An Efficient scheme for Water Leakage Detection using Support Vector Machines (SVM) – Zig

Akshay Kothari, M. Balamurugan

Abstract— Water is one of the most essential and valuable resources for all living beings, yet in the present day, there is a scarcity of it. Half of the water loss in large cities and industries is due to leaks and illegal lines. 10%-20% of water loss can be reduced by detecting leaks but without the presence of advanced monitoring systems, this problem is typically worsened. Monitoring the consumption and leak detection for such large areas is a challenging task. To overcome this issue a small prototype is prepared called Zig. Zig is designed for both household and industrial purposes. Its main aim is to monitor the flow and consumption of water at different levels of a building like a first-floor and so on which may represent some industrial and household situation. This work focuses on pressure/flow monitoring method to reduce the operational cost and also to detect leakage. One of the machine learning algorithms, Support Vector Machines (SVM) has been applied to detect the leakage and it is compared with Random Forest algorithm to show that proposed scheme is detecting water leakage better.

Keywords: Machine Learning, Support Vector Machines, Random Forest, Water Leakage Detection

1. INTRODUCTION

Water is one of the most essential and valuable resources for all living beings, yet in the present day, there is a scarcity of it. Every year more than 32 billion m3 of treated water is wasted by leakage from water distribution systems and 16 billion m3 is distributed to people without proper invoice in terms of corruption, theft, or improper metering [11].

Half of the water loss in large cities and industries is due to leaks and illegal lines. Generally, Water loss is very costly problem because water is one of the precious natural resources [10]. 10%-20% of water loss can be reduced by detecting leaks but without the presence of advanced monitoring systems, this problem is typically worsened.

Monitoring the consumption and leak detection for such large areas is a challenging task. To overcome this issue a small prototype is prepared called Zig. Zig is designed for both household and industrial purposes. Its main aim is to monitor the flow and consumption of water at different levels of a building like a first-floor and so on which may represent some industrial and household situation.

The main aim of this paper is to study various leakage detection techniques which help in detecting leaks in the water pipes. The paper focuses on pressure/flow monitoring method to reduce the operational cost and also to detect leakage.

Artificial Intelligence is one of the ways to find out the method based on the information of water flow meter readings to detect leakage. The Support Vector Machine (SVM) can be used as identified pattern to solve water leakage problem. Further, SVM works on Structural Risk Minimization (SRM) in order to find out the appropriate hyperplane which divided 2 classes of input and will deal with the data in higher dimension. [2].

2. LITERATURE REVIEW

2.1 Study about different leakage detection techniques.

In [3], Obeid, Abdulfattah M., et al. presents a comprehensive survey on various leakage detection and leakage localization techniques like ultrasound, acoustic, electromagnetic, pressure/flow monitoring methods and so on. Also, the research work carried out to implement these various techniques using various platforms such as a microcontroller, application-specific integrated circuit and others are reviewed and discussed with their advantages and disadvantages [3]. Furthermore, an integrated energy aware system-on-chip solution has been identified obtain more reliable results. In [10], Adedeji, Kazeem B., et al. had mentioned that there are lots of scope is still available to do further research in water leakage detection problem. Further, they had highlighted few existing techniques for the same problem also. In [11], Darsana, P., et al. summarizes the recent studies carried out for the leakage detection. The study concludes that most of the existing approaches have faced the difficiluties due to several reasons such as lack of available data , expensive sensors and lack of available attributes. From the above study, the leakage detection techniques can be classified as follows:

- Externally-based techniques
- Internally-based techniques

The earlier techniques to detect water leakage require more involvement from the human being due to personal inspection at every place. It costs more human oriented and time consuming method and its reliability is very low while detecting leakages. The externally-based techniques generate a leak alarm by using sensors that are installed on the pipe. However, installation is very complex and cost of the system

Revised Version Manuscript Received on April 05, 2019.

Akshay Kothari, M.Tech Student, Department of CSE, Christ (Deemed to be University), Bengaluru, Karnataka, India

M. Balamurugan, Associate Professor, Department of CSE, Christ (Deemed to be University), Bengaluru, Karnataka, India (E-Mail: Balamurugan.m@christuniversity.in)

is high in these types of leakage detection techniques. On the other hand, internally-based techniques monitor internal pipeline through various parameters such as pressure, flow, and fluid temperature by using field sensors. The complexity of installation and system costs are usually low in this type of leakage detection techniques.

2.2 Study about leakage detection without applying machine learning to it.

In [1], Berglund, Andrew, et al. describes the measuring of pressure at various points to detect leakage. To determine the leaks that most closely approximate the observed pressure pattern absolute differences between observed and simulated pressure values at the sensors are calculated. No machine learning algorithm was used. Instead, the framework consists of successive linear approximation methods based on linear programming.

In [4], Sithole, Bheki, et al. carries out a research to present a low-cost Smart Water Meter Device which is capable of providing reports of the current household consumption with possible leakage detection. Here flow meter sensors are deployed to obtain the consumption of the water. No machine learning algorithm has been used to detect leakage.

