constant pressure is maintained within the system, some precautions have to be taken. Among them include;

- ii) Constant filling of reservoirs with water
- iii) The distribution channels have to be well maintained. Illegal connections, pipe bursts and leaks, and debris have to be addressed before affecting the distribution.
- iv) The pumps have to be switched on and off in a controlled manner to avoid surges. This ensures a gradual change in water flow without affecting the pipes.
- v) Using gate valves, pressure can be well regulated to meet the demands.

2.2.9 Uphill and downhill Pressure Maintenance

The area supplied with water by the distribution network requires a minimum pressure to drive the water across the network. For the raised reservoirs where water flows freely under gravity, a height of at least 120ft is required to generate the necessary pressure. This elevation determines what we refer to as the *pressure zone* which is an area within a given geographical setting that uses same pressure to drive water across the network. Note that water flow from one pressure zone to the other is only possible from higher to lower pressure zone through a valve called the pressure reduction valve [43]. Within a pressure zone, the pressure reduction valves can be added to create a pressure district. Since leakage and aging are partly a result of very high pressures, a distribution pressure of about 40 to 80psi is maintained with the help of pressure reduction valves. Another alternative is to use gate valves especially in very high pressure zones such as outlet points of reservoirs and overhead storage tanks.

CHAPTER 3: LITERATURE REVIEW

3.0 Chapter Summary

This chapter gives a summary of the related work and the various applications that are currently being applied for water theft detection. A brief discussion of the Ensemble learning techniques has also been presented. The chapter is concluded with an introduction to Decision Trees.

3.1 Related Work

There have been various contributions by different researchers formulating deeper understanding of the water theft problem and how to overcome it. A significant number of approaches for detecting water theft have been proposed by [30] [36] [44] [45] [46]. The existing work present a water theft detection using hardware prototypes consisting of flowrate sensors and some kind of control which has been majorly PLC (Programmable Logic Circuit).

Some initiatives have utilized IoT technology to access the remote sensor data collected by flowrate sensors and store it in remote database that serves a web application [47] [48]. With the ability to remotely monitor and control what happens on the distribution network, the volume consumption values can be acquired and stored such that the meter readings taken from the consumer's water meter can be compared to those taken by the web application. Other technologies such as SCADA and fuzzy logic have also been proposed for remote monitoring of the distribution system [44] [47] [48]. The major drawback with these approaches is the failure to audit the water supply and detect theft in real-time since water audit can only be done after a period of time. Hence this cannot rule out other illegal practices that the consumers may do during the course of the water distribution. Cases of illegal connections are detected using the flowrate differences between two flowrate sensors placed at strategic branch points along the network [49] [50] [51].

The sophistication of the water theft problem and the calibration errors within such sensors makes it the entire approach prone to errors as the theft detection is heavily dependent on the supply pressure which may not be well regulated due to various factors such as failure to ensure that the reservoirs are always full of water, inefficiency in water

pumps for driving water into the distribution system, booster pumps for additional pressure enhancement in the system, water leaks and pipe bursts among others.

The motivation behind these works is that water distribution companies greatly suffer from wastage of water by the users which leads to loss of incomes, compromise in water quality and lack of sustainability in water distribution. The works of [44] aims at developing mechanisms for monitoring water distribution and usage along the service pipes. The work looks at ways of minimizing water wastage hence propose a system for automating water storage and distribution to ensure sustainability. Different sensors such as water level sensors, and flow rate sensors were used to capture data of water level and flow rates respectively. The work also proposed SCADA and Fuzzy technologies to monitor water storage and distribution. The major contribution of their work can be viewed in the ability to address overflow from the overhead tank, leakage and empty flow experienced in traditional methods by regulating the level of water and distribution pressure.

Due to urbanization resulting partly from increasing population growth, the need for safe piped water is high hence managing water distribution efficiently can serve as one of the mechanisms for ensuring a steady supply to all the consumers [45]. A framework for automatic monitoring of water supply and storage has been proposed in [45], which is closely related to that of [44], where a network of flow rate and water level sensors are used to monitor the distribution pressure and level of water in the storage tank/reservoir respectively. The sensors are interfaced with PLC controlled with a software program which is rule based. Remote monitoring has been enabled by use of SCADA. Other components of the network that have been used include proximity sensors for theft detection, solenoid valve for turning off the water supply once an error is detected, and pumping systems which include stations for pumping, water treatment, storage and distribution. Conclusively, water wastage resulting from poor usage, water theft and leakages has been discussed. A closely related approach is presented in [46] in which water automation control using Programmable Logic Controller (PLC) as the central processing unit has been emphasized. The water level sensors interfaced with PLC have been used to detect the level of water in the storage tank. A disconnection is done using the solenoid valve in case of overflow. Pressure manipulation in the distribution system is detected by the time base feature which aids in detecting water theft.

The notion that if a specific volume of water is supplied within a stipulated period of time, any fluctuations in the distribution pressure can easily be detected has been applied in [49]. The system makes use of PLC ladder, solenoid valve, high and water level sensors to detect the amount of water in the tank, flow rate sensors for theft and leakage detection basing on the flow rate differences between two sensors. The flow transmitter captures the volume within a given service pipe and relays this data to a central remote database for billing. Additionally, allowing just enough amount of water in the distribution network to ensure proper utilization can be used control and monitor water distribution [50]. This idea has been used in [50] to design a control water supply into the distribution network using the water valve and relay interfaced with a controller. Flow rate sensors and flow switch have been positioned at every home unit controlled by Arduino microcontroller. The water supplied and consumed is calculated using flow data captured by the flow meters obtained over time [50].

A differing initiative from the above is that undertaken in [51]. Water quality is an aspect that is crucial in the water distribution cycle. The process of water treatment, storage and pumping to the distribution system can impact water quality directly or indirectly [51]. Unlike the work in [44] [45] [46] [49] [50], [51] proposes a real-time water distribution and quality monitoring system to ensure continuous uninterrupted supply. The proposed flow monitoring system comprises of three units; (i) automated water distribution unit which turns off the pump when the water level in the tank is below a certain minimum level, (ii) water theft detection unit which makes use of flow rate sensors for monitoring water flow and in case of abnormal flow, the solenoid valve turns off the water supply and (iii) PH unit which uses PH electrodes to measure water PH. A network of sensors is used for measuring the chemical and physical parameters of water such as turbidity, PH, temperature, conductivity and dissolved oxygen. The sensor data is received and interpreted by ARM core controller. Since theft and leakage leads to variations in these parameters which conversely alter water quality, a flow monitoring system has been proposed [51].

A related research has been undertaken by [52]. The work proposes a framework for water distribution monitoring. The flow rate sensors are interfaced with PLC to monitor flow rate fluctuations hence detect theft. The system comprises of a network of GSM, PLC and PC system, PH sensor, Remote Terminal Units (RTUs), actuators and

transducers. The data acquired through SCADA (Supervisory Control and Data Acquisition) can give real-time state of the distribution network.

In [53] atomization is used for monitoring water distribution and anti-theft management. The authors design a self-power generation to use valves for automatic opening of the overhead tank. Two flow rate sensors have been used to take record of water flow patterns and hence use the data to detect theft. The supply then cuts off the solenoid valve once theft is detected. Other additional functional requirements met by the system include; water sedimentation, water filtering using granular media filter and chlorination/treatment. Another initiative for water supply monitoring has been undertaken by [54]. Water distribution is monitor basing on flow variations recorded by flowrate sensors. The authors make use Remote Terminal Units (RTUs) transducers and actuators placed over a given pressure zone. Data is acquired through SCADA technology. HMI running on a remote server computer is used for monitoring network for water distribution in addition to remote control tasks using OPC technology and data transfer using WAN wireless communication [54].

Regulating water consumption at client end at intervals of specific time can also be used as a methodology for water theft detection [55]. In their work, flow rate measurements exceeding set limit turns on the solenoid valve to shutdown water supply. The GSM communication is used for sending alert sms to the control center in case of any suspicious change in the network state. However, in [56], the GPS is used to obtain the precise location of the water theft after receiving a sms alert via GSM.

Kinge and Pranam [57] designed and implemented an application for automatic billing (smart meter) and prepaid recharge. The network of flow rate sensors is used to detect high pressures and using the Solenoid valve, the distribution network is cut off when the pressures go beyond a certain nominal value. A web application is used for billing purposes. Due to high population growth leading to increased water demand, such an application can help to reduce the amount of non-revenue water. Soundhrya [58] proposed a management system for water distribution. A prototype for water supply monitoring and control consisting of a network of flowrate sensors, actuators, piping and valves have been included for demonstration purposes.

The high level of water misuse/wastage has also contributed to its scarcity [59]. A better management technique through automation methods can enable quick fraud detection.

Image analysis has been used to discover water meter fraud [59]. A three step approach has been used in their implementation. (i) OPF classifier and HOG descriptor have been used to detect water meter fraud. (ii) Segmentation and morphological methods and image processing have been used to detect seals (iii) Assessing the condition of the water meter by fraud classification. The water meter inspection dataset comprising of various water meter images has been used to validate the framework resulting into an accuracy of 81.29%. Computer vision strategy is therefore a promising area that can be used for fraud detection in water meters. Their work tackles a very crucial aspect of water theft which tends to be more sophisticated. This methods has been applied in a various filed such as pre –paid electricity to detect fraud in electricity meters.

Related work by [47] proposes a server-end monitoring and control framework for piped water. Arduino was used as a minicomputer, flow rate sensor for flow rate monitoring and solenoid valve for disconnection. A specialized web application called Cayenne for billing and payments has been designed and implemented in their work. The data is relayed to the remote database that serves the web application via Ethernet Shield VI. Wiznet W5100 provides IP stack for handling TCP and UDP packets.

The work in [60] presents a case study on model based procedures for detection of attacks on cyber-physical systems (CPS). A simulation tool for distribution systems of water called EPANET (software program used by American EPA to track water flow and pressure at every note of the pipe, water level in the reservoir and chemical concentration within the distribution network) [24] has been used to simulate a water supply Network. A Linear time Invariant (LTI) model which is an input-output model was obtained. This was based on the simulation data and subspace techniques of Identification. System dynamics are estimated using Kalman filter derived by the LTI model and the data that comes from EPANET and Kalman filter estimates have a difference which produces what is referred to as residual variables. For attack detection, change detection procedures involving dynamic Cumulative Sum (CUSUM), residual variables and bad-data are used. The model is validated using subspace Identification method.

The challenge of excess water sucking using pumps is a major issue in most water distribution networks around urban centers [61]. Their research uses flow rate readings that are taken by the sensors and the deviations from the normal flow are calculated and used to monitor theft. Such remote techniques can be used to detect but also prevent theft in an early manner. The system comprising of microcontroller, flow rate sensor and

solenoid valve is deployed at the consumer's end. Data captured is transmitted to the officer's in-charge by sms by GSM modem.

