Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Modeling and Predicting Piped Water Theft using Machine Learning Approach

Simon Peter Khabusi
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India

Rajni Jindal
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India

***Abstract*— Water theft is a major challenge to water distribution companies in most communities around the world. This results into unbalanced water flow and damage to the water pipes. The resulting high level of nonrevenue water does not only lower income generation but also the quality of service. The recent advances in the field of Machine Learning have seen its latest application to a wide range of fields. This work proposes a Random Forest (RF) prediction Model to accurately detect water theft along the distribution network. Data for training and testing purposes was collected through experimentation over a 3 hour period in 10 seconds intervals using a system prototype of two Arduino microcontrollers programmed and interfaced with water flow rate sensors with the ability to adopt a normal flow rate value after a small time delay. Therefore flow rate fluctuations outside a range of ±2 litres/minute were taken as abnormal flow. The proposed RF model was evaluated on four statistical measures namely; accuracy, precision, recall, and F-measure, and compared with three competitive approaches that is, logistic regression (LR), Support vector machine (SVM) and K-nearest Neighbour (KNN). Experimental results show no significant differences observed in accuracy and F-measure among the four models, while the proposed RF gives a higher value of recall. Conclusively, the proposed classifier has advantages compared with the other approaches in terms of reliable feature importance estimate and efficiency in test error estimation without incurring the cost of repeated model training associated with cross-validation.**

Keywords—Water theft, flowrate, Arduino, machine learning, supervised learning

I. INTRODUCTION

Water scarcity is a major challenge around the globe [1]. The United Nations 2015 report on world water development highlighted limitation of clean and safe drinking water as a major issue across Europe and North America, Asia and the Pacific, the Arab region, Latin America and the Caribbean, and Africa [1]. The United Nations Department of Economic and Social Affairs [UNDESA] report published in 2013 states that water crisis is not only a natural but also a human-made phenomenon [6]. Physical scarcity has resulted from natural phenomenon, that is; climate change and global warming whereas people's inability to utilize adequate water sources has resulted from economic scarcity which is common among low income earners in developed countries and more prevalent in under developed communities, especially in Africa [2].

World Health Organization (WHO) report of 2015 [3] shows that only 50% or less of people in rural Africa have access to clean drinking water and improved sanitation in spite of the fact that this is an issue that concerns all aspects of the wellbeing of persons, including their health, agricultural activities, educational development, economic productivity, even peace and stability [2]. In most communities, water is not only scarce but also of low quality [4] due to rapid urbanization and inefficiency in Government systems to supply safe and clean piped water to all citizens [5].

Despite these challenges, African states are striving hard to extend piped water to all communities [5]. The sector reforms and Investments report published by Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) in 2015 showed a significant increase in piped water connections over the recent years in Africa. Among the countries surveyed, Burkina Faso had the highest piped water coverage and this was at 76%, followed by Kenya at 70%, Zambia at 66%, Tanzania at 59% and Uganda at 56% [5].

Therefore, due to financial inability to afford and limited piped water supply, water theft is prevalent in most communities in Uganda especially the urban centers [6]. Its usage in almost each and every aspect of life has made its distribution and supply of utmost importance to researchers. Homes, Hospitals, public offices, construction companies and Industrial centers need a steady constant supply of water to ensure sustainability of their activities. Hence any scarcity could lead to great losses of money and even lives. Large water consuming entities allocate a great percentage of their budget on water [7] and even still, the normal daily water distribution per service pipe could most likely not meet their water demands. These result in a number of illegal activities such as illegal connections, meter bypass, meter tampering, meter reversals, and vandalism to ensure that more water can be obtained at lower or no costs [8].

Some customers use suction pumps to increase water supply to their service lines there by rendering low/no supply to other clients and eventually destroying pipes due to extreme uncontrolled pressures. Such water volumes could under normal circumstances not be availed [9] considering the normal water supply procedure in a given pressure zone. Such attempts interfere with the distribution water pressure which results into poor water supply and leaks resulting from pipe bursts due to extreme pressures

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

exerted by the suction pipes [9][10][11]. Piped water theft is widely spread and it is perpetrated by all classes of people [12].

In Uganda for example, National Water and Sewerage Corporation (NWSC), a government parastatal body mandated with supplying clean and safe piped water to citizens incurs about 30% to 35% of its total revenue in water theft [13]. The initial attempts involving use of police to enforce the law against piped water theft, community policing and regular staff supervisions to reduce the Non-Revenue water (NRW) have not yielded much success. In response to this dire need, a lot of initiatives have been undertaken by different researchers to provide a mechanism for automated water supply, monitoring and timely reporting of water theft activities along the distribution network. The current state of research involves use of remote flow rate sensors laid along water distribution pipelines to capture the water flow rate values. Figure 1 shows the conceptual framework of the process of water theft detection in existing applications.