2.3 Study about leakage detection by applying machine learning to it.

In [9], Mashford, John, et al. describes the use of SVM analysis to detect the leakage in simulated water pipe networks. The water pipe network was simulated using a simulation tool EPANET. SVM is applied to the data obtained from pressure sensors or flow monitoring devices which monitors the pipe network. In [2], Salam, A. Ejah Umraeni, et al. describes the use of SVM method on the drinking water data from a local company that had been modelled using EPANET 2.0 software. Here, in this research, the pattern of the pressure change is analyzed computationally using SVM method to detect the leakage. The accuracy obtained by using SVM method for the leakage location is 76.14%. In [7], Kang, Jiheon, et al. describes the leakage detection using the sound and vibration produced by the water leaking from pressurized pipes. The leakage detection system proposed is the fusion of one-dimensional convolutional neural network and SVM. The proposed method achieved the leakage detection accuracy of 99.3%.

In [8], KEMBA, Joseph, et al. describes the leakage detection in the water network where highly sensitive pressure sensors are deployed. The pressure data was obtained from EPANET simulation tool. SVM with RBF kernel is applied to the pressure data to detect leakage. As a result, 90% leakage detection accuracy was achieved. From the above study, it is known that the leakage data has some pattern in it which is to be identified. One of the ways to perform pattern identification is to apply machine learning algorithm. In this proposed work, we have applied one of the machine learning algorithms is Support Vector Machine (SVM) which can work on SRM principle and divided 2 classes of input and will deal with the data in higher dimension [2].

3. IMPLEMENTATION

The purpose of this research is to detect leakage from the flow meter readings which are deployed on a Zig by using Support Vector Machine method.

3.1 Zig

Zig is a small prototype designed for both household and industrial purposes. Its main aim is to monitor the flow and consumption of water at different levels of a building like a first-floor and so on which may represent some industrial and household situation.

One of the leakage detection technique used here is flow monitoring technique.

The data is obtained from the readings of flow meter installed in each floor. For the implementation of this paper, we simulated the zig in LabVIEW.

3.2 Schema of Research

The main aim of this research is to detect leakage from the data collected from the flow meter readings which are deployed on the zig. The general stages of the research are described in Figure 1.

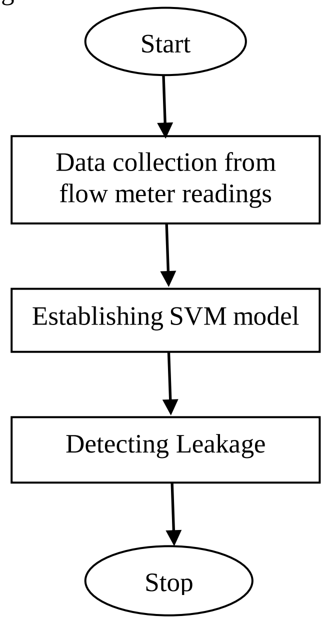

Figure 1 Schema of Proposed Work

3.3 Type of Data

The data utilized in this research is the flow meter readings which are utilized to detect leakage by applying Support Vector Machine method.

For processing data and detecting leakage, SVM is used to obtain good prediction results with a high accuracy. To detect leakage, a lot of leakage data is to be obtained from the actual data which continuously intends to create many leakages in the pipe. So the data is obtained from the simulation of zig in LabVIEW. The simulation of the zig is shown in Fig. 2.

3.4 Data Collection

The first step in collecting data is to simulate the zig in LabVIEW wherein the flow of water is 300 liters per hour. The water flow meter gives the exact reading of 5 liters per minute if there is no leakage or 0 liters per minute if there is no water flow. A reading less than 5 liters per minute describes that there is leakage which is to be detected using SVM method.

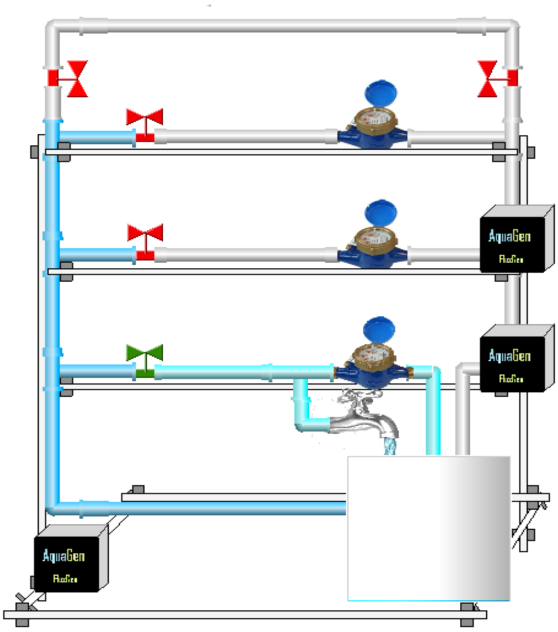

Fig. 2 Simulated Zig

3.5 Support Vector Machines (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. The algorithm creates an optimal hyperplane based on the labelled training data provided.

