Machine Learning Model Comparison for Leak Detection in Noisy Industrial Pipelines

Dimitrios Kampelopoulos
Department of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
dkampelo@physics.auth.gr

George N. Papastavrou
Department of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
gepapast@physics.auth.gr

George P. Kousiopoulos
Department of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
gkousiop@physics.auth.gr

Nikolaos Karagiorgos
Department of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
nkaragio@physics.auth.gr

Sotirios K. Goudos
Department of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
sgoudo@physics.auth.gr

Spyridon Nikolaidis
Department of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
snikolaid@physics.auth.gr

***Abstract*— In this paper, two machine learning techniques are applied and compared in order to model leak detection in pipelines in noisy environments. A set of accelerometers, mounted on the surface of the pipes, was deployed for the data acquisition process. Measurements of noise during normal operating conditions were recorded as well as measurements of leaks, generated on various distances from the sensors. Using these measurement data, a training set was created from their time-domain and frequency-domain features. The leak detection process is then modeled as a binary classification problem (leak detection or not). For this problem, two machine learning classification techniques were evaluated, the support vector machines and the decision trees. The results for each learner are compared with the original data from the test dataset using representative performance indicators and, overall, high levels of accuracy are achieved.**

Keywords— machine learning, model analysis, leak detection, noisy environments, data-acquisition, feature extraction, model accuracy

I. Introduction

It is well known that, pipeline network systems are established for the transportation of fluids, like water, petroleum liquids or even gases. The problem is that because of their broad use, not only for industrial but also for urban purposes, their maintenance and continuous monitoring is considered critical. Leak events could cause disastrous consequences, both environmental and economic, that are globally perceived [1].

Many methods have been proposed in the past for the problem of leak detection in pipes. Some of the most applied techniques include time-domain reflectometry [2], negative pressure methods [3], as well as acoustic-based methods [4]. Acoustic methods are non-invasive techniques that depend on the vibro-acoustic signature of the noise and leak signals on the pipe. A variety of sensors can be utilized like piezoelectric, hydrophones, and accelerometers with the latter being a common choice due to their relatively low cost, their sensitivity in lower frequencies, flat response and the sharpness of its peaks when correlated [5].

For leak detection, feature extraction and pattern recognition are the main approaches to the problem. Previous research includes extraction from time-domain features such as mean value, variance, peak value, kurtosis, skewness, shape factor etc [6,7], and frequency-domain approaches like the cepstrum analysis [8] and the filter diagonalization method [9].

From the extracted features, in order to make an educated decision about the presence of a leak, some researchers employ machine learning techniques [10,11]. The problem of leak detection is a reduced down to a classification problem, either binary to determine the presence of a leak or not or multi-label to classify the magnitude of the leak as well [12].

In this work, a monitoring system for leak detection is proposed utilizing machine learning algorithms to detect potential leaks based on variances from the pipe's normal operation noise. The described method is non-invasive and relies on the acoustic signals taken from a set of accelerometers, mounted on the pipe's surface. When the sensors are set, the pipe's noise is measured and kept as a reference. From the raw data, several time and frequency-domain features are extracted and, after being compared to the reference, they are used to build the training dataset for the algorithms. The dataset contains measurements of noise during pipe's normal operating conditions, as well as measurements of artificial leaks. So, the leak detection problem is reduced to a binary classification task (noise or leak). The focus of this paper will be to test two machine learning models, Support Vector Machines (SVM) and Decision Trees (DT) and evaluate their results in terms of accuracy and recall.

The proposed approach is utilized as part of a larger pipeline monitoring system, currently under development, and is only responsible for the leak detection process. Once a leak is detected, another independent process of the system is activated to locate it. The localization procedure is achieved by utilizing wave propagation properties and calculating the difference in time of arrival between the local sensors' acoustic signals. The current study will be focused on the machine learning approach for the leak detection process and the localization procedure will not be discussed in this paper.

II. Experimental Setup

The experimental procedure includes noise measurements of the pipes under normal working conditions and measurements of artificial leaks. More specific, the facility of the Hellenic Petroleum S.A. in Thessaloniki, Greece was chosen to conduct the measurements. Two different pipeline configurations were selected, including the line Dr-1452 (2-6 inches diameter, 8.3kg/cm^2 pressure, 40°C temperature) carrying atmospheric air and the line E-1404 (8-inch diameter, 4kg/cm^2 pressure, 30-35°C temperature) carrying water. These pipes' geometry is complex with some of their sections containing machinery (automatic valve reliefs, compression chambers, cooling structures, etc) and their length can reach up to 300 meters. To produce the artificial leaks, there are multiple valves across the pipes each with different diameter.