Water leakage, distribution and theft ought to be monitored to ensure efficiency in water supply and minimize losses. [48] Proposes a smart city IoT based system for monitoring water distribution to detect leakage and water theft. A number of services are rendered by the proposed system which integrates Cloud and fog computing. These services include; detection of water theft, fault prediction and localization, quality control, consumer utilization. Their work entails design of water distribution network for analysis using EPANET, water distribution monitoring using sensor network interconnected with IoT, use of sensor connectivity for monitoring underground pipes, Integration of cloud and fog computing to design IoT based architecture for water distribution monitoring. Their work makes use of flow rate sensors for pressure monitor, external vibration sensor equipped with MI transceiver and pressure sensors alongside M1 Hub. RF based communication is used for device-device communication over the ground and IP-based communication is used for data transmission over the cloud to central server.

The water content and mineral levels are also monitored using sensor networks situated underground, whereas the volumetric contents are obtained using AG-II sensors. To obtain data from various sensors, CR500 data logger has been used. With this, up to 40 sensors can be used connected.

3.2 Existing Water theft Applications and Prevention

3.2.1 Main metering

The main metering is the traditional approach to water billing and theft detection. After a specified period of time, say a month, the meter readings are taken and compared with the volume supply to that given service pipe. Any discrepancies in the volume readings are interpreted accordingly hence the existence or nonexistence of theft can be detected. Any negative deviation will prompt the authorities to investigate the matter. These systems are the initiative of JICA (Japanese International Corporation Agency) and Yangon City is one of the cities whose usage has been applied. Under the Technical Transfer Program, over 300 meters were to be installed in 280 homes, schools and other residential places by the end of the program. The recording capacity of these meters in wider areas is quite high [62]. The major drawback of this approach is the inability for timely remote monitoring and alerting.

3.2.2 Prepaid metering Systems

With the introduction of prepaid meters, clients are able to procure the quantity of water of their choice prior to usage. Water dispensers under the prepaid systems are able to control water flow based on the client's purchased volume and needs. This form of metering and theft control makes use of three basic components, that is; a prepaid card, internal unit and meter. Three basic modules make the prepaid system, these are; latching valve, electronic model and water meter having a pulse output [63]. The advantageous benefits of prepaid system can be seen in the ability to regulate water consumption by the consumers, leakages can easily be identified based on the amount consumed and that supplied, the credit at the instant can be known, the state of the network can easily be traced and finally, the simplicity of smart card ensures efficient use of water systems.

Conversely, the prepaid systems have some limitations such as meter tampering since the prepaid meters are not theft proof, poor network may interrupt data access from the remotely controlled prepaid meters, the cost of the meters is quite high for the low income earners, remote alerting is not possible and lastly, cases such as meter reversals and illegal connections still remain unsolved.

3.2.3 Physical Monitoring

Most water distribution companies have employed physical methods for piped water theft detection. In Uganda, this is the major approach to detecting water fraud used by National Water and Sewerage Corporation. Such physical methods are herein discussed as follows.

3.2.3.1 Meter bypass detection

Certain procedures have been laid out to detect a meter bypass scenario. These measures involve analyzing the meter and the connecting sides with variations in flow to view the flow patterns which can give a hint on the state of the flow through and past the meter.

Closing the meter's stop clock would mean stopping flow through the meter. However, if this is done and there still exists a vibration in the connecting pipe to the meter, then this implies an alternative path has been created which is a clear indication of a meter bypass or T-junction

The ball valve below the water tank once pressed stops water flow into the tank if a single path of water is available. However, in cases where a meter bypass exists,

pressing the ball valve will not stop the flow, instead, water will flow and reach the roof tank which gives an indication that a meter bypass exists.

The legal distribution lines once closed are expected to cut off water supply to the entire building. However, this is not the case under scenarios where meter bypass exists. If water still flows in the taps even after closure of the supply valve, then there may be a meter bypass. Note that water flow from the taps cannot take longer if indeed there is not alternative flow despite the fact that back flow from the pipes can also be the reason for flow in the taps.

Reversing the above procedure and opening all the supply lines to the building would ensure water supply to all the taps. This may not be the case if there exists a meter bypass. Some tap(s) may not have supply in cases where the customers have bypassed the meter. This could be because of the closure of a gate valve.

The process can again be reversed and those taps that previously didn't have water are checked. The closure of the control valve will be affirmed if all the taps at this point have water flowing through them [34].

3.2.3.2 Detecting Illegal Connections

Water supply companies greatly depend on police and communities to detect cases of illegal connections [34]. In most countries, a penalty is levied on such a client that is caught with an illegal connection on their service pipe. This encourages community vigilance and corporation with the water supply companies.

Another technique of illegal connection practiced by consumers is where the water is drawn just before the meter. With a clear record of clients with consistent and inconsistent consumption rates, a surprise visit can be made to such clients. Such a water theft method is normally perpetrated over weekends and public holidays when the client least expects the monitoring team patrol through the area. Therefore, such days can be utilized to visit the clients.

3.2.3.3 Meter reversal

Meter reversal is the commonest water theft method used by most customers. A surprise check to the clients' meters can ensure the meter is not reversed. The meter readings are taken by the pointer which is reversed. This is more visible and less sophisticated to detect if the monitoring is done at the right time. The flow direction can also be indicated

by the stop cork position. Additionally, having extremely lower meter readings than the previous readings may indicate on the existence of a meter reversal.

3.2.3.4 Meter Tampering

Meter tampering is very sophisticated to detect because there is no clear indicative and confirmatory signs to the problem. However, physical marks on the meter are used for the initial detection of water theft by this approach. Firstly, If a meter is being manipulated, it may appear shiny since its being held over time. Secondly, the meter fittings are always tight if the meter is not being tampered with. However, this may not be the case if the customers keep on trying to tamper with the meter. The meter fittings tend to loosen and if this is observed, then there is a reason to suspect that meter tampering would be going on. And thirdly, the fittings on the pipe will tend to also loosen. Marks of pipe wrench may also be observed.

Generally, the physical method of water theft detection bases on a number of assumptions that may not give conclusive evidence of the theft occurrence. Most of the physical monitoring methods involve the presence of the monitoring team at the client's premises. It is practically difficult to detect theft since customers who indulge in such acts are also sophisticated. And lastly, with the massive corruption levels in most countries, use of police and communities is too ineffective as majority of them can easily be bribed.

3.3 Ensemble Methods

It is possible to combine various techniques of Machine Learning to form one predictive model with the intent of decreasing the variance referred to as bagging, improving prediction performance referred to as stacking or bias which is referred to as boosting. This is what is called Ensemble [64]. With Ensemble methods, the same problem can be solved by training multiple learners. Different learners are constructed singly and then combined as opposed to the ordinary learning techniques that focus on constructing single learner from the dataset used for training [64]. Ensemble learning is also referred to as the *multiple classifier systems* or *committee-based learning*.

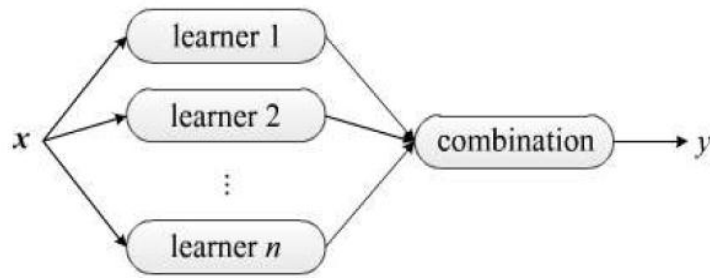


Figure 5: Common Ensemble architecture

The Ensemble Architecture comprises of a set of learners called *base learners* which are normally generated by base learning algorithm. The generation of the base learners is done from the raining data. The base learning Algorithm used in this case could be neural network, decision tress or other types of algorithms used in learning. Some learning algorithms (single base) are used by most ensemble methods in producing base learners that are homogenous in nature. Such base learners are called homogenous because they are of similar type. The base learners that are produced result into ensembles that are also homogenous [65]. Heterogeneous ensembles are also produced through other methods though in such cases, there no learning algorithm (single base) exist hence such leaners are sometimes referred to as *component* or *Individual learners* instead of calling them base learners.

Ensembles differ from base learners in their generalization ability which is much stronger compared to base learners'. The need for ensembles can be best seen in their ability to improve the performance of the weak leaners. The ability to combine several models by Ensemble learning is crucial in improving the accuracy of results produced by machine learning approaches. The predictive performance of the ensemble methods is higher compared to that of single models. However, it should be noted that to attain the highest accuracy in a given ensemble compared to the Individual members, the accuracy and diversity of the base leaners has to be guaranteed.

There are two major categories of ensemble methods, that is; *sequential* and *parallel* ensemble methods.

Sequential ensemble methods are those where the generation of base leaners takes a sequential pattern. An example is the ***AdaBoost (Adaptive Boosting)***. In these methods, the dependence that exists between base leaners is used as the basis for generation of the base leaners. Boosting the overall performance can be achieved by obtaining the high weight examples that were mislabeled and attaining their weight.

Parallel ensemble methods on the other hand involve generating base learners in parallel. An example of such methods is the random forest. Unlike the sequential ensemble, the independence relationship of the base learners is exploited in the parallel ensemble. This is possible because through averaging, the error reduction to the lowest value can be achieved.

3.3.1 Bagging

This stands for bootstrap aggregation. An estimate's variance can be reduced by obtaining an average of various estimates. An ensemble $f(x)$ can be computed from training m number of disparate trees using divergent datasets that are selected at random with replacement.

$$f(x) = 1/M \sum_{m=1}^M f_m(x) \quad (3.1)$$

In base learner training, a sampling technique called bootstrap is used for selecting the subsets of data used to train base learners. Bagging further utilizes the voting procedure in classification task and averaging procedure in regression to build up the output of resulting from the base learners. In cases where the learners are stable, there is no necessity to combine them because the performance of generalization cannot be improved by the ensemble methods.

One of the commonest categories of ensemble methods is the random forest in which the process of constructing the ensemble involves building each tree using the bootstrap sample obtained from the training set. The features are also randomly selected to form a subset of the features; that is, not all features are selected hence these makes the tree more random. This leads to a slight increase of the forest's bias and decrease of its variance as a result of obtaining the average of the trees that are less correlated to each other [66].