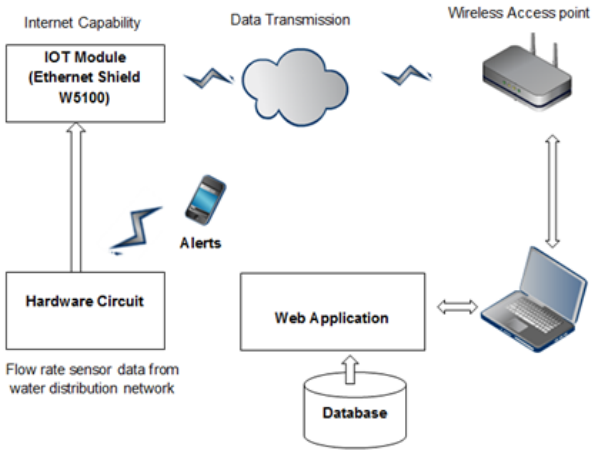

Fig. 1. Conceptual framework of existing piped water theft detection systems

A network of flowrate sensors interfaced with PLC is used to capture the flow patterns along the water distribution system. A range of normal water distribution flowrate is determined and coded on the PLC. Any variation outside the set range is treated as theft and reported to authorities via a sms using a GSM modem interfaced with PLC. Different technologies such as IOT, SCADA are used to relay the sensor data to a remote database serving a web application which takes a record of the daily patterns in water distribution [14]. The remotely controlled server running web application enables the authorities to look into the state of the network whenever it is required. The programs that run on these systems are fixed and rule-based yet the water theft problem is very dynamic and sophisticated in nature. This calls for more intelligent approaches to the problem hence the proposal of this Machine Learning framework for water theft prediction using Random Forest classifier.

Prediction models have been used greatly in different disciplines to address different issues. However, their performance are focused on similar aspects, that is, (i) use of data sets that are appropriate and with enough data (ii) prediction model performance assessed using appropriate performance metrics (iii) using statistical tests to evaluate the reliability of the results (iv) using data different from the training data to validate the predicted models[15].

In this work, we model and predict piped water theft using data collected over a period of 3 hours during an experiment. Water from a reservoir was let in the interconnected distribution network and its state was varied using suction pump and diversionary flow. The proposed model can be incorporated in existing systems at the server end to address the shortcomings of drawing conclusions based on fixed hardcoded rules. 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.

II. RELATED WORK

Significant number of research initiatives has been undertaken in response to water theft challenges that face water distribution companies around the globe. To the best of our knowledge, till to-date, no work has been done on application of Machine Learning Predictive Models in piped water theft detection. The authors in [14] use Arduino for central processing interfaced with water flow rate sensors, solenoid valve and flow meter. The microcontroller captures water flow rate readings which are compared with the fixed set value of water flow. Any reading that deviates from the set norm value is regarded as theft. A mobile application called Cayenne is also proposed for prepaid water purchase. With the use of Ethernet Shield W5100, data from the system can be uploaded to the application. This IOT module is dependent on the WizNet5100 Ethernet chip datasheet with a system IP stack for handling TCP and UDP packets [16].

The proposed system tries to address over flow and over utilization issues. Incorporating SCADA and PLC into the system has been recommended for future work. Remote access to the monitoring system by use of the IOT module and secure data transmission through use of the different Ethernet protocols is a major contribution of this study though some forms of water theft including meter reversals, meter tampering and meter bypass have not been fully addressed.

A closer approach to that of [14] is the use of PLC and SCADA for monitoring and the integrated system based on flow rate readings captured using the PLC interfaced flow rate sensors, water level readings in the reservoir taken by the level sensor, proximity sensors and pressure transmitters [17]. Water distribution pressure is maintained using the level reading of the reservoir as the reference value and maintaining a water level above the set value. However, the conclusive inference based on pressure differences remains unanswered because in the conventional distribution

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

networks, pressure rise and fall is based on water usage and time of day or rate of water consumption [17]. Therefore, designating a nominal value or range could lead to misleading theft alerts.

Similar approaches have been applied by A. Mancharkar, Rakesh and IShan [18] and A. Gantala and P. Nalajala [19] who interfaced flow rate sensors with microcontroller and using a fixed set value of flow rate, water theft is deducted. Similar work by A. Rahat saw the use of PLC, level sensor, Solenoid valve and time base feature to detect water theft perpetrated by way of pressure manipulation [20]. These works also greatly make use of Solenoid valves for automatic shutdown of water as soon as theft is identified.