The acquisition system consists of a set of three accelerometers (PCB 352C33), the National Instruments NI-9232 data acquisition card with the NI LabView software and a laptop for the program execution. The accelerometer has an almost flat frequency response in the range of 0 – 15 kHz, thus a sampling rate of 25.6 kHz was chosen. The data acquisition card has an on-board anti-alias filter of 0.4Fs, so the system can measure frequencies up to 11 kHz with a flat response [13,14]. The data were obtained in intervals of 1 sec and the dataset contains about 4 hours of measurements for both noise and leaks.

III. Feature extraction – Noise Reference

As the literature suggests, there are many features for leak detection, including time and frequency-domain features [7,8,15]. As a first approach, a variety of the suggested features, including rms, variance, skewness, kurtosis, shape factor, crest factor, integral in time, integral in spectral density, entropy, clearance factor and peak frequency, were examined. Some of them are easily computable in real time and others more computationally expensive. In this work, the focus will be on features that can be derived in real time and can be easily implemented in hardware, for future development. Also, since there is high correlation between certain features, like rms with variance, some of them were discarded without any negative impact in accuracy. In fact, highly correlated features tend to reduce the algorithm's reliability and introduce bias [16]. Therefore, some of the features had to be eliminated, and the remaining are:

- Rms
- Skewness
- Kurtosis
- Clearance factor
- Crest factor
- Peak frequency

Apart from them, the correlation function and the sum of the absolute difference in spectral energy between each band of 250 Hz (labeled as slope sum) were also used as features and will be described later in this section.

As mentioned, the system's focus is to monitor the pipe and detect potential leaks after its installation. To achieve that, a noise measurement of certain duration is kept as a reference. In this study's case, based on the periodicity of the pipe's noise, a 10 sec measurement was chosen as reference. This measurement's spectrum is then divided in bands of 250Hz and the energy of every band is calculated. Any band with energy higher than the average is filtered. These filters will then be applied to the raw data that will be used for the feature extraction for noise and leaks. In general, the noise signals generated by the machinery usually appears in lower frequencies than the leak [17]. By this adaptive filtering procedure, this difference in frequency can be taken advantage of, and by filtering the noise, the leak signal becomes more apparent and eventually easier to detect.

Apart from that, the noise characteristics of the pipe should also be considered, in a way that the reference for feature extraction is updated accordingly. More specific, the variance in noise levels is mostly caused by the operation of heavy machinery (automatic valve reliefs, compression chambers, cooling structures, etc) as well as the geometry of the pipe's structure. Before measuring the reference, the periodicity of the machinery's operation as well as their distance from the sensors, should, also, be examined. For example, in the Dr-1452 gas pipeline, a relief valve is activated every 5 minutes, changing the pressure on certain sections and thus, the noise levels. Knowing the period of the machinery, as well as the difference in noise levels between each consecutive valve relief event, for each section of the pipe, allows the system to update the reference when needed. However, if the new reference levels are higher than expected, that reference is discarded, since a transient event or a potential leak might have occurred during the measurement. As a result, the reliability of the system, also, relies on how frequently the reference is updated and the duration of the reference measurements.

The noise reference was not only used for the filtering but also for the estimation of the correlation feature. More specific, from the noise reference, a signal of 1 sec duration is saved to be correlated to every next measurement. The noise that is present on the pipes is not random since it is generated from machinery operating in certain frequencies. This results in the noise being correlated to a large extend. Filtered noise signals from different time frames, when correlated, the sum of their cross-correlation function squared values is relatively high. However, correlating a noise signal with a leak signal results in a lower sum value. That difference in the results allows for the exploitation of the correlation function as a metric for leak detection.

Another metric is how the spectrum variates for leak and noise. When a leak is present, certain frequency bands have increased energy compared to the noise spectrum. To measure that change in spectrum, the absolute difference in energy between each band and its next was calculated. The sum of these values is used as a feature and labeled for this study as the slope sum.

Apart from that, the rms values and the energy in each band are different for different pipe structures or even different sections of the same pipe. In order to make these features more generally applicable, their percentage change from the noise reference is calculated and used as a feature. In other words, the rms and slope sum features are the percentage change from the noise reference and not the measured values. As a result, a low change in the rms values is an indication for the absence of a leak and vice versa.