In two dimensional spaces, this hyperplane divides the plane into two parts where in each class deceits on the either side of the line. Support vectors are the data points nearest to the hyperplane and are considered as the critical elements of a dataset, because the position of dividing the hyperplane will change if they are removed.

SVM is a machine learning technique which can be used to perform both binary classifications and regression tasks. They are correlated to Artificial Neural Networks (ANNs) and have various properties that make them better than artificial neural networks.

Reasons that make SVM better than artificial neural networks are as follows:

- SVMs can work effectively on high dimensionality input spaces.
- SVMs can manage small samples of training and testing set.

The advantages of SVM over neural networks and drawbacks of neural networks are stated in [6].Some of the drawbacks of ANNs to which SVMs are less sensitive are as follows:

- Lack of generalization
- Local optima
- Uncontrollable Convergence

Additionally they have generated excellent performance in generalization on a variety of problems such as Bioinformatics, Fault diagnosis, Image detection, Power systems and Text categorization [8]. Both regression and classification can be performed using SVM. SVM will have two forms of output such as real number and predicted class which are associated with input pattern while acting as regressor and classifier respectively. While SVM performing as a regressor, it behaves like function approximate. SVMs will be trained on trained data set which has number of input patterns and the associated output values or categories. They can then be evaluated on a testing set to determine the appropriate performance [8].

The purpose of the training in Support Vector Machine (SVM) method is to make the system able to identify the data pattern of each leakage data. The data used for training is the input data which is obtained from the flow meter readings.

In practice, kernals have been used to implement SVM algorithm. A kernel is a similarity function. It is a function which is provided to machine learning algorithm that takes two inputs and describes how similar they are. There are different kernels that SVM can use to process the data. They are as follows:

- Linear Kernel
- Polynomial Kernel
- Radial Basis Function (RBF) Kernel

Linear Kernel:

The Dot-product is called as kernel in linear kernel. Kernel function can be written as:

$$K(x,x') = 1 + sum(x * x')^d \tag{1}$$

Where, *x* and *x*` are input vectors and support vectors respectively.

The kernel defines a distance measure or the similarity between new or input data and the support vectors. The dot product is the similarity measure used for linear kernel SVM because the distance is a linear combination of the inputs.

Other kernels such as polynomial and radial basis function kernel can be used that transform the input space into higher dimensions.

Polynomial Kernel:

Polynomial kernel can be used instead of dot-product. Kernel function can be written as:

$$K(x,x') = 1 + sum(x * x')^d \tag{2}$$

Where, d is the degree of the polynomial and *x* and *x*` are input vectors and support vectors. This is same as the linear kernel when d = 1. The polynomial kernel allows for curved lines in the input space.

Radial Basis Function (RBF) Kernel:

The more complex kernel that can be used is RBF kernel. Kernel function can be written as:

$$K(x,x') = \exp(-\gamma * ||x - x'||^2) \tag{3}$$

Where, γ is a parameter that is specified to the learning algorithm, *x* and *x*` are input vectors and support vectors and is the squared Euclidean distance between the input vectors and support vectors. The value of g ranges from 0 < γ < 1. γ can be defined as:

$$\gamma = \frac{1}{2\sigma^2} \tag{4}$$

Where, σ^2 is variance.

$$\gamma = \frac{1}{2\sigma^2}$$

3.6 Development of SVM Analysis Method

There are different kernel functions which SVM uses to process the data such as Linear Kernel, Radial Basis Function (RBF) Kernel and Polynomial Kernel. To ease the learning process of SVM and to define the support vector, kernel determination is must. In this research the input data obtained from the flow meter readings is processed by the Radial Basis Function (RBF) kernel. RBF kernel has mainly two parameters:

C – the misclassification penalty parameter

The C parameter tells the SVM optimization how much one wants to avoid misclassifying each training example. It trades of misclassification of training examples against simplicity of the decision surface. The low value of C makes the decision surface smooth, while the high value of C gives the model freedom to select more samples as support vectors which aims at classifying all training examples correctly.

γ – the kernel function parameter

The γ parameter describes how far the influence of a single training example reaches, where, low value means ‘far’ and high value means ‘close’.

The values of C and γ which can provide best results may differ for different problems. To find the best pair of C and g for a given problem, the experiment conducted in [5] recommends sequences of C and γ that is exponentially growing.

In this research, the different values of C and γ were applied to obtain the optimal results. Also for different values of C and γ, Accuracy, RMSE and Correlation Coefficient can be calculated as :

Accuracy:

Accuracy is calculated in % which is the measure of the total correctly predicted output. In other way, accuracy can be calculated by dividing the number of predictions over the total number of a test sample and multiplying the result by one hundred.