H.A Gaikwad and P.V.G Puranik [8], J. Tharanyaa, A. Jagadeesan and A. Lavanya [9] and P.D.B Madihali and P.SS Ittannavar have proposed different applications that make use of an array of flow rate sensors interfaced with microcontrollers as processing units to detect and report theft. Mechanisms for remote alerting have also been proposed in their work [9].

The different research initiatives herein discussed conform to the idea that any form of distribution pressure alteration or interruption in the distribution network can significantly affect water supply. However, it should be noted that not all forms of water theft directly interferes with distribution pressure. The major form of pressure disruption is illegal connection. With other issues such as pipe bursts and leakages, water pressure can still be interfered with and hence reported as theft [21]. With activities such as meter tampering, meter reversals and meter bypass where the water pressure is not affected directly, such systems are rendered inefficient in detecting the problem. This remains a major research gap which this work aims at closing. Using data captured from a distribution network over a period of time, a trained Machine Learning model can be used to predict the existence of theft more accurately.

Till now, no work has been done on piped water theft prediction using Machine Learning models. In this work the latest technology is utilized to efficiently solve the piped water theft problem using server side processing.

A Machine Learning theft prediction model using Random Forest Classier has been proposed. The model was trained and tested on data collected by remote flow rate sensors distributed over the network over a period of 3 hours and its performance evaluated on four indexes, that is; accuracy, precision, recall and F-measure. We further evaluated our model using benchmark methods, that is; Support Vector Machine (SVM), Logistic Regression (LR) and K-Nearest Neighbour (KNN).

A. State of the Art

The existing traditional piped water theft detection systems are not comprehensive in nature; that is, they can only detect theft in cases where pressure is interrupted. In scenarios where the client themselves perpetrate the theft by manipulating the water meters, such methods are unable to detect the theft since no form of pressure disruption may be caused in the distribution network. These methods are also based on hard coded rules running on the microcontrollers which can fail in many cases. Additionally, their inability to adapt to the continuous changes in the water distribution patterns is another drawback. Therefore, in this project, we track the patterns of water distribution over time and train a Random Forest Classifier which can use the received data from flow rate sensors laid along the distribution network to predict changes in flowrate along the network.

III. MATERIALS AND METHODS

In this section, a thorough description of the proposed random forest classifier has been made followed by a discussion of the Benchmark models that is SVM, LR, and KNN.

A. Random Forest

RF is a statistical learning model which works well with small to medium data compared to Neural Network which require large training data [22]. Random Forests can be precisely defined as a combination of tree predictors in which each tree is dependent on values of an independently sampled random vector whose distribution is same for all the trees within the forest.

The traditional classifier Model is vulnerable to over fitting and has quite a limited accuracy. To improve the classification accuracy, multiple models are aggregated; a method referred to as ensemble or classifier combination [23]. This involves integration of several single classifiers also referred to as weak classifiers. The results of classification of all weak classifiers determine the final classification result. This involves an ensemble of trees which are grown and allowed to vote for the majority class. Random vectors governing the tree growth in the ensemble are generated. This is a way of growing the ensembles.

Bagging is an early example whereby to grow each and every tree in the ensemble, random selection is made without replacement from the training set examples. Secondly, the split selection involves selecting the split randomly from among the p best splits and this is done at each node [24]. The third approach involves selecting the training set from a random set of weights on those examples within the training set.

RF has the capability of processing continuous data and data of high-dimensionality and discreteness. The results of random forest are less affected in noisy data. A great number of theories and research experiments have proved random forests to have the ability of generating high accuracy and efficiency. Additionally, random forest is one of the most popular frontier research fields in data mining and bioinformatics, among others [25].

B. Benchmark Methods

i) Support Vector Machine

Support Vector Machine (SVM) is a supervised learning model that performs classification by constructing hyperplane in a high dimensional space that maximizes the margin between classes [26].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

The most important idea here is to obtain a partition hyperplane that has a maximum margin in sample space. The main aim of SVM method is mapping the sample space into a high dimensional feature space or even infinite dimension sometimes referred to as Hilbert space.