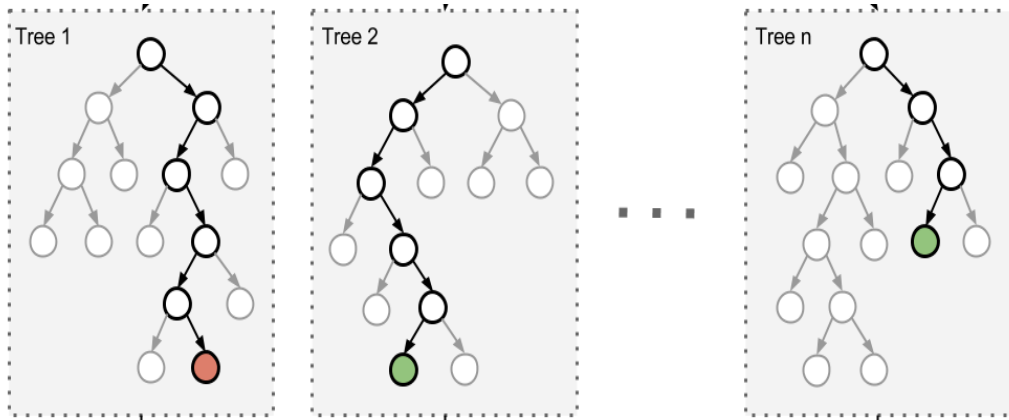


Figure 6: Bagging Example

The randomization of the splitting thresholds leads to increase in random state of the algorithm in trees that are highly randomized.

The decision boundary visualization of one decision tree with Bagging and the component DT on the three Gaussians dataset has been done to better understand the concept of Bagging. The observation made from the graphical representations affirms that the flexibility of the bagging decision boundary being greater than that of one DT. This helps in error reduction by about 1% from 9.4% of a single DT.

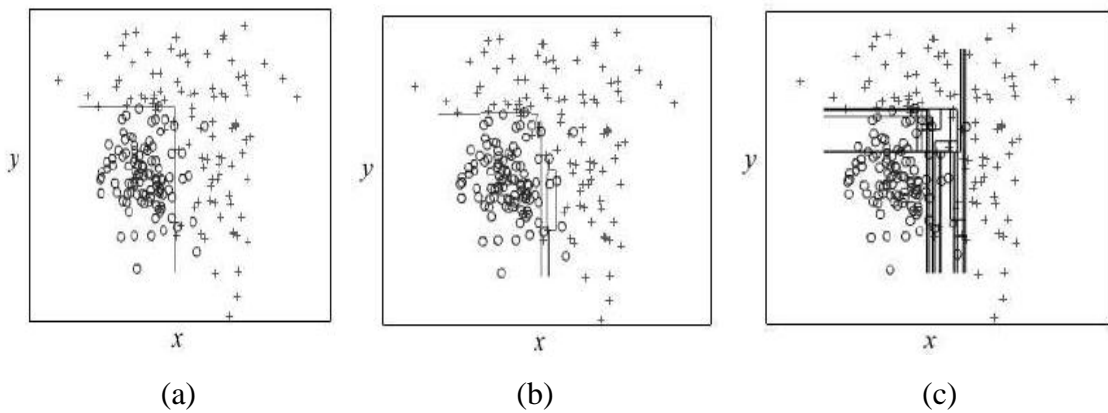


Figure 7: Decision boundaries of (a) single DT (b) Bagging (c) 10 DTs used by Bagging on the three Gaussians dataset

3.3.2 Boosting

Converting weak learners to strong learners involves a certain category of algorithms under which “boosting” lies. Boosting works on the principle of taking a list of weak learners and fitting them to weighted versions of data. The weak learners denote those models quite better than the randomized guess. Such an example is the small DTs. Some examples that could have been wrongly classified in the preceding rounds are assigned more weight. To obtain the final prediction, a combination of the independent

predictions is done through classification (weighted majority voting) or regression (the weighted summation).

AdaBoost is the most widely used category of the boosting algorithms. Its Algorithm is given below.

Adaptive Boosting algorithm

1. Initialize data weights $\{w_n\}$ to $1/N$
 2. **for** $m = 1$ to M **do**
 3. fit a classifier $y_m(x)$ by minimizing weighted error function J_m :
 4. $J_m = \sum_{n=1}^N w_n^{(m)} 1[y_m(x_n) \neq t_n]$
 5. compute $\epsilon_m = \sum_{n=1}^N w_n^{(m)} 1[y_m(x_n) \neq t_n] / \sum_{n=1}^N w_n^{(m)}$
 6. evaluate $\alpha_m = \log\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$
 7. update the data weights: $w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m 1[y_m(x_n) \neq t_n]\}$
 8. **end for**
 9. Make predictions using the final mode: $Y_M(x) = \text{sign}\left(\sum_{m=1}^M \alpha_m y_m(x)\right)$
-

The AdaBoost Algorithm illustrates the classification of data basing on the threshold of the feature variables in every base learner comprising of a DT whose depth is one. The feature threshold that determines the classification divides the space into two subspaces in which a linear decision plane separates them and the plane is in parallel with one of the two axes.

Gradient Tree Boosting refers to the rationalization of boosting to random differentiable loss functions. This can be employed in classification and regression problems. Gradient Boosting uses the sequential approach to build the model.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (3.2)$$

For every stage, the DT $h_m(x)$ is selected to reduce the loss function L for the given current model $F_{m-1}(x)$.

$$F_m(x) = F_{m-1}(x) + \underset{h}{\text{argmin}} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + h(x_i)) \quad (3.3)$$

The distinguishing feature of regression and classification algorithms lies in the class of the loss function utilized.

3.3.3 Stacking

Stacking uses meta-regressor and classifier to combine multiple models, that is classification and regression model respectively. It is an example of ensemble learning in which the training of the models at base level depends on a whole training set. On the other hand, the training of the meta-model is done on the features of the base level model results. Since the base level comprises of various learning algorithms, this type of ensemble learning is heterogeneous.

Below is the stacking algorithm.

Algorithm

1. Input: training data $D = \{x_i, y_i\}_{i=1}^m$
 2. Output: ensemble classifier H
 3. *Step 1: learn base level classifiers*
 4. **for** $t = 1$ to T **do**
 5. learn h_t based on D
 6. **end for**
 7. *Step: construct new dataset of predictions*
 8. **for** $i = 1$ to m **do**
 9. $D_h = \{x'_i, y_i\}$, where $x'_i = \{h_1(x_i), \dots, h_T(x_i)\}$
 10. **end for**
 11. *Step 3: learn a meta-classifier*
 12. Learn H based on D_h
 13. return H
-

3.4 Decision Trees

Divide and conquer is the technique used in Decision Trees (DT) for the performance of the constituent set of decision tests organized in a tree structured format. The split test feature is related to each and every node other than the leaf node. Split in DTs refers to the process of dividing node data into multiple various subsets based on the different values indicated on their feature test. The node has some instances that may fall on it and such instances are assigned the different labels associated to the leaf nodes [64]. During prediction tasks, the root node is the starting point from which feature tests are done. Once the leaf node is arrived at, the result is achieved. In the DT tree example shown in figure 8, the starting point of classification is at the evaluation of the y-coordinate feature whose value if larger than 0.73, then the classification of the instance will be denoted as “cross”. However, if the value is less than 0.73, then the x-coordinate feature whose value if greater than 0.64, the classification of the instance will again be “cross” and if

the value is less than 0.64, then the classification is “circle”. Note that DT learning algorithms are recursive because the split selection is done for the dataset at every step and used in dividing dataset into subgroups where ever subgroup is used as the dataset for the proceeding step. The way of selecting the splits is the core of DT algorithms.

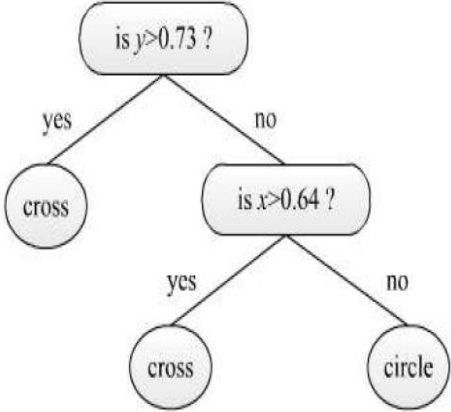


Figure 8: Decision Tree Example

CHAPTER 4: METHODOLOGY

4.0 Chapter Summary

This chapter discusses the approach and techniques used to undertake the project work. A description of the RF classifier and other benchmark techniques such as LR, KNN and SVM has been given. This is followed by system design and experimental procedure for data collection. The description of the various components used in the design of the hardware prototype has also been presented. The chapter is concluded with some data analysis and preprocessing.

4.1 Materials and Methods (Description of the algorithms)

This section gives a thorough discussion of the supervised machine learning algorithms used in the study.

4.1.1 Random Forests

Random Forests (RF) are supervised learning algorithms which use ensemble learning to build various small DTs with small number of features which is computationally affordable into a stronger learner [67]. This is done in parallel manner using the majority vote. Generally, RFs have proven to achieve the highest accuracy among all the learning algorithms. The Pseudocode of the RF algorithm is show below.

4.1.1.1 Algorithm 1 Random Forest

Precondition: A training set $S = (x_1, y_1) \dots (x_n, y_n)$, features F , and number of trees in forest B .

```
1 function RandomForest( $S, F$ )
2  $H \leftarrow \emptyset$ 
3 for  $i \in 1, \dots, B$  do
4  $S(i) \leftarrow$  A bootstrap sample from  $S$ 
5  $hi \leftarrow$  RandomizedTreeLearn( $S(i), F$ )
6  $H \leftarrow H \cup \{hi\}$ 
7 end for
8 return  $H$ 
9 end function
10 function RandomizedTreeLearn( $S, F$ )
11 At each node:
12  $f \leftarrow$  very small subset of  $F$ 
```

```
13 Split on best feature in  $f$ 
14 return The learned tree
15 end function
```

The following procedure lays out the algorithm functionality.

Given a sample S , a subset of S known as bootstrap sample is chosen for every tree that is in the forest, $S(i)$ refers to the i th bootstrap. The modified version of the DT learning algorithm is then used for learning for the DT. The modification to the algorithm proceeds as follows; the feature subset $f \subseteq F$, where F is the total number of features, is randomly chosen in all the feature splits that can be made. This is done at every tree node and the splitting is then done basing on the best f feature instead of the features in F . Note that the selected subset, that is f is far much smaller than F . The most computationally demanding aspect of DTs lies in determining the feature on which to splitting can be based. The learning curve of the tree can be accelerated by minimizing the feature set.

4.1.1.2 Advantages of Random Forests

Algorithms that use bagging as an ensemble method tend to output results with reduced variance. In building the DT ensemble, Random Forest utilizes bagging in its operation. To understand the advantages of RF, it's plausible to justify why the performance of ensembles which involve randomized selection of subsets is way much better than that of traditional techniques. This is even way much better when the constituent models are not correlated. Splitting bootstrap samples in bagging (traditional) may make the individual DTs correlated since similar features may be used repetitively. Therefore this correlation between the DTs can be minimized by doing a restriction on each split and narrowing it down to only small sample features that are randomly selected.