RMSE:

RMSE is frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed [8]. RMSE can be calculated by using

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - yi)^2} \tag{5}$$

Where, number of test samples is denoted by *n*, *y* is target and *yi* is predicted output.

Large values of RMSE means poor predictions of the targets whereas small values of RMSE mean good predictions of targets.

4. RESULTS AND ANALYSIS OF SVM MODEL

4.1 Results

As discussed earlier, SVM can be used to perform both classification and regression. When SVM is used for classification, the output is the predicted class associated with the input pattern and when it is used for regression, the output is the real number associated with the input pattern.

Also, there are three different kernels that SVM uses to process the data. They are Linear kernel, Polynomial kernel and RBF kernel. As SVM are learning machines, kernel determination is must to ease the learning process. The following research uses RBF kernel.

There are 1440 sets of input data and 1440 sets of target data. These data were divided randomly into a training set of 1080 and a testing set of 360 which is in the ratio of 70:30. The SVM model then was trained using the training set.

The different sets of values were applied to the training parameters of the SVM which are C – the misclassification penalty parameter and g – the kernel function parameter, to find the best pair which gives optimal results. Also for different pairs of values of C and g, %accuracy, correlation coefficient and RMSE values are calculated.

Table 1 presents a list of the different set of values used for C and g with %Accuracy, Correlation Coefficient, and RMSE values. The different set of values for C and g which are highlighted in the table denotes the highest range of accuracy achieved with the lowest range of RMSE. The highest accuracy achieved was 91.66% for values C =2 and g = 0:003 and the RMSE achieved for this pair of values is 0.28 which is lowest than all the other pairs. Furthermore, the value of correlation coefficient is 0.82 which is close to 1. It means that there is a strong positive relationship between the predicted and the actual target. Also, large RMSE conveys poor prediction whereas small RMSE describes good predictions of targets.

TABLE 1: PARAMETER VALUES FOR SVM

Sr. No	*C*	γ	%Accuracy	Correlation Coefficient	RMSE
1	0.1	0.0002	86.1	0.67	0.37
2	0.25	0.0009	88.6	0.74	0.33
3	**0.5**	**0.0006**	**90**	**0.78**	**0.31**
4	**1**	**0.0006**	**90.2**	**0.78**	**0.31**
5	**2**	**0.003**	**91.66**	**0.82**	**0.28**
6	**4**	**0.005**	**91.1**	**0.80**	**0.29**
7	8	0.02	88.8	0.74	0.33
8	16	0.02	89.1	0.75	0.32
9	32	0.04	88	0.72	0.34
10	64	0.01	88.8	0.75	0.33

After obtaining the best parameter values of C and g the results i.e. actual and the predicted target are described in various forms such as confusion matrix, roc-curve graph and line graphs.

Confusion Matrix

The predicted results of a classification problem can be best described or summarized by confusion matrix. The key to confusion matrix is that the number of correct and incorrect predictions are summarized with count values and broken down by each class.

For example, in a binary classification, suppose class 1 is positive and class 2 is negative and if observation in class 1 it means observation is positive and observation in class 2 means observation is negative, then the confusion matrix summarizes the result in the following terms.

- ***True Positive (TP):*** Actual observed value is positive and is forecasted to be positive too.
- ***False Negative (FN):*** Actual observed value is positive, but it is categorized to be negative.
- ***True Negative (TN):*** Observation is negative and is also predicted to be negative.
- ***False Positive (FP):*** Analysis Result is negative, but it is predicted to be positive.

Figure 3 describes the confusion matrix without normalization of the predicted target values and the actual target value and figure 3 describes the confusion matrix with normalization of the predicted target values and the actual target values.

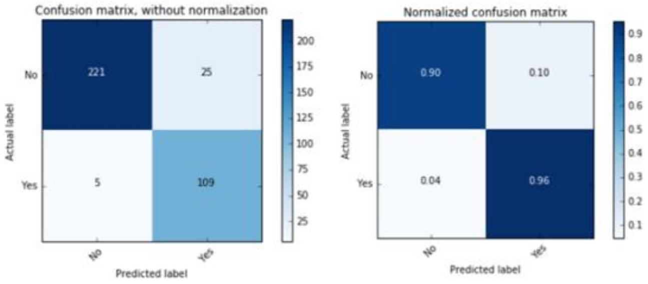

Figure 3 Confusion Matrix with and without Normalization

In the above figure 3 of confusion matrix without normalization, 0 and 1 on the labels describes that there is no leakage detected at value 0 and the value 1 means there is leakage detected. As mentioned before there are 360 test dataset, it can be observed from the figure that 221 TP values were predicted correctly for no leakage, and 109 TN values were predicted correctly where leakage is detected.

Also, 25 FN values were predicted as leakage detected but actual values shows that there is no leakage detected. These are the values where the valve of the zig was suddenly opened and closed when the tap was kept on to produce leakage. And finally, 5 FP values were predicted as no leakage detected whereas the actual value shows that there is leakage which in this case is very less compared to FN values. Figure 5.2 shows the same confusion matrix as in figure 5.1 but with normalization where the values are ranged from 0 to 1.