This is done through nonlinear mapping techniques such that nonlinear separation problem in original sample space is transformed into a linearly separable problem in the feature space [27].
Given a training dataset $T = \{(x_1, y_1), ..., (x_N, y_N)\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})T$ is a sample, y_i is the label of the sample, N is the number of samples, and m is the number of attributes. The partition hyperplane can be described as
$f(x) = w^T x + b$, where $w = (w_1, w_2, \ldots, w_m)T$ (1)
is the normal vector and b is the displacement.

ii) Logistic Regression

Logistic Regression is a supervised learning classification algorithm well known to be the gold standard prediction method [28]. LR is a statistical model that describes the relation between predictor variables denoted by $x' = (x_1, x_2, \ldots, x_p)$ and a response variable, which is a two value categorical variable [29], that is, the "normal flow" or "abnormal flow" as investigated in the study. The conditional probability of occurrence of water theft (abnormal flow) can be written as $K(Y = 1|x) = \pi(x)$; then, the LR model for k predictor variables can be written as;

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}} \quad (2)$$

Where $0 \leq \pi(x) \leq 1$
The logit transformation is a useful transformation of logistic regression, defined as below;

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (3)$$

The odds ratio (OR) associated with one unit change in x_j is represented with $e^{(\beta_j)}$

iii) K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a simple instance-based classifier that stores all cases and assigns the document to the majority class of its k nearest neighbours [26]. KNN algorithm is sometimes also referred to as; case based reasoning, example based-reasoning, instance based learning, memory-based reasoning and lazy learning. KNN has been used in many applications such as pattern recognition and statistical estimation, among others.
The method used by KNN to classify data is non-parametric in nature that is; does not exploit the underlying probability distribution assumptions of the data set. Two broad categories of KNN are recognized; structure less NN techniques where entire data is grouped into training and test sample data and structure based NN techniques which are based on some of the structures of data such as; orthogonal structure tree (OST), axis tree, ball tree, nearest future line and central line [30]. Distance is evaluated from training point to sample point with the lowest distance point referred to as nearest neighbor.
The commonest application of KNN has been seen in datasets with continuous attribute data. The algorithmic steps of KNN proceed as follows.

1. Finding the *k*-training instances that are closest to an unknown instance
2. Picking the classification that is most commonly occurring for the *k* instances.

Various ways are used to measure similarity between two instances of n attribute values. For each and every measure, the following three requirements apply.
Assume dist (p, q) be the distance measure between two points p, q then;

$$dist(p,q) \geq 0 \; and \; dist(p,q) = 0 \; iff \; p = q \quad (4)$$

$$dist(p,q) = dist(q,p) \quad (5)$$

$$dist(p,r) \leq dist(p,q) + dist(q,r); \quad (6)$$

Equation (6) is referred to as the "triangle in equality". It states that the shortest distance between any two points is a straight line [31]. Z score standardization and minimum-maximum normalization are used for continuous [31] data. KNN suffers from a number of drawbacks. These include; low efficiency and dependency on the selection of good values for *k* [33].

IV. PROJECT DESCRIPTION

A circuit prototype was designed in proteus, with two Arduino microcontrollers designated as slave and master programmed and interfaced with flowrate sensors, GSM (Global System for Mobile Communication) modem, and LCD (Liquid Crystal Display). A brief description of these components is presented in the next section. Experimentation was done over a 3-hour period and water flow rate data collected in 10 seconds interval. This data was tabulated, preprocessed and analyzed. 80% of the data was used for training a Random Forest classifier and 20% was used for testing purposes. The model was evaluated on accuracy, precision, recall and F-measure. Training and Testing was also done on benchmark models that are SVM, LR and KNN and the resulting performance measurements were compared with the proposed model.

A. Design

i) Block Diagram

An experimental prototype was designed as shown in figure 2 and used for data collection. The system adopts a nominal value of flow rate after a delay of about 2 minutes. This is in line with the water supply mechanisms by water

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

distribution companies which strives to supply water in a given pressure zone at a range of predefined pressures. Though some fluctuations may occur, the distribution pressure cannot fall above or below a certain value [32].