Additionally, learning each DT quickly involves minimizing the number of features considered at every single node. This quickens the learning process and ultimately the construction of more trees. The less correlation between the trees leads to increased accuracy and therefore this justifies for the high performance of RFs.

4.1.1.3 Properties of Random Forests

Breiman [68] described a method to naturally rank variable importance in classification and regression problems.

Step 1 involves fitting a RF to data by measuring the value of the variable in the dataset $D_n = \{ (X_i, Y_i) \}_{i=1}^n$. A record for the “out of bag error” is taken for every data point and the average of the records is obtained over the entire forest. This is done during the process of fitting. Not that the errors on test set (independent) may be replaced so long as during model training, bagging was not applied.

Permutation is then done on training data and a computation of out of bag error is again done on the adjusted dataset. This is done after training the features value of j for the purpose of measuring the significance of the j th feature. The error (out of bag) difference prior and after doing a permutation over all the DTs is computed. Its average is used for the computation of the j th feature’s value score. Standard deviation of the error differences is then applied to normalize the score.

The high value features which produce this score are classified as more valuable scores and likewise those that result into lower values are ranked least important. The drawbacks of this method include; biasness of the RF on categorical variables that have disparate levels. Other alternative methods that can be used to address the above problem include; use of unbiased trees [69] and partial permutations [70][71]. In cases where data comprises of classes of related features of similar application for output, smaller groups are more advantaged compared to bigger ones [72].

4.1.1.4 Random Forest and K-Nearest Neighbour

Jeon and Lin [73] identified the connection between RF and K-Nearest Neighbour (KNN) algorithm which reveals weighted neighborhood property in both schemes. The prediction models are constructed from the set $\{(x_i, y_i)\}_{i=1}^n$ (training set) whose result of prediction \hat{y} looks at that point and its neighbours for x' new points and W , a formalization weight function.

$$\hat{y} = \sum_{i=1}^n W(x_i, x') y_i \quad (4.1)$$

The non-negative relativity in terms of weight of training point (i th point) relative to x' which is a new point is $W(x_i, x')$. The weight functions can be defined as follows.

If x_i is closest to x' which is among the k many points, weights $W(x_i, x') = \frac{1}{k}$ applies in KNN.

1. $W(x_i, x')$ is selection of part of the data used for training that and can be assigned to the same leaf just like x' , this applies to a tree.

Noting that a forest computes the averages of the m tree set of predictions, alongside W_j which are the functions of the weights, the following are the resulting predictions.

$$\hat{y} = \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n W_j(x_i, x') y_i = \sum_{i=1}^n \left(\frac{1}{m} \sum_{j=1}^m W_j(x_i, x') \right) y_i. \quad (4.2)$$

This affirms the neighborhood weightiness of the entire forest with weights that are a result of the individual weights. In this view, the points close to x' are the x_i points that are in same leaf node as the x' in a single tree or more within the forest.

4.2 Benchmark Methods

4.2.1 K-Nearest Neighbour (KNN)

KNN is a neighborhood-based classifier in which all cases are stored and the documents are assigned to the group with more neighbours (k nearest) [74]. It is alternatively called the instance based, example-based or lazy learning and sometimes also referred to as memory based reasoning. KNN has been widely used in various applications, among these include; statistical estimation, pattern recognition, etc. The classification process done by KNN is not dependent on the assumptions of probability distribution that underlies in the dataset. This is referred to as being non-parametric. KNN has two recognized variants, (i) nearest neighbour techniques that are unstructured: here, all the data is categorized into two, that is training and test samples. (ii) Nearest neighbour techniques that are structured: these base on a number of given data structures such as ball tree and Orthogonal Structure Tree (OST) [75]. The “nearest neighbour” concept in KNN originates from the approach used to perform the classification, that is; finding the least distance between the training to sample points.

In circumstances where we have data with continuous attributes, KNN is the best fit for classification tasks. To perform classification, KNN selects an instance which is not defined and the agreed numbers of instances used for training are selected. Secondly, the k -instances will have various classifications; hence the one that is commonest to all the instances is picked.

Given a number of various values of attributes, say m , two instances of those attributes are chosen and the closeness between them is determined using various techniques. Certain conditions have been found to apply in the different measures used in KNN.

Assuming u and v are two points and $x(u, v)$ is the distance between the points, then;

$$x(u, v) \geq 0 \text{ and } x(u, v) = 0 \text{ iff } u = v \quad (4.3)$$

$$x(u, v) = x(v, u) \quad (4.4)$$

$$x(u, r) \leq x(u, v) + \text{dist}(v, r); \quad (4.5)$$

The triangle quality illustrated in equation 6 states that the minimum distance from one point to another forms a straight path [76]. Working with attributes of data which is continuous in nature involves the Z score standardization and normalization of lowest-largest value [76]. The KNN disadvantages are routed from basing on the value selection which may not be accurate sometimes hence leading to low efficiency [77].

4.2.2 Support Vector Machine (SVM)

SVM is a supervised classification and regression model in which various classes are represented in hyperplane in a model space of multiple dimensions. The multidimensional 1 space is employed to widen the class margin difference [74]. To reduce the error, SVM generates the hyperplane iteratively. SVM obtains multiple classes by diving datasets hence a largest margin can be obtained in the sample space.

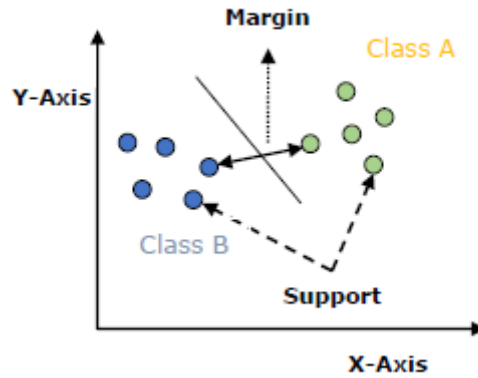


Figure 9: SVM Plane

In this technique, the goal is to map the sample space to the feature space of the highest possible dimension, also termed the Hilbert space. The nonlinear techniques are used to attain the Hilbert space. Dividing the plane formed by the original sample is challenging because its nonlinear. To overcome this drawback, transformation is done from nonlinear to linear plane which is easily separable [78].

To illustrate the above scenario, assume a set $S = \{(x_1, y_1), \dots, (x_M, y_M)\}$ of training data, where a sample $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$ with sample label, y_i , and M being the

sample numbers, and m being the attribute numbers. The description of the partition hyperplane is illustrated as shown with w being the normal vector and the displacement being denoted as b

$$f(x) = w^S x + b, \text{ where } w = (w_1, w_2, \dots, w_m)S \quad (4.6)$$

4.2.3 Logistic Regression (LR)

LR is a prediction method categorized under supervised learning algorithms that used classification [79]. LR algorithm stipulates the relationships that define the predictor variables. The predictor variables can be represented as $y' = (y_1, y_2, \dots, y_l)$, in addition to the response variable. In the case of this work, “normal flow” and “abnormal” variables which are categorical are the responses being predicted.

The existence of water theft which is depicted as abnormal flow has a conditional probability which can be defined as $L(P = 1|x) = \pi(x)$. LR model for l predictor variables can be written as;

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_l y_l)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_l y_l)}} \quad (4.7)$$

$$\text{Where } 0 \leq \pi(y) \leq 1$$

The LR can be transformed into what is referred to as the logit transformation which is described as follows.

$$g(y) = \ln \left(\frac{\pi(y)}{1 - \pi(y)} \right) = \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \dots + \beta_l y_l \quad (4.8)$$

A single unit change in y_j is defined by the odds ratio illustrated as $e^{(\beta_j)}$

4.3 Project Description

This section gives a brief description of the work that was done in the study.

The prototype of piped water theft detection system was developed in hardware using two Arduino Uno defined slave and Master. These were programmed in Arduino and the software was uploaded to the memory. The Master was interfaced with one flow rate sensor, Global System for Mobile Communication (GSM) modem, buzzer for alert, and the Liquid Crystal Display (LCD). The slave microcontroller was interfaced with single flow rate sensor. These hardware components have been described in the proceeding section. Water pipes of about 20mm diameter were tightly connected to the flow rate sensors in a branched mode and the other end of the pipe was tightly connected on the

tap of an overhead tank. Using a stop clock, water from the overhead tank was released into the distribution network (experimental system) and the state of the network (Normal flow or abnormal flow) recorded in intervals of 10 seconds. Flow rate values and volume of water at every instant were read from the LCD and recorded. The experiment was run for about 3 hours which generated 1017 data entries used to create the water theft detection dataset. Some variables such as normal flow value and absolute difference values of normal flows and the flow rate every instant were derived. This data was preprocessed, visualized and some feature engineering was also done. The resulting dataset which was free of any missing values was used to train Random Forest classifier in which 80% of the total data was utilized in training and 20% for testing the trained model. Then accuracy, precision, recall and F-measure were used to as the basis for evaluating the model performance. To appreciate the results of the Model, an evaluation was done on LR, KNN and SVM for benchmarking purposes.

4.4 System Design for data collection

4.4.1 Block Diagram

Data collection was done using a system prototype for piped water theft detection designed with Arduino Uno interfaced with two flow rate sensors (figure 2). The basis for theft detection is the ability to adopt a fixed value of flowrate designated as the nominal value within a short time delay of 5 seconds. The basis for incorporating this kind of operation is the fact that pressure zones in the default piped water distributions systems have a fixed range of flow rate values which are maintained. Flow rate variations could occur but within a s define range of values [80]. Any unauthorized activities along the piping network such as illegal connections could lead to a wide variation of flowrate values.

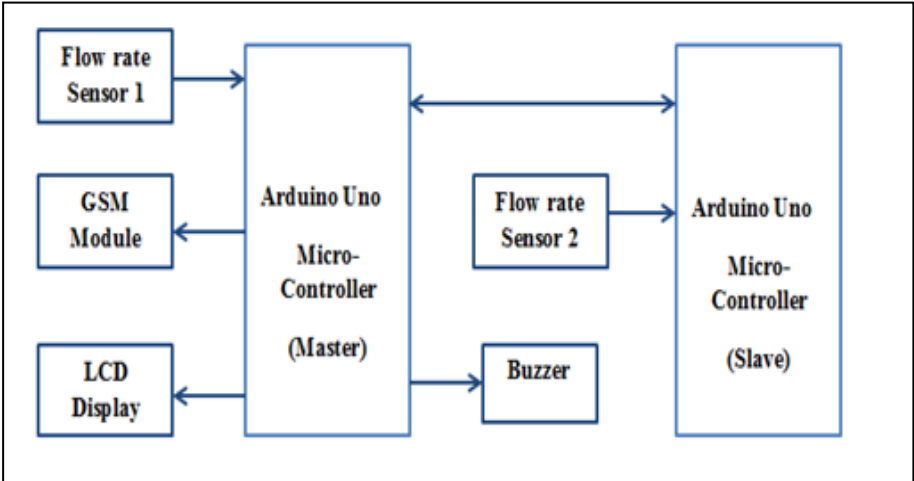


Figure 10: Block diagram

4.4.1 Flow chart

The flow of events followed in operation of the prototype for data collection is as shown.