Confusion matrix without normalization:

$$\begin{pmatrix} 221 & 25 \\ 5 & 109 \end{pmatrix}$$

Confusion matrix with normalization:

$$\begin{pmatrix} 0{:}90 & 0{:}10 \\ 0{:}04 & 0{:}96 \end{pmatrix}$$

4.2 Receiver Operating Characteristic (ROC) Curve

ROC curve known as Receiver Operating Characteristic curve is a metric used to evaluate the classifier quality output. In ROC curves,Y axis carries the true positive rate and X axis has the false positive rate. The top left corner of the plot indicates the “ideal” point where zero for FP rate and one for TP rate. This is not very realistic, but it does mean that a larger Area under the Curve (AUC) is usually better [12].

The “steepness” of ROC curves is also key factor where is ideal to maximize the TP rate while minimizing the FP rate. ROC curves are usually used for binary classification to study the output of a classifier which in this case is used as the binary classification includes whether leakage is detected or not.

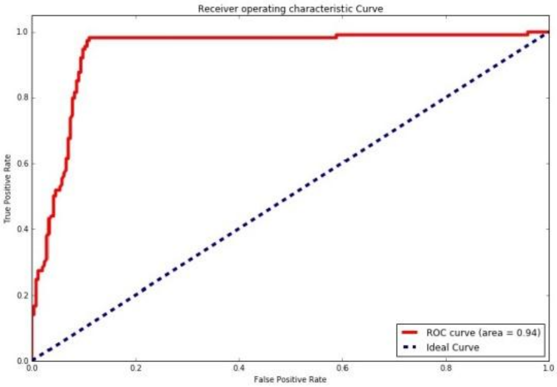

Figure 4 ROC curve

Figure 4 describes the ROC curve obtained by plotting the true positive rate against the false positive rate as described in the confusion matrix at various threshold settings.

The dashed line in figure 4 also described as ideal curve as it shows the true positive rate of 1 and false positive rate of 0. The ROC curve as shown in figure IV describes that the larger area under the curve is usually better.

The results are described by using line graphs in various forms as follows:

- Line graph showing actual target versus predicted target of different sets of test data.
- Line graph showing leakage data and leakage detected

4.3 Line graphs for actual target versus predicted target for different sets of test data

Figure 5,6,7 and 8describes the line graph for actual target versus predicted target for different sets of test data. Here, in each graph the total number of test data is 40 and the value for leakage detection is 1 whereas value 0 indicates that there is no leakage. The blue line in the graph describes the predicted target and orange line is the actual target. Both the predicted and actual target are plotted together. If the predicted target values line is exactly on top of the actual target line then the predicted target are same as actual target. On the other hand, the point at which both the predicted target values line and

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

actual target values line do not meet, it means that the predicted target is not same as actual target.

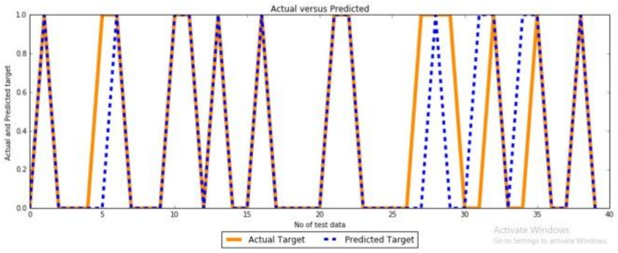

Fig 5 Actual vs Predicted for first test dataset by SVM

The above figure describes the case where the model detected that there is no leakage at point 5, 27 and 29 but the actual values detected leakage and at point 31 and 34, the model detected leakage but there is no leakage according to actual values.

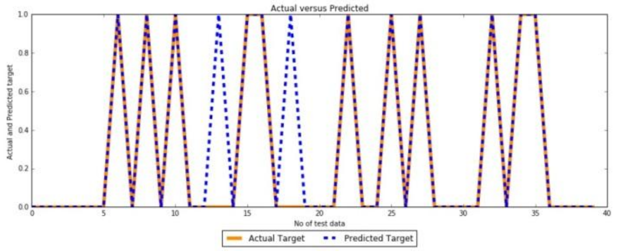

Fig 6 Actual vs Predicted for Second test dataset by SVM

The above figure describes the case where the model detected leakage but there is no leakage detected according to actual values.

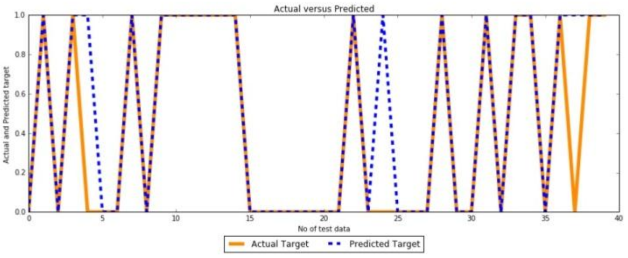

Fig 7 Actual vs Predicted for Third test dataset by SVM

The above figure describes a case where the actual value shows that there is no leakage detected but the predicted values detected leakage.