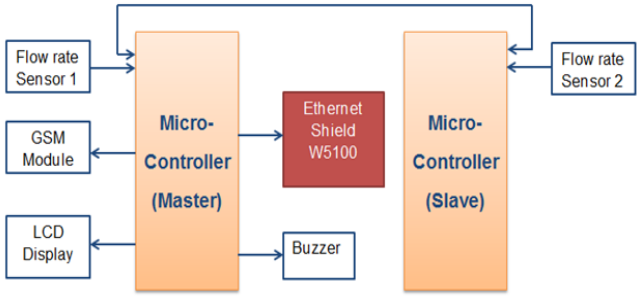

Fig. 2. Block diagram of the designed prototype

ii) Flow chart

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

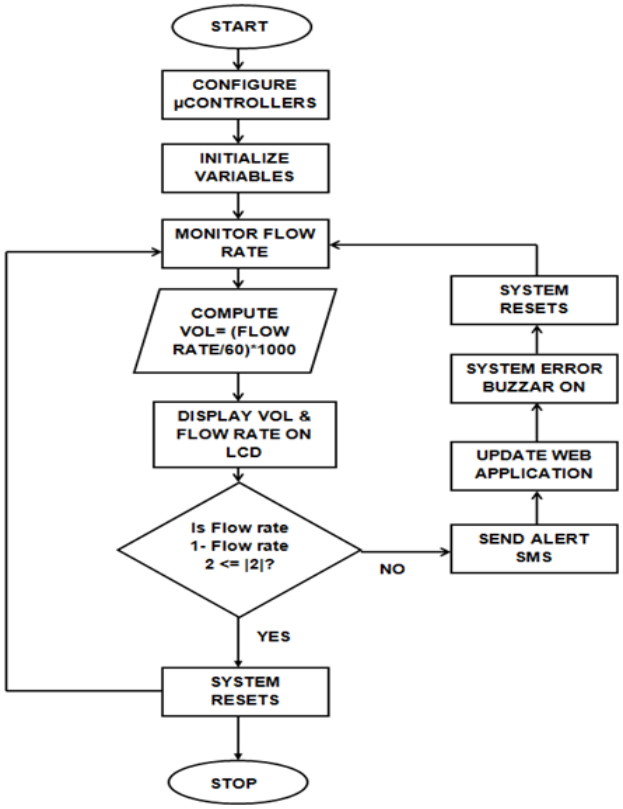

Fig. 3. Flow chart

B. Components Description

i) Arduino Uno

The functionality of Arduino is based on ATmega328 datasheet. It has 14 pins (input or output pins) referred to as digital pins. There are 6 analog input pins, 16 MHz ceramic resonator, power jack, USB connection, an ICSP header, and a reset button. With the use of the Arduino board, the microcontroller can be powered by a computer using a USB cable or directly with a battery or AC to DC adapter. The distinctive feature of the Arduino Uno from other boards that precedes it is that it has the FTDI USB serial driver chip [33].

Fig. 4. Arduino Uno

ii) Flowrate Sensor

The flow rate sensor functions on the principle of Hall effect which states that a voltage difference is induced in a conductor transverse to the electric current and the magnetic field perpendicular to it [34]. This idea is applied in the flow rate sense using a rotor shaped like a propeller. A liquid flowing through the sensor pushes against the fins of the rotor thereby causing a rotation. The rotation of the rotor induces a voltage producing an output of about 4.5 pulses from the sensor for every volume of 1 litre of water passing through the rotor per minute. The magnet attached to the rotor shaft generates a changing magnetic field which induces this voltage [37]. The flow rate sensor's data cable (yellow) is connected on pins D2 and D3 of the Arduino Uno which are also called interrupt pins.

Fig. 5. Flowrate sensor

iii) GSM Modem

SIM900 is a quad band GSM modem. The modem board communicates via UART as it has pins labeled TXD and RXD for UART communication. The board is powered by 12V but it also has a 5V line for Arduino power. It uses AT (attention) commands for its communication control.

Fig. 6. GSM Modem

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

V. CIRCUIT DESIGN AND EXPERIMENTATION

The water theft detection system was designed and simulated in proteus software with two Arduino Uno microcontrollers programmed as Master and Slave because every Arduino Uno can support a single flowrate sensor. The Master controls the entire circuit including the slave microcontroller which is instructed by the master.

The GSM module was interfaced with the Master and uses serial communication to send sms alerts to a Mobile phone. The SMS states the flow rate along a given service line and the volume consumption at that instant.

We further performed system testing on the complete integrated system for validation purposes. To test the compliance of the complete system to the system requirements, we fitted water pipes of 20mm diameter with flow rate sensors connected at branch points as shown in figure 7. Water was then supplied through the system using a water reservoir placed at an elevation from the ground surface.

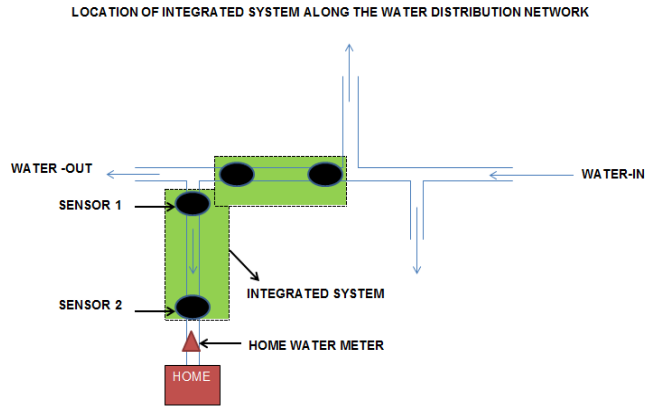

Fig. 7. Conceptual Water Distribution Network

A. Data Collection

Experiments were ran to depict the real life scenario of water distribution and flow rate values were read from the LCD in a 10 seconds interval. The water flow rate was varied by connecting a suction pump to the distribution Network and alternating it with a diversion to vary the pressures. The alert messages of abnormal flow were received on the Mobile phone and a record taken accordingly.