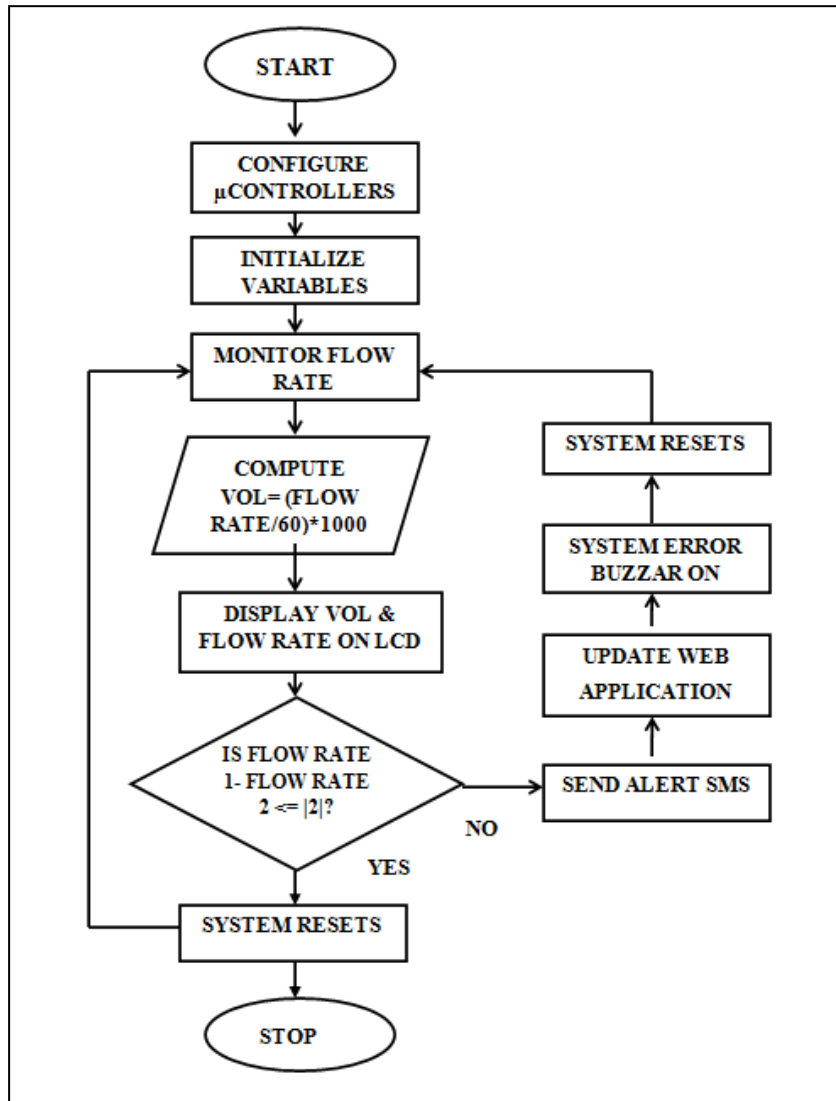


Figure 11: Flow chart

4.5 Components Description

i) Arduino Uno

Arduino Uno microcontroller falls under the microcontroller family that uses ATmega328 datasheet on which its operation is built upon. Arduino Uno has a total of 14 digital pins categorized as input and output. The total number of analog pins is 6, with a ceramic resonator of frequency 16 Megahertz. Other components include; the ICSP header reset button, USB connector and Power jack for powering the microcontroller with 5volts of direct current (DC). This power can be generated from Computer USB slots or with a batter source. The powering can be done when the microcontroller is on

its board (Arduino board). A power adaptor that converts AC to about 5V of DC can also be used to power the board. Arduino Uno has a FTDI USB serial chip which distinguishes it from other earlier boards [78].

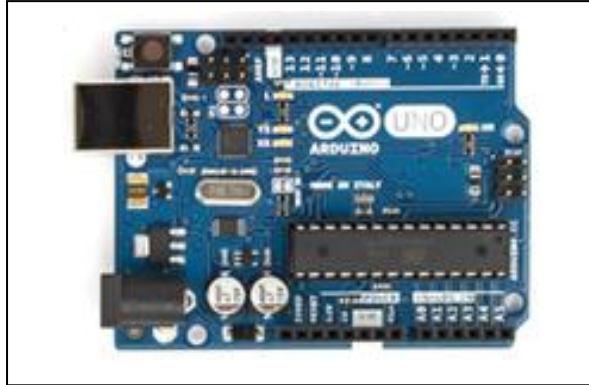


Figure 12: Arduino Uno Board

ii) Flowrate Sensor

Flow rate sensor consists of a rotor which is shaped in form of a propeller. Its operation is based on the Hall Effect principle which explains the induction process of a conductor that in relation to electric current and the perpendicular field of a magnetic force [81]. For a conductor that is in transverse to an electric current and a perpendicular field of magnetic force, it experiences an induction of a voltage difference. The rotor in the sensor uses this principle to measure the flowrate of water passing through it which causes the rotor to rotate by pushing the rotor fins. As the rotor rotates, a voltage is induced which leads to a sensor output of 4.5 pulses. This occurs in every instance when a volume of 1000cm^3 of the fluid passes through the sensor in every minute. The voltage induction is brought about by the magnet which is connected to the shaft of the rotor bringing about the varying field of the magnet [82]. Interfacing the flow rate sensor to the Arduino Uno is as follows. The data cable which is marked yellow is connects to interrupt pins of the microcontrollers (D2 & D3).



Figure 13: Flowrate sensor/Flow meter

iii) GSM Modem

The SIM900 variant of the GSM modem was used in the study. This category of modem is a quad band type of GSM modem. Signals are sent from the system to the phone as a sms using UART which is the main communication channel. This makes use of the TXD and RXD for UART pins. The Modem is normally powered by 12volts though it also bears the 5volts line used for Arduino power. For controlling the communication, the board uses certain commands called the attention (AT) commands.



Figure 14: GSM Modem

4.6 Experimentation for Data Collection

A prototype of the system for detecting water theft was designed using proteus software. The software program that drives the Arduino was written and uploaded to the memory of the two microcontrollers. Only one flowrate sensor was supported by an individual controller due to the issue of timers and clock which justifies why two controllers were used in the study. The main controller of all other circuit components is the master controller, which also gives instructions to the slave.

To send an alert sms to the administrator's phone the GSM modem was used. The sms contains the water consumption details, that is; volume and flow rate values and the location of the specific system deployment.

The system was tested to ensure it meets the functional requirements. Pipes a diameter of about 20mm were fitted to the sensors and arranged as shown in figure 15. The other end of the pipe was connected to the tap an overhead tank and water was released into the piping network. The flow rate sensors were laid as shown.

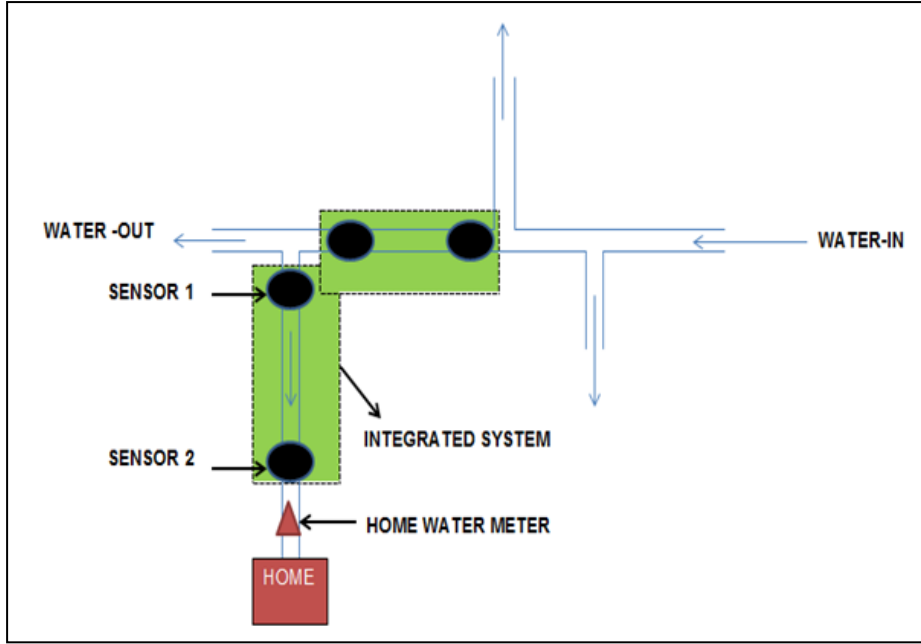


Figure 15: System Prototype positioning along the water pipeline

4.6.1 Data Collection

The values of flowrate of the water passing through the system were recorded as read from the LCD in the intervals of 10 seconds. A suction pump was connected to the piping system and used to vary the flow in the upward dimension. However, to decrease, the flowrate, a diversion pipe was connected to the system just before the sensors. The state of the network was determined based on the SMS alerts received on the phone that stated any abnormalities in water flow and for the case of normal flow, no sms was received. The experimentation generated a total of 1017 entries and the derived variables such as volume flow and the difference of the flowrate and adopted norm value in absolute figures were obtained and incorporated in the dataset.

$$A_d = |f - f_n| \quad (4.9)$$

$$V_{inst} = f * t \quad (4.10)$$

Where A_d is the absolute difference, f is the flowrate at that instant, f_n is the adopted norm value and V_{inst} is the instant volume flow. The final dataset consists of 6 variables.

4.6.2 Data Analysis and Processing

The recorded data had some missing values whose tuples were discarded. The derived variables were rounded off to the nearest decimal value to ensure uniformity in the dataset.

The values recorded indicated a certain pattern of water flow after a period of roughly 2 minutes. This implies that the distribution system adopts a certain range of flowrate and stability in the flow. The system was configured to reset whenever abnormality flow was reported which resulted into a fresh count and adjustment in the flow. A normal flowrate value would again be attained after a short delay.

The variations in water flow were used as an indicative procedure for deducing the existence of water theft. A higher deviation from the normal flowrate value intimated the likeliness of water theft existence. A certain pattern that showed a fluctuation of flow within certain limits was observed after visualizing the adopted flowrate values designated as the norm flowrate values, as shown in figure 16.