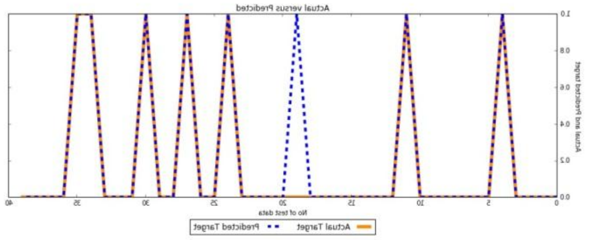

Fig 8 Actual vs Predicted for Fourth test dataset by SVM

The above figure shows the case where the model predicted leakage but there is no leakage according to actual value.

Line graphs showing leakage data and leakage detected

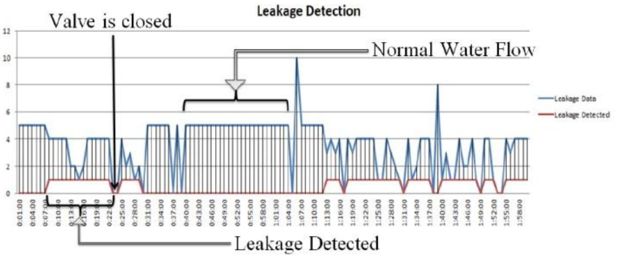

Fig 9 Leakage Detection Graph

Figure 9 describes the line graph for the leakage detection. The blue line in the graph is the processed data obtained from the flow meter readings and it also contains leakage data. The blue line at zero indicates that valve is closed. The red line shows the possible detected leakages according to the input values. The red line at zero indicates there is no leakage or the valve is closed whereas red line above zero indicates leakages.

4.4 Validation of Results

Studying the attributes of the estimation function and evaluating it on the same set of data is procedurally false: a method that would replicate the identifiers of the samples that it has seen would have a proper score but it may fail to predict anything useful on yet-unseen data. This situation is called overfitting. The part of the data set would have been kept as test data to avoid over fitting problem in supervised algorithms experiments.

The basic approach k fold CV has been used here. The following procedure is used for each of the k “folds”:

- A model is trained using *k-1* of the folds as training data;
- The proposed model is evaluated on the remaining part of the data

In the proposed research, k-fold cross validation is used to validate the results where k = 10. The cross-validation score obtained for 10-fold CV is as follows:

Cross validation score: [0.91, 0.88, 0.85, 0.87, 0.87, 0.9, 0.87, 0.89, 0.86, 0.84]

4.5 Final Overview of Results obtained by SVM model

- RMSE: 0.288675134595
- Correlation coefficient: [[1. 0.82] [0.82 1.]]
- Accuracy: 0.916666666667 (91.6%)
- Cross validation score: [0.91 , 0.88 , 0.85 , 0.87 , 0.87 , 0.9 , 0.87 , 0.89 , 0.86 ,
- Confusion matrix, without normalization: [[221 25] [5 109]]
- Normalized confusion matrix: [[0.9 0.1] [0.04 0.96]]

4.6 Comparing Results of SVM Model and Random Forest Model

As discussed in the previous sections, both SVM and Random Forest can be used to perform both classification and regression. The following section compares both the model in terms of accuracy, RMSE and correlation coefficient. Also, comparison is done to show the number of correctly predicted output by both the model.

There are 1440 sets of input data and 1440 sets of target data. These data were divided randomly into a training set of 1080 and a testing set of 360 which is in the ratio of 70:30. SVM and the Random Forest model then was trained using the training set.

Table 2 describes the comparison of SVM model and Random Forest model established on the same training and testing dataset. Also the results have been validated by performing 10-fold cross validation.

TABLE 2: COMPARING RESULTS OF SVM AND RANDOM FOREST MODEL

Measures	SVM	Random Forest
%Accuracy	91.6	87.7
RMSE	0.28	0.34
Correlation Coefficient	0.82	0.72
No. of True Positive values	221	220
No. of False Negative values	25	26
No. of True Negative values	109	96
No. of False Positive values	5	18

From the above table it can be observed that SVM achieved better accuracy then the random forest model by setting the parameters of C = 2 and γ = 0:003 for SVM. Also, it can be seen that RMSE value for SVM is less compared to random forest. Small value of RMSE means good predictions of output which in this case can be understood that the predictions made by SVM model are better than the predictions made by the random forest. The correlation coefficients for both the models are close to 1 but correlation coefficient of SVM model is closer to 1. Correlation coefficient value closer to 1 means that there is a strong positive relation between the actual and the predicted targets.