After 3 hour period, 1017 values were captured and tabulated. Some variables such as absolute difference between the established norm value at the instant and the recorded flow rate, volume consumption were computed and included in the dataset.

Absolute difference, A_d = |flow rate – established norm value|

Volume flow at the instant = flow rate * time

A total of 6 variables were recorded in our dataset.

B. Data Analysis and Processing

We noted that after about 2 minutes of water flow, there was stability in the flow rate values read. Every time the system detected abnormality in the flow, there was reset and therefore the system would adopt a new value of flow rate designated as the normal flow. Water theft was based on these fluctuations; the bigger the fluctuation, the more likely was the conclusion that there was error in the distribution system. Taking the adopted normal values at every system reset point; the water flow pattern shown in figure 8 was observed.

This pattern aligns with the idea that water distribution companies supply water at a given fixed range of water pressures in each and every pressure zone which is the major basis for theft detection [9]. A pressure district is a collection of varying pressure zones and there can be varying distribution pressures within a pressure district, however, the fluctuation does not go outside the normal ranges as depicted in our flowchart presented in figure 3.

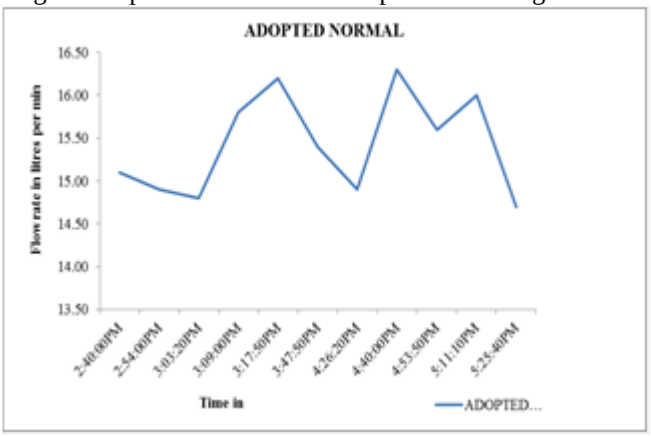

Fig. 8. Water flow pattern

From the graph obtained, it can be observed that the maximum fluctuation of the flow rates throughout the entire experiment was (16.3-14.7= 1.6litres/min).

VI. MODELING AND PREDICTION

The model training and testing took a number of tasks which involved Importation of the necessary libraries that is; numpy, pandas, matplotlib, and seaborn after installing dependencies in anaconda. Jupyter Notebook was used with Python 2. A thoroughly cleaned dataset in csv format was read. Understanding the data patterns is an Important step in data analysis [35] hence data exploration and visualization was done. To improve the performance of the classifiers, some feature engineering was undertaken. Here, domain knowledge and data mining techniques are used to extract features from raw data.

Feature engineering is also considered as some form of applied machine learning. In our experiment we have constructed Random Forest Classifier and other benchmark predictive models which included; SVM, KNN and LR. We used 80% of the data for training purposes and 20% for testing. Finally predictions were drawn from test data and performance measures, that is; accuracy, precision, recall and F-measure were computed from the confusion matrix.

A. Model Evaluation

The water theft prediction model performance was evaluated using accuracy, precision, recall and F-measure

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

using the confusion matrix. A preliminary discussion of confusion matrix presented here is necessary in enabling us understand the performance metrics used for performance evaluation in our study.

Table 1 shows a confusion matrix illustrated with four main quadrants, which include; True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). The number of correctly classified positive samples is indicated my TP, the classification model may predict some negative samples incorrectly as positive, these are referred to as FP where as those positive samples that are predicted incorrectly as Negative are represented by FN. However, the numbers of negative correctly predicted samples are represented as TN [36].

Table 1: Confusion Matrix

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

Accuracy refers to the ratio of the number of correctly classified samples to the total number of samples. The harmonic mean of precision and recall of the classification is called F-measure [36]. In this definition, precision is described as the ratio of all positive results correctly classified to the total number of positive results in the classification and recall describes the ratio of correctly classified positive results to the actual number of positive results that would have ideally been returned. These evaluation performance metrics were calculated using the following formulae.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Accuracy = \frac{TP}{TP+FN} \quad (9)$$

$$F-Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (10)$$

VII. RESULTS AND DISCUSSION

The results of our experiment showed that the proposed classification model achieves the highest accuracy and recall. KNN achieved a similar accuracy though its recall was lower than that of the proposed mode as shown in table II.