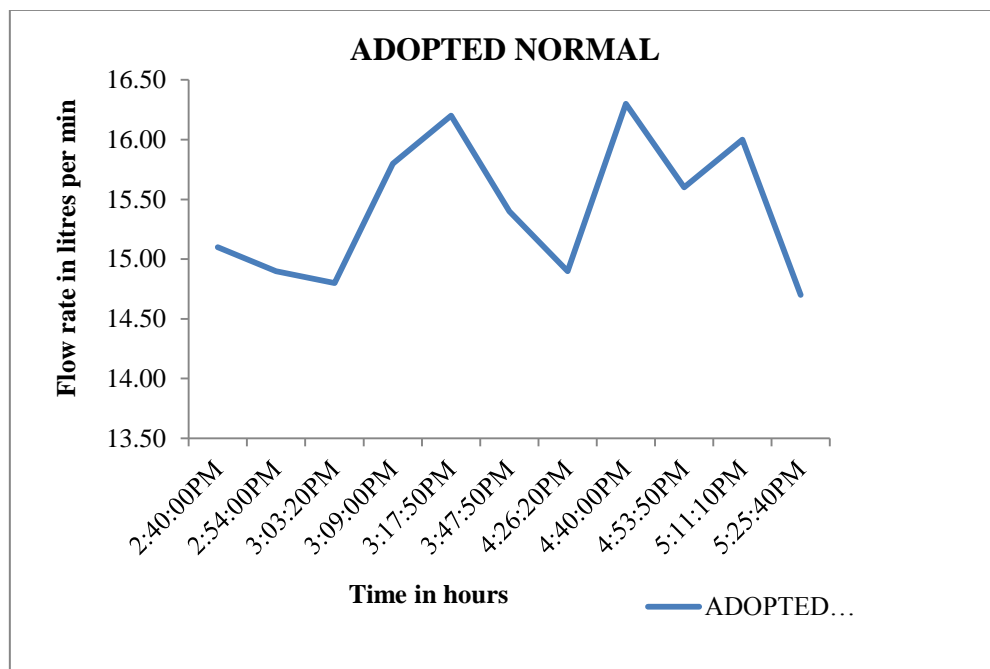


Figure 16: Water flow pattern in experimental system

The basis for detecting water theft in most of the works discussed in chapter three is the concept of water distribution which explains that for a given pressure zone, the water distribution pressure is maintained within certain limits. This is something that is hard to maintain by distribution entities strive to stabilize a given range of pressure to ensure uninterrupted supply hence efficiency in the distribution systems [29]. However, the

experimental pattern satisfies this concept. These fluctuations within certain ranges are not guaranteed in real life applications due to various reasons such as low water levels in the reservoirs, pipe bursts, leaks, blockage in the pipes, inefficiency in pumps and gate valves among other reasons.

The distribution pattern for the experiment as illustrated in the line graph shown in figure 16 indicates the minimum value of flowrate being 14.7 litres per minute and the maximum value being 16.3 litres per minute. The entire fluctuation was with the range of 1.6litres per minute.

CHAPTER 5: RESULTS AND DISCUSSION

5.0 Chapter Summary

The results of the experiment have been discussed in this chapter. Some introductory theory on the various steps taken to obtain the results has also been elaborated. The chapter is concluded with discussion of the results obtained and their implication.

5.1 Modeling and Prediction

In this section, a summary of the various steps taken to train and test the prediction models has been presented.

The libraries that were used in the experiment included pandas, numpy, matplotlib and seaborn. These were imported in Anaconda with Jupyter Notebook, the integrated IDE working with Python 2 after the relevant dependencies were installed.

The dataset was converted into a csv file and uploaded, the read with pandas. Exploring and visualizing data is a vital step in data analysis. This was done to get acquainted of the patterns that exist in the data [83]. The next step was feature engineering which was done to improve the classification accuracy. Feature extraction requires expert knowledge to mine the different features from data using various techniques. This was performed on the water theft dataset. The dataset was partitioned into training and testing data in ratio of 4:1 respectively. The Random Forest Algorithm was used to train the classification model used for prediction of existence of water theft. The other Benchmark methods that are KNN, LR and SVM were also used to construct the respective predictive models. Model testing was done on the 20% data and their performance was evaluated based on the confusion matrix used to compute the model accuracy, precision, recall and F-score.

5.2 Model Evaluation

The confusion matrix was generated and used to compute the various performance measures. A confusion matrix is an array of values that indicate the correctness in classification in relationship to the total number of values used in the testing sample.

Table 1: Confusion Matrix

Confusion Matrix		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

The matrix consists of four partitions, which are TP (true positive), TN (true negative), FP (false positive) and FN (false negative) as depicted in table

The samples that are truly positive and classified accurately are described by the true positive value of the confusion matrix. The samples that are truly negative but wrongly classified in the positive class are referred to as the false positives. On the other hand, there are samples that are actually positive but predicted wrongly as belonging to the negative class. Such samples are the false negative samples. And lastly, the negative samples that are indeed negative and further classified as being negative are what we call, the true positive samples [84].

The samples that are correctly classified form the correctly classified class which consists of the TP and TN, whereas the entire sample size class consists of TP, FP, FN, and TN. The calculation of the various performance measures only differ in terms of their mathematical computation using the four values. Therefore, accuracy which shows the ability of the classification model to correctly classify the positive the samples is ratio of TP and TN class to that of the entire sample, that is TP, FP, FN, and TN class. Precision describes the correctness in classification of all positive samples as compared to the entire class sample of positive samples and recall on the other hand, defines the classification correctness of positive samples in relation to the accurate number of positive results expected to have been output, expressed as a fraction. Finally, F-Score is the harmonic mean score of precision and recall achieved by the model [84].

The performance measures can be mathematically represented as follows.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.1)$$

$$Precision = \frac{TP}{TP+FP} \quad (5.2)$$

$$Recall = \frac{TP}{TP+FN} \quad (5.3)$$

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (5.4)$$

5.3 Discussion

The model accuracy of Random Forest was considerably high at 97% which affirms its potential in being used for predicting water theft along water distribution pipelines. The competing benchmark model of KNN also attained a similar accuracy value unlike the remaining models whose accuracies were slightly lower. From the bar graph in figure.....used to visualize the performance measures and how they vary with various models, it can be observed that the highest values were those of precision of SVM and LR models which was at 98% and equal to the Recall attained by the RF model,

Table 2: Performance Measure

Method	Accuracy	Precision	Recall	F-Score
RF	0.97	0.95	0.98	0.96
SVM	0.96	0.98	0.93	0.95
LR	0.96	0.98	0.93	0.95
KNN	0.97	0.95	0.96	0.95

The values of performance measures of the models have been visualized in bar graph.

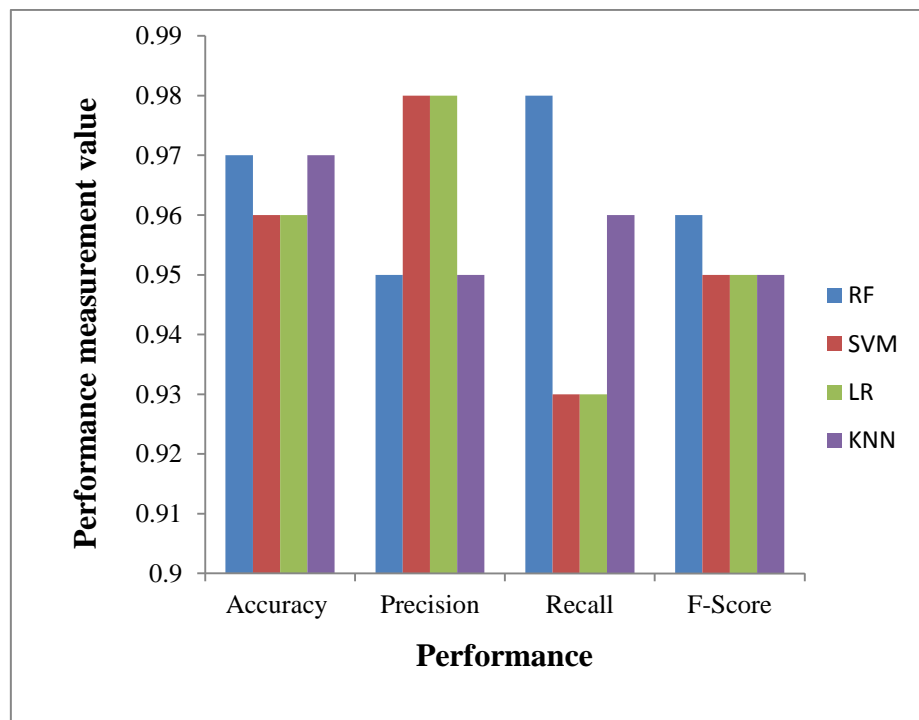


Figure 17: Performance of prediction models

The ability of KNN to achieve the same accuracy as that of RF mode can be associated to using the same training and testing sizes of data. In training the RF model, 20 estimators were used and the increase of the number of estimators did not impact on the accuracy value obtained.

In [85], a discussion of how the number of trees affects the variable importance has been done. The result of variable importance is more stable with increased number of trees (estimators). However, the stability increase is not exponential. Using unnecessarily large number of trees may be time wasting but the accuracy value will not be affected [86]. In our experiment, the same value of accuracy achieved with 20 estimators was the same as that achieved with 40 estimators. However, a slight drop in the accuracy value was observed when the number of estimators was further increased to 50. A further discussion with [87] intimates that with n_{tree} being 200, the Random Forest can attain accurate results though this greatly varies with the number of variables used and the total number of training samples.

The number of data samples used were about 800 only which led to an accuracy of 97%. This implies that with more data samples and the accuracy of the model could even improve. Due to variations in atmospheric pressure exerted on the water, some flowrate values attained were totally in line with the flow pattern earlier established in figure 16. It implies that more training data is required to train the model.

The data used in the study reveals a certain pattern which aligns with ideal water distribution systems whose distribution pressure is maintained within a range of pressures. The fluctuations are confined within a certain range of flowrate (1.6 litres per minute) though this pattern is not attainable in most real world distribution systems due to various reasons such as low water levels in the reservoir hence exerting less atmospheric pressure to drive the water to the extreme end of the pipeline, blockage in the channels, pipe bursts and leaks, inefficiency of driver pumps and pressure boosters, and ineffectiveness of gate valves which often require manual control among others. Therefore, an intelligent prediction model like this is the fitting solution for intelligent prediction of water theft cases in water pipelines.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.0 Chapter Summary

In this chapter, a summary of the work done has been given in the conclusion section and lastly, the possible areas for future work have been highlighted in the last section.

6.1 Conclusion

Water theft is still a challenge to water distribution companies across the globe. It does not only deprive the company of her revenue but also leads to compromise of water quality, insufficient water supply to other consumers, pipe bursts and leaks among other problems. The research work so far done builds the detection mechanism using hardware. The software programs fed in the controller memory computes the variation in water flow using the flowrate data collected by flowrate sensors from the piping network. This approach has various drawbacks as the water theft problem gets more sophisticated. The software program which is rule based may not be able to make accurate conclusions in case of erroneous data and other forms of theft such as meter tampering and reversals that do not directly alter water flow. In this study, hardware prototype, which is a network of flowrate sensors interfaced with Arduino has been designed and used to collect data used to model a random forest classifier for prediction of water theft at the server end. The total number of variables summed to 5 after cleaning the data.