Also, it can be seen that SVM model correctly classified the predicted target for leakage detection more than the Random Forest model. Finally, SVM model predicted only 5 target as leakage not detected but in actual target there is leakage which is very less in compared to Random Forest.

4.7 Line graphs for actual target versus predicted target obtained by SVM and Random Forest for different sets of test data

Figure 10 and 11 describes the line graph for predicted target by SVM versus predicted target by Random Forest for different sets of test data. Here, in both graph the total number of test data is 90 and the value for leakage detection is 1 whereas value 0 indicates that there is no leakage. The black line in the graph describes the predicted target by Random Forest and orange line is the predicted target by SVM. Both the predicted target are plotted together. If the line is exactly on top of the other line then the predicted targets are same. On the other hand, the point at which both the predicted target values line do not meet, it means that the predicted targets are not same.

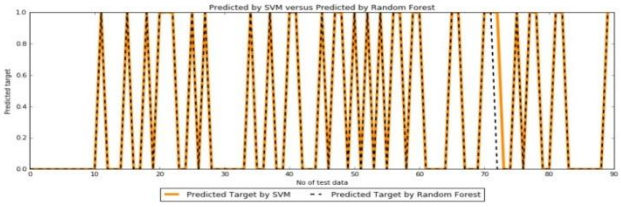

Fig 10 Predictions of SVM versus Predictions of Random Forest for first test dataset

As seen from the above figure, the predicted values by both the model is almost same except at point 72 where the predicted targets are different.

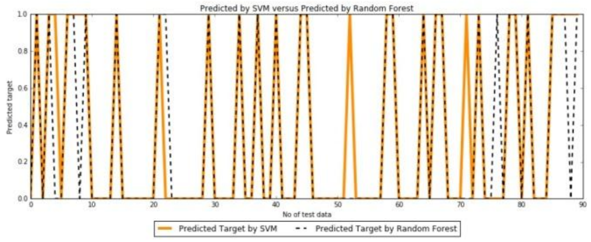

Figure 11 Predictions of SVM versus Predictions of Random Forest for second test dataset

The above figure describes the case where at many different points the predicted targets by both the model are not same.

4.8 Line graphs showing leakage data and leakage detected by both SVM and Random Forest models

Figure 12, 13, 14 and 15 describes the line graph for leakage data and leakage detected by both SVM and Random Forest for different sets of test data. Here, in all the graphs the total number of test data is 90 and the value for leakage detection is 1 whereas value 0 indicates that there is no leakage. Also for the leakage data, the value 5 and 0 indicates there is no leakage and any value less than 5 indicates there is leakage which is to be detected by both the models i.e. SVM and Random Forest. The blue dotted line in the graph describes the leakage data. The green line in the graph describes whether the SVM model detected the leakage or not and red line describes the leakage detection done by Random Forest classifier.

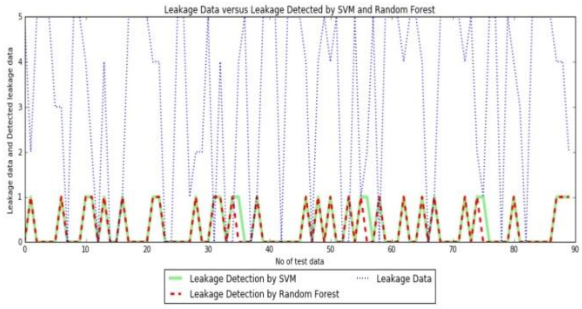

Figure 12 Leakage data versus Leakage detected by both SVM and Random Forest for first test dataset

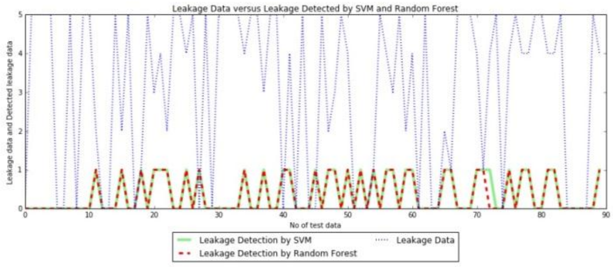

Figure13 Leakage data versus Leakage detected by both SVM and Random Forest for second test dataset

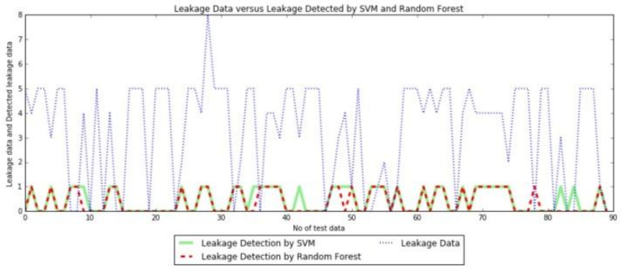

Figure 14 Leakage data versus Leakage detected by both SVM and Random Forest for third test dataset