Table- II: Performance metrics values

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

A visual graphical representation of the variation of performance scores of the four classification models is shown below as a bar graph.

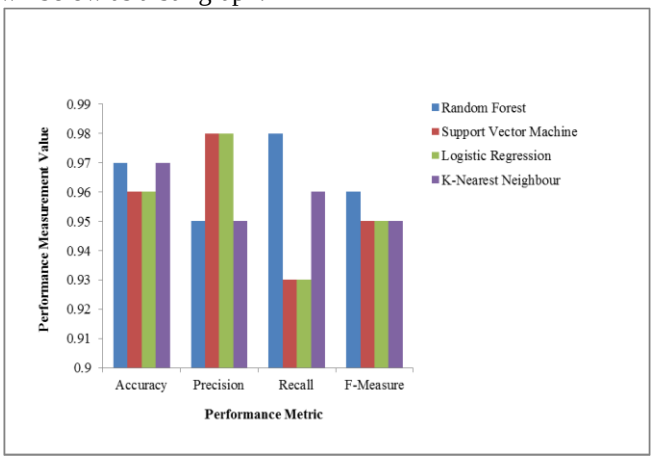

Fig. 8. Performance metric results of the prediction models

VIII. CONCLUSION

The existing research initiatives depend solely on hardware to deduct the occurrence of piped water theft. These works have basically made use of flow rate sensor data collected from water distribution and fed into the microcontroller memory to deduce the state of the network based on hardcoded rules. In this study, we have resolved the issue by integrating hardware and software approaches to intelligently predict theft. The programmed electronic circuit was used as a data collection module for the purposes of training the Random forest classifier. The data collected over a 3 hour period in 10 seconds intervals gives a fair representation of real life scenario of piped water distribution. This data was cleaned and preprocessed which gave a total of five variables used in the study. These were; flowrate, adopted norm value, absolute difference between adopted norm and flowrate value at the instant, volume and status of the distribution system. The data was then divided into training set and testing set which was 80% and 20% of the dataset respectively.

The Random Forest classifier was then constructed and evaluated on four performance metrics, that is; accuracy, precision, recall and F-measure. Comparing the results with three competitive approaches that is, SVM, LR and KNN used in our experiment, no considerable differences were noted other than the recall value which was highest in RF model. Reliable feature importance estimate and estimation of test error with efficiency that no repeated model training is required are two notable advantages of RF compared to the benchmark approaches. Our trained model can be incorporated in existing systems at the server end to intelligently deduce the existence of water theft. For deployment purposes, water flowrate values along distribution lines can be collected over a period of time and used to generate a dataset that can then be applied in the model training. In such a case, unsupervised techniques may be applied.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

FUTURE WORK

The data collected depicted a common pattern of water flow which enabled quick model training. The high accuracy obtained ascertains the efficiency that can be achieved with Machine Learning approaches. Flow rate data from water distribution networks can be collected for a reasonable period of time and be used to model and predict water theft using our model or unsupervised learning techniques. As an initiative to eradicating piped water theft problem, a complete system can be further developed.

REFERENCES

[1] The United Nations World Water development report (UNWWD), water for sustainable world, UNESCO, United Nations Chapter 4, 2015

[2] Pradeep K. Naik, 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

[3] World Health Organization, WHO *World Water Day Report*, 2015. Retrieved from http://www.who.int/water_sanitation_health/takingcharge.html

[4] W. Gumindoga, G. Takawira, L. Fengting, N. Innocent & M. Clifton, Health Safety of Drinking Water Supplied in Africa: A Closer Look Using Applicable Water Quality Standards as a Measure, Springer Science+Business Media B.V. 2017

[5] 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

[6] United Nations Department of economic and social affairs (UNDESA), "International Decade for Action 'Water for life'," United Nations, New York, 2005-2015.

[7] Ministry of water and Environment, "Uganda Water and Environment Sector Performance Report," Government of Uganda, Kampala, 2018.

[8] National Water and Sewerage Corporation (NWSC), "National Water and Sewerage Corporation (NWSC)," NWSC, 23 06 2013. [Online]. Available: http:www.nwsc.com [Accessed 7 05 2015].

[9] H. A. Gaikwad & P. V. G. Puranik, "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.

[10] J. Tharanyaa, A.Jagadeesan & A.Lavanya "Theft identification and Automated Water Supply System Using Embedded Technology," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 2, no. 8, pp. 3727-3734, 2013.