The variables used in model training are; flowrate, volume consumption and status of flow, the acquired flow rate value designated as normal flow, absolute difference between acquired normal flow and the flow at the instant. The missing values were deleted resulting in a total of 1017 entries. Prior to model training and testing, the data was divided in the ratio of 4:1 respectively for the purpose of constructing the classifier.

The model performance was based on accuracy, precision, recall and F-score and compared with other three competing methods, that is; KNN, SVM and LR. The four trained models achieved considerably high accuracies which affirm their ability to predict water theft more intelligently. The accuracy of the existing system has not been documented; therefore we cannot be able to compare our model performance with the performance of the hardware systems that have been designed. Random forest is

effective in estimating the value of the features and test error without necessitating repeated training of the model. This work formulates a new direction in research on water theft detection being the first initiative to introduce machine learning techniques in predicting water theft. The model can be integrated with the IoT based systems of water theft detection to predict water theft more accurately with minimal errors. To apply this approach in the field, data on water flowrate and volume consumption can be collected over a reasonable period of time with various controls such as ensuring that the piping system is in good condition, ensuring that no leaks, illegal connections or any form of interruption happens to distribution system over the period of data collection to ensure accurate values are obtained. This data can be used to train the model, in this case, unsupervised learning model be used since the response variable is not known.

It is recommended that more accurate flowrate sensors be used in data collection due to various drawbacks in the existing sensors which include; limited flowrate range that can be read and high calibration errors.

6.2 Future Work

The defined pattern observed in the collected data made is easier to train the classification model. The high performance measure attained by the models reaffirms a promising direction of water theft prediction research using Machine Learning. The next direction of this work involves incorporating the model in a complete system with remote database and web application.

APPENDICES

Appendix 1: System Prototype for Data Collection

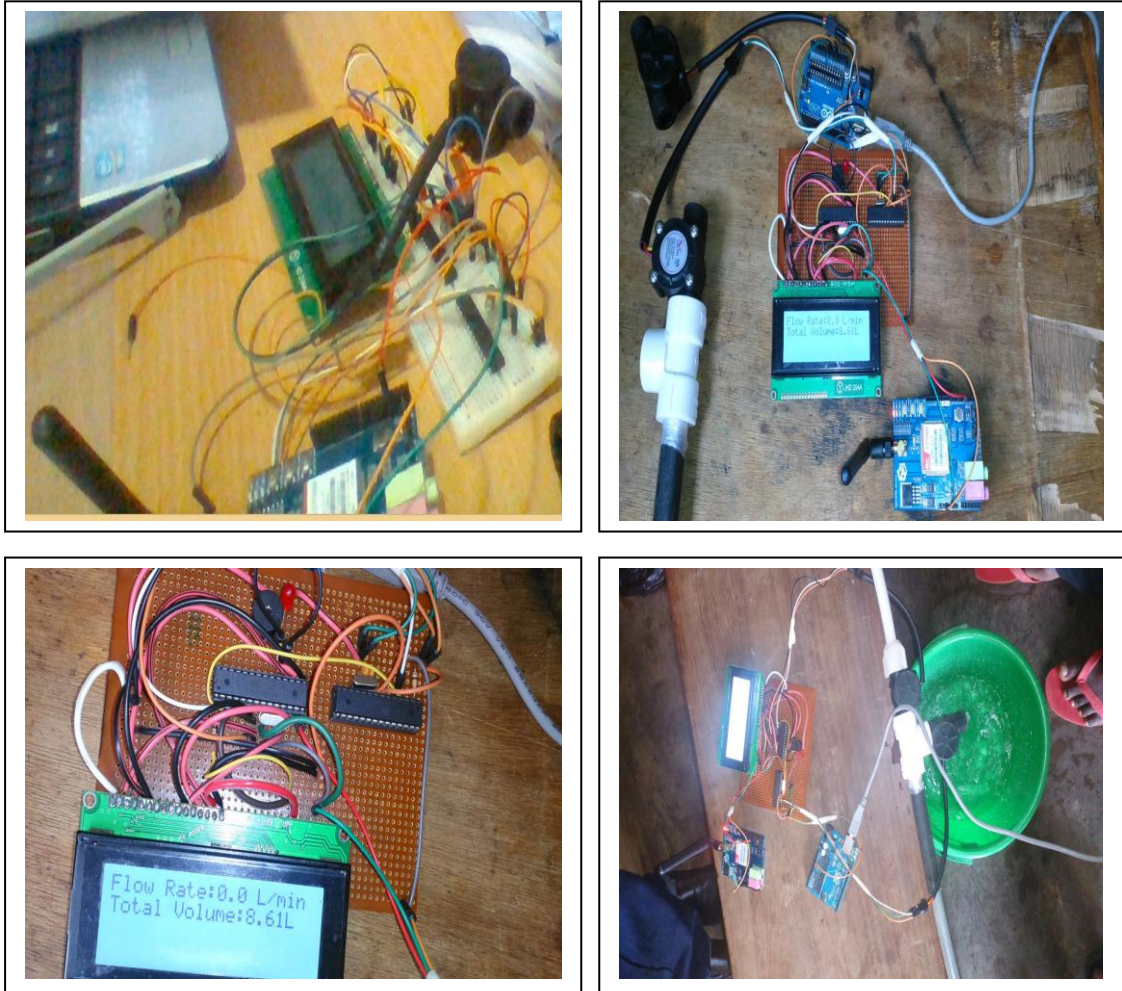


Figure 18: System Prototype for data collection

REFERENCES

- [1] The Wikipedia Guide, *Introduction to Machine Learning*, Chapter 1.
- [2] A. Smola and S.V.N. Vishwanathan, *Introduction to Machine Learning*, pp. 3-11, Cambridge University press, 2008
- [3] T.A Oladipupo, *Types of Machine Learning Algorithms*, pp. 20-47, Research gate, 2010
- [4] S. Banerjee, A.Y. K. Chua and J. J Kim, *Using Supervised Learning to Classify Authentic and Fake Online Reviews*, pp. 1-7 IMCOM '15, January 08 - 10 2015, BALI, Indonesia
- [5] S. Benzel and A. Stanescu, *Histogram Methods for Unsupervised Clustering*, ACM Southeast Conference, ACMSE, pp. 248-251, Tampa, FL, USA, April 2-4, 2020
- [6] Z. J Leibo, P. Julien, E. Hughes, S. Wheelwright, A. H. Marblestone, E. DuéñezGuzmán, P. Sunehag, D. Iain, and T. Graepel, *Malthusian Reinforcement Learning. In Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, Montreal, Canada, May 13–17, 2019, IFAAMAS, 9 pages.
- [7] Liaw and Andy, *Documentation for R package random forest (PDF)*, Retrieved 15 March, 2013.
- [8] Leo Breiman, *Random Forests - Machine Learning*, 45 (1): 5–32, 2001
- [9] Y. Amit and G. Donald, *Shape quantization and recognition with randomized trees (PDF)*, Neural Computation 9 (7): 1545–1588, 1997
- [10] T. Kam Ho, *Random Decision Forest*, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282
- [11] T. Kam Ho, *the Random Subspace Method for Constructing Decision Forests*, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8): 832–844, 1998
- [12] The United Nations World Water development report (UNWWD), *water for sustainable world*, UNESCO, United Nations Chapter 4, 2015
- [13] UNDESA (United Nations Department of Economic and Social Affairs), *International Decade for Action Water for life*, United Nations, New York, 2005-2015.

- [14] Naik K Pradeep, *Water crisis in Africa: myth or reality?* International Journal of Water Resources Development, ISSN: 0790-0627 (Print) 1360-0648, 2016 (Online) Journal homepage: <http://www.tandfonline.com/loi/cijw20>
- [15] WHO (World Health Organization), *World Water Day Report*, 2015. Retrieved from http://www.who.int/water_sanitation_health/takingcharge.html
- [16] M. Clifton, L. Fengting, N. Innocent, W. Gumindoga and G. Takawira, *Health Safety of Drinking Water Supplied in Africa: A Closer Look Using Applicable Water Quality Standards as a Measure*, Springer, 2017
- [17] Review of Sector Reforms and Investments, *Access to Water and Sanitation in Sub-Saharan Africa, Key Findings to Inform Future Support to Sector Development*, Synthesis Report, GIZ Competence Center Water, Wastewater, Solid Waste, Eschborn, January 2019
- [18] B. V Felbab, *Brookings Mountain West Lecture Series*, University of Nevada, Las Vegas, 20 February 2015, [Accessed 2020].
- [19] Ministry of water and Environment, "Uganda Water and Environment Sector Performance Report," Government of Uganda, Kampala, 2018.
- [20] National Water and Sewerage Corporation, *National Water and Sewerage Corporation*, NWSC, 23 06 2013, Available: <http://www.nwsc.com> [Accessed 2020].
- [21] G. Puranik and A. Gaikwad, *Automated urban water supply system and theft identification*, International Journal of Electronics and Communication Engineering & Technology (IJECET), vol. 6, no. 6, pp. 145-156, 2015.
- [22] A. Lavanya. J. Tharanyaa and A. Jagadeesan, *theft identification and Automated Water Supply System Using Embedded Technology*, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue. 8, pp. 3727-3734, 2013.
- [23] P. Ittannavar and D. Madihalli, *Smart Water Supply Management*, International Journal of Emerging Trends in Electrical and Electronics (IJETEE – ISSN: 2320-9569), Vol. 10, Issue. 9, pp. 77-79, 2014.
- [24] Lewis A. Rossman, *EPANET2 User Manual*, United States Environmental Protection Agency (EPA), Water Supply and Water Resources Division National Risk Management Research Laboratory Cincinnati, OH 45268, Chapter 1, 2000
- [25] National Water and Sewerage Corporation, Littlegate publishing, 01 04 2015. [Online]. Available: <http://www.littlegatepublishing.com>. [Accessed 2020].