It must, also, be noted, that during the noise measurements (a total of 4200 seconds of the recorded data), about 200 (4.8%) of them contain the transient event of a relief valve occurring near the sensors. Events like that are characterized by a sudden increase in kurtosis and crest factor value, in such a way that the output of the system can be ignored temporarily until the event dies down. However, those transient events are present in the training data and were labeled as noise. In some cases, these events' duration is higher than one second and the system is only able to determine their beginning and their end. The intermediate seconds might be falsely classified as leaks while evaluating the algorithms and may reduce their resulting accuracy scores. Although this is no concern, since the system will not produce a classification decision until the end of the transient event.

IV. Model Comparison – Results

The fist dataset contains 13 extracted features from the accelerometer noise and leak data. From the correlation matrix (fig. 1) calculated over all the features of the dataset, a lot of them appear highly correlated, like rms, variance, entropy, integral and psd integral, as well as kurtosis and shape factor. From the above, only rms and kurtosis were selected, and the resulting dataset is comprised of 8 features.

By using a grid search cross validation process, the most optimal hyper-parameter values for the Support Vector Machines (SVM) and Decision Tree (DT) classifiers were derived. More specific, each combination of parameter values was tested multiple times and the one with the consistently highest accuracy score was selected. The resulting parameters for each classifier are shown below:

- SVM: C = 60, kernel = polynomial, degree = 7, gamma = scale.
- DT: max_depth = 7, max_features = 7, min_samples_split = 9, min_samples_leaf = 6.

The two resulting classifiers were, then, trained, tested and compared with each other. As shown in the accuracy score boxplot of fig. 2, while the DT seems to be a better fit (97.9%) than the SVM (97.1%) based on the accuracy score, the SVM is more consistent, exhibiting a lower standard deviation value (0.57%) than the DT's (0.63%).

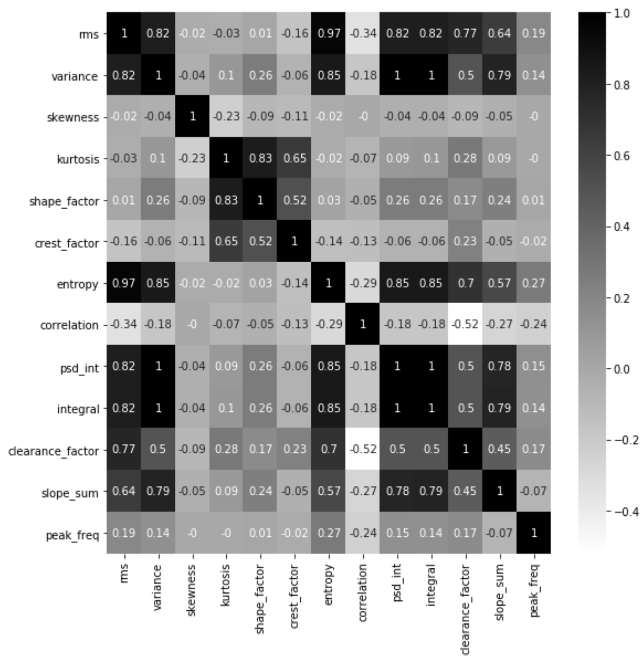

Fig. 1. Correlation matrix for the 13 extracted features of the overall data.

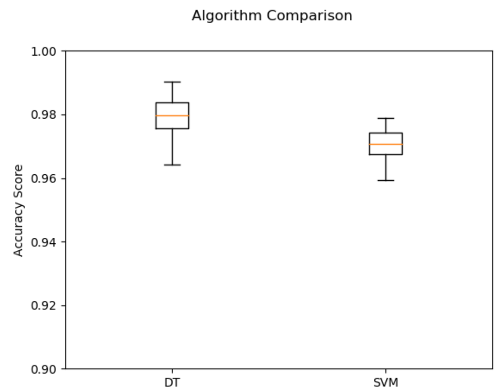

Fig. 2. Boxplot of the accuracy score for each machine learning algorithm.

A more detailed review of the classifier's quality of predictions can be derived from the correlation matrices for each class of the tested data (fig. 3,4). The confusion matrix indicates the ratio of true and false predictions for each class. As illustrated, both classifiers exhibit a low percentage in false noise predictions (1.6% for SVM and 1.2% for DT). In contrast, false leak prediction percentage is higher (3.9% for SVM and 4.2% for DT). As mentioned, this can be attributed to the transient events present in the dataset, that the system during normal operation can detect and temporarily disable its output until the end of that event.

A more detailed review can be derived from the classification reports for each algorithm (Tables I, II). The classification report contains metrics about the quality of the algorithm's prediction, like the precision, the recall and the f1-score. First, the number of true positives, false positives, true negatives and false negatives is calculated for each class. For example, for the noise class, a true positive means