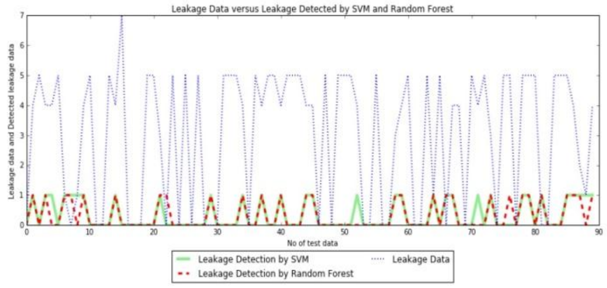

Figure 15 Leakage data versus Leakage detected by both SVM and Random Forest for fourth test dataset

The proposed work carried out in this research is to detect leakage of water in a zig which is simulated in the LabVIEW and to detect the leakage machine learning algorithm is applied on the dataset obtained from the simulated zig. From the literature survey, it was found that the internally-based technique for leakage detection is more suitable as it is easy to deploy and the system cost os low.

This research uses pressure/flow monitoring technique where flow meters are deployed at each level of zig. The leakage data was then collected through the LabVIEW where the zig is simulated.

Machine learning algorithm like SVM is applied on the leakage data and for the parameters of C = 2 and γ = 0:003 the model achieved 91.6% accuracy.

The SVM model was compared against the Random Forest model which was also applied on the same set of leakage data and the random forest model achieved the accuracy of 87.7%. Hence, SVM proved to be a better model to detect the leakage. The results are described in various forms like confusion matrix, roc-curve and line graphs.

5. CONCLUSION AND FUTURE WORK

This paper has described water leakage detection using a machine learning algorithm named Support Vector Machines. The main aim of this study was to detect leakage by using less number of attributes in data.

Also, the Zig which was constructed for obtaining water consumption and detecting the leakage uses flow meters because of which the operational cost is reduced.

The SVM model was trained using the data of the flow meter readings using RBF kernel function the best parameter pairs with values $C = 2$ and $\gamma = 0.003$. The model obtained the RMSE = 0.28 which is lowest than the other RMSE of the different parameter values which describes good prediction of the targets. The accuracy achieved for detecting leakage was 91.6%.

The future work of this research is to increase the accuracy by combining more than one machine learning techniques by using less number of attributes in the data and also to predict leakage localization using the same less attributed data.

REFERENCES

1. A. Berglund, V. Areti, D. Brill and G. Mahinthakumar (2017), ”Successive Linear Approximation Methods for Leak Detection in Water Distribution Systems”, Journal of Water Resources Planning and Management, vol. 143, no. 8, p. 04017042.
2. A. E. U. Salam, M. Tola, M. Selintung and F. Maricar (2014), ”A leakage detection system on theWater Pipe Network through Support Vector Machine method,” 2014 Makassar International Conference on Electrical Engineering and Informatics (MICEEI), pp. 161-165.
3. A. Obeid, M. BenSaleh, S. Manzoor Qasim, M. Abid, M. Jmal and F. Karray (2016), ”Towards realisation of wireless sensor network-based water pipeline monitoring systems: a comprehensive review of techniques and platforms”, IET Science, Measurement & Technology, vol. 10, no. 5, pp. 420-426.
4. B. Sithole, S. Rimer, K. Ouahada, C. Mikeka and J. Pinifolo (2016), ”Smart water leakage detection and metering device,” 2016 IST-Africa Week Conference, pp. 1-9.
5. C. Hsu, C. Chang, and C. Lin, ”A practical guide to support vector classification (2003)”, Department of Computer Science and Information Engineering, National Taiwan University, pp. 1-16.
6. D. Silva, J. Mashford, and S. Burn (2009), ”Computer aided leak location and sizing in pipe networks”, Urban Water Security Research Alliance.
7. J. Kang, Y. Park, J. Lee, S. Wang and D. Eom (2018), ”Novel Leakage Detection by Ensemble CNN-SVM and Graph-Based Localization in Water Distribution Systems”, IEEE Transactions on Industrial Electronics, vol. 65, no. 5, pp. 4279-4289.
8. J. Kemba, K. Gideon and C. N. Nyirenda (2017), ”Leakage detection in Tsumeb east water distribution network using EPANET and support vector regression,” 2017 ISTAfrica Week Conference (IST-Africa), pp. 1-8.
9. J. Mashford, D. De Silva, S. Burn and D. Marney (2012), ”Leak Detection in Simulated Water Pipe Networks Using SVM”, Applied Artificial Intelligence, vol. 26, no. 5, pp. 429-444.
10. K. Adedeji, Y. Hamam, B. Abe and A. Abu-Mahfouz (2017), ”Towards Achieving a Reliable Leakage Detection and Localization Algorithm for Application inWater Piping Networks: An Overview”, IEEE Access, vol. 5, pp. 20272-20285.
11. P. Darsana and K. Varija (2018), ”Leakage Detection Studies for Water Supply Systems— A Review”, Springer, Water Resources Management, vol. 78, pp. 141-150.