[11] P. D. B. Madihalli & P. S. S. Ittannavar, "Smart Water Supply Management," International Journal of Emerging Trends in Electrical and Electronics (IJETEE – ISSN: 2320-9569), vol. 10, no. 9, pp. 77-79, 2014.

[12] V. Felbab-Brown, "Brookings Mountain West Lecture Series, University of Nevada, Las Vegas," University of Nevada, Las Vegas , 20 02 2015.[Online].Available: https://www.unlv.edu/brookingsmtnwest/past-lectures-events. [Accessed 25 10 2015].

[13] National Water and Sewerage Corporation, "National Water and Sewerage Corporation (NWSC)," Littlegate publishing, 01 04 2015. [Online]. Available: http://www.littlegatepublishing.com. [Accessed 11 10 2015].

[14] 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.

[15] 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

[16] A. E. shiled, "Arduino Ethernet shield," 24 04 2010. [Online] Available: https://docs.rs-online.com/0084/0900766b80db991d.pdf. [Accessed 27 04 2020].

[17] A. Panchal and K. Degate, "Automated water supply system and water theft identification using PLC and SCADA," International Journal of Engineering Research and Applications ISSN : 2248-9622, vol. 4, no. 4, pp. 67-69, 2014.

[18] A. Mancharkar & K. Rakesh, "Automated water distribution system for smart city using PLC and SCADA," International Journal of Emerging Technologies and Engineering (IJETE), vol. 3, no. 3, 2016

[19] A. Gantala & P. Nalajala "Public water supply monitoring to avoid tampering and water man Fraud using smart water meter system," nternational Journal of Mechanical Engineering and Technology (IJMET) ,vol. 8, no. 8, pp. 1194-1201, 2017.

[20] A. Rahate & G. Ashwini, "Automated water distribution system and theft detection," International Journal of Advance Engineering and Research Development, vol. 4, no. 4, pp. 650-654, 2017.

[21] M. Panwar "Issues, Challenges and Prospects of Water Supply in Urban India," IOSR Journal Of Humanities And Social Science (IOSR-JHSS), vol. 20, no. 5, pp. 68-73, 2015.

[22] R. A. Berk, "Statistical Learning from a Regression Perspective," Springer, Switzerland, 2019

[23] F. K. N. &. Z. J. P. Jian-Bina W U, "A review of Technologies on random Forests," Statistics and Information Forum , vol. 3, pp. 32-38, 2011.

[24] L. Breiman, Random Forests, CA 94720 : Statistics Department University of California Berkeley, 2018.

[25] H. L. L.-W. &. L. M. R P Adams, "A physiological series dynamics based approach to patient monitoring and outcome prediction," EEE journal of Biomedical and Health Informatics,, vol. 19, no. 3, pp. 1068-1076 , 2015.

[26] K. S. R. &. D. M. Harikrishna, "Children's Story Classification in Indian Languages Using Linguistic and Keyword-based Features," ACM Trans. Asian Low-Resour. Lang. Inf. Process, vol. 19, no. 2, p. 22 pages, 2019.

[27] M. Z. &. H. Shi, "A Novel Support Vector Machine Algorithm for Missing Data," ACM, 2018.

[28] T. B. &. D. P. 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.

[29] B. M. &. K. M. W. K. J Wang, "Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: A case study of Taiwan," Computers in Biology and Medicine, vol. 47, p. 147–160 , 2014.

[30] N. B. &. Vandana, "Survey on nearest neighbor techniques," International Journal of Computer Science and Information Security (IJCSIS), vol. 80, no. 2, 2010.

[31] M. &. B. D. 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.

[32] 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

[33] A.Uno-Farnell, Arduino Datasheets, Ivrea: Italy, https://www.farnell.com/datasheets/1682209.pdf.

[34] Arduino, "How to interface an Arduino Uno with a Flow rate sensor to measure a liquid," Arduino, 23 03 2018. [Online]. Available: ttps://maker.pro/arduino/tutorial/how-to-interface-arduino-with-flow-rate-sensor-to-measure-liquid. [Accessed 16 09 29].

[35] A. Z. &. A. Casari, Feature Engineering for Machine Learning Principles and Techniques for data scientists, Sebastopol, California: O'Reilly Media, April, 2018 First Edition.

[36] Z. X. &. F. Y. S. Yang, "A patient outcome prediction based on Random Forest," ACM ISBN 978-1-4503-7640-2/19/1 , 2019.

[37]"Working with water flow sensors and Arduino," 05 08 2019. [Online].Available: https://www.electroschematics.com/12145/working-with-water-flow-sensors-arduino. [Accessed 23 12 2019].