- [26] A. Kyotalengerire, *Challenges facing National Water and Sewerage Corporation*, New Vision paper, Kampala, September 15, 2015.
- [27] G. M Tamilselvan and V. Ashishkumar, *IOT Based Automated water distribution system with water theft control and water purchasing system*, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, vol. 7, no. 4, pp. 151-156, 2018.
- [28] R. Malhotra, *An empirical framework for defect prediction using Machine Learning techniques with Android software*, Applied Soft Computing, Elsevier, vol. 49, pp. 1034-1050, 2016
- [29] Dictionary.com, *Dictionary.com*, 2020.
- [30] S. Ittannavar and B. Madihalli, *Smart Water Supply Management*, International Journal of Emerging Trends in Electrical and Electronics (IJETEE – ISSN: 2320-9569), vol. Vol. 10, Issue. 9, pp. 77-79, October, 2014.
- [31] Shahrour, *Sustainable and Smart City*, Lille University, February 2014. [Online]. Available: <http://www.lillel.fr.com>. [Accessed 2020].
- [32] *Introduction to flow measurement*, Omega, 2014. [Online], [Accessed 2019].
- [33] A.S Benigno, *Design Manual, Water Partnership Program, Rural water Supply, Vol. 1*, Manila, Philipines, February, 2012.
- [34] United Nations Human Settlements Programme (UN Habitat), *Reduction of Illegal Water*, Nairobi, 2012.
- [35] Board of Water and Sewer Charge Review (BWSCR), *Water Service Rules and Regulations*, City of Dayton
- [36] H. A. Gaikwad and P. V. G. Puranik, *Automated urban water supply system and theft identification*, International Journal of Electronics and Communication Engineering & Technology (IJECET), pp. Volume 6, Issue 6, June 2015.
- [37] *Flow meters*, Omega, 2014, [Accessed 2020].
- [38] US Environmental Protection Agency, *Control and Mitigation of Drinking Water Losses in Distribution Systems*, 10-019, November, 2010.
- [39] Z. Satterfield and B. Vipin, *Tech Brief, Water Meters*, vol. Vol. 4, no. Issue. 2, Summer 2014.
- [40] Z. Satterfield and B. Vipin, *Water Meters*, National Environmental Services Center, 2014

- [41] A. M Maher and T. Nemanja, *Impacts of supply duration on the design and performance of intermittent water distribution systems in the West Bank*, Water International, 38:3, pp. 263-282, 2013
- [42] Washington State Department of Health, *Environmental Public Health Office of Drinking Water*, June, 2014. DOH 331-338.
- [43] The Molitor, *Underspace*, 22 October 2012, Available: <http://underspace.com>. [Accessed: 2020].
- [44] A. Mancharkar, R. Kulthe and I. Shewale, *Automated water distribution system for smart city using PLC and SCADA*, International Journal of Emerging Technologies and Engineering (IJETE) Vol 3 Issue 3, pp 11-14, 2016
- [45] A. Panchal, K. Dagade, S. Tamhane, K. Pawar and P. Ghadge, *Automated water supply system and water theft Identification using PLC and SCADA*, International journal of Engineering Research and applications, Vol 4 Issue 4 (version 6), pp. 67-69, 2014.
- [46] R. Ankita, A. Gaikwad, D. Rehu, R. Ashutosh Raichurkar, R. Jadhar, *Automated water distribution system and theft detection*, International Journal of Advance Engineering and Research Development, Vol 4 Issue 4, 2017
- [47] G.M Tamilselvan, V. Ashish Kumar, S. Jothi Prasath, S. Mohammed Yusuf, *IoT based Automated water distribution system with water theft control and water purchasing system*, International Journal of Recent Technology and Engineering (IJRTE), Vol 7, Issue 4s, 2018.
- [48] K. N Lakshmi and S. Sankaranarayanan, *IoT enabled Smart Water distribution and underground pipe Health monitoring Architecture for smart cities*, 5th International Conference for Convergence in Technology (I2CT) 2019 IEEE Pune, India, pp. 1-7, 2019.
- [49] R. Namrata, W. Rohan, G. Dinesh, T. Amol, *Automatic water distribution and leakage detection using PLC and SCADA*, International Research Journal of Engineering and Technology (IRJET), Vol 4, Issue 02, pp. 1-3, 2017.
- [50] S. Sundaresan and M. Nivetha, *Automated drinking water distribution using Arduino*, SSRG International Journal of Civil Engineering (SSRG-IJCE), Vol 5, Issue 5, pp. 66-69, 2017.
- [51] S. Gopalakrishnan, V. Hemalatha, *An embedded based monitoring and distribution system for water supply in urban areas*, International journal of Engineering Science and Computing, Vol 7, Issue 5, pp 11326-11328, 2017.

- [52] P. Prashant, P. Shrinivas, B. Pooja, C. Ashish, *Automation in drinking water supply distributed system and testing of water*, IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), pp. 36-38.
- [53] R. Gowtham, M. C Varunkumar, P. M Tulsiran, *Automation in urban drinking water filtration, water supply control, water theft Identification using PLC and SCADA and self-power generation in supply control system*, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Vol 3, Issue 7, 2014, pp. 698-703.
- [54] J. Gouthaman, R. Bharathwajanprabhu and A. Srikanth, *Automated urban drinking water supply control and water theft Identification system*, proceeding of the 2011 IEEE students' technology symposium, IIT Kharagpur, pp. 87-91, 2011
- [55] H. Adityaraj, B. Patel, S. Sheth, *Automatic water distribution system using Arduino Uno*, International Journal of Engineering and development (IJERD), Recent trends in Electrical, Electronics and Communication Engineering, pp. 20-24, 2016.
- [56] G. Chaitanya, M. J. Tabassum, S. Akila, *Smart urban water supply scheme with water theft detection system*, International Journal of Pure and Applied Mathematics, Vol 119, Issue 15, pp. 951-958.
- [57] S.R. Kinge, N. Nibhona, P. Singh, R. Kumar, *Automatic water distribution*, Journal of Emerging technology and Innovative Research (JETIR), Vol 4, Issue 05, pp. 31-35, 2017.
- [58] V. G Kumar and M. Soundhrya, *Automation in water board using PLC with SCADA*, International Journal of Advanced Research Biology Engineering Science and Technology (IJARBEST).
- [59] P. D Juliana and A. Tavares Da Silva, *Fraud detection in water meters using pattern recognition techniques*, ACM, pp. 77-82, 2017.
- [60] M.A Chuadhry, R. Justin, *Model based attack detection scheme for smart water distribution Networks*, ACM, pp. 101-113, ACM 2017
- [61] A. Lavaya, P. J Shri Tharanyaa, and A. Jagadeesan, , *theft Identification technology and automated water supply system using embedded system*, Journal of Electrical, Electronics and Instrumentation Engineering, Vol 2, Issue 8, pp. 3727-3733, 2013.
- [62] Z. Nyein, *Japan International Cooperation Agency (JICA) to help Yangon City Development Committee (YCDC) Battle Water theft*, Vol. 3, Issue 7," 2015, [Accessed 2020].

- [63] Elster Metering, 2015. [Online]. Available: <http://elstermetering.com>.
- [64] Zhi-Hua Zhou, *Ensemble Methods*, Foundations and Algorithms, pp. 56-58, Taylor & Francis Group, LLC, 2012
- [65] G. Louppe, *Understanding Random Forests from theory to practice*, PhD dissertation Chapter 4, University of Liège, arXiv preprint arXiv: 1407.7502, July 2014.
- [66] D. G Thomas, *Ensemble Methods in Machine Learning*, pp. 1-15, Oregon State University, Corvallis, Oregon, USA
- [67] Re Matteo and V. Giorgio, *Ensemble methods: A review*, CRC Press LLC, 2014
- [68] L. Breiman,, *Random Forests*, Machine Learning 45 (1): 5–32, 2001
- [69] A. Boulesteix, C. Strobl and T. Augustin, *Unbiased split selection for classification trees based on the Gini index*, Computational Statistics & Data Analysis: 483–501, 2007
- [70] G. Runger, H. Deng, and E. Tuv, *Bias of importance measures for multi-valued attributes and solutions*, Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), pp. 293–300, 2011
- [71] L. Tolosi, A. Altmann, O. Sander, T. Lengauer, *Permutation importance: A corrected feature importance measure*, Bioinformatics, 2010
- [72] T. Lengauer and L. Tolosi, *Classification with correlated features: unreliability of feature ranking and solutions*, Bioinformatics, 2011
- [73] Y. Lin and J. Yongho, *Random forests and adaptive nearest neighbors (Technical report)*, Technical Report No. 1055, University of Wisconsin, 2002
- [74] D. M. Harikrishna, *Children’s Story Classification in Indian Languages Using Linguistic and Keyword-based Features*, ACM Trans, Asian Low Resource. Lang. Inf. Process, Vol. 19, Issue. 2, p. 22 pages, 2019.
- [75] Vandana and N. B, *Survey on nearest neighbor techniques*, International Journal of Computer Science and Information Security (IJCSIS), Vol. 80, Issue 2, 2010.
- [76] P. Chandra, *Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm*, in International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), Delhi, 2013.
- [77] *Arduino Datasheets*, Ivrea: Italy, <https://www.farnell.com/datasheets/1682209.pdf>.
- [78] Shi. H, A, *Novel Support Vector Machine Algorithm for Missing Data*, ACM, 2018
- [79] S. Briggs, *Decision trees for predicting risk of mortality using routinely collected data*, International Journal of Social and Humanistic Computing, vol. 6, no. 6, p. 303–306, 2012.

- [80] M. J. P. C. University, *Performance assessment System, Basics of Water Supply System, Training Module for Local Water and Sanitation Management*, Maharashtra Jeevan Pradhikaran (MJP) CEPT University, Maharashtra, 2012.
- [81] Arduino, *How to interface an Arduino Uno with a Flow rate sensor to measure a liquid*, Arduino, 23 03 2018. [Online] Available: <https://maker.pro/arduino/tutorial/how-to-interface-arduino-with-flow-rate-sensor-to-measure-liquid>. [Accessed 2020].
- [82] *Working with water flow sensors and Arduino*, 05 08 2019. [Online].Available: <https://www.electroschematics.com/12145/working-with-water-flow-sensors-arduino>. [Accessed 2019].
- [83] A. Casari, *Feature Engineering for Machine Learning Principles and Techniques for data scientists*, Sebastopol, California: O'Reilly Media, April, 2018 First Edition.
- [84] S. Yang, *A patient outcome prediction based on Random Forest*, ACM, 2019.
- [85] M. Wiener and A. Liaw, *Classification and regression by random forest*, R News pp. 18–22, 2002
- [86] P. N Thanh and K. Martin, *Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery*, Sensors, Vol 18, Issue 18, 2018.
- [87] J. Liu, Q. Feng, and J. Gong, *UAV remote sensing for urban vegetation mapping using random forest and texture analysis*, Remote Sens.7, 1074–1094, 2015

LIST OF PUBLICATIONS

1. Simon Peter Khabusi and Rajni Jindal, *Pressure dependent piped water theft detection with remote billing and location alert*, IEEE/IIAI International Congress on Applied Information Technology (IEEE/IIAI AIT 2019), Yogyakarta, Indonesia, 4th -6th November, 2019
2. Simon Peter Khabusi and Rajni Jindal, *Modeling and predicting piped water theft using machine learning approach*, International Journal of Engineering Research and Technology (IJERT), Vol. 9 Issue 05, pp. 304- 311, May-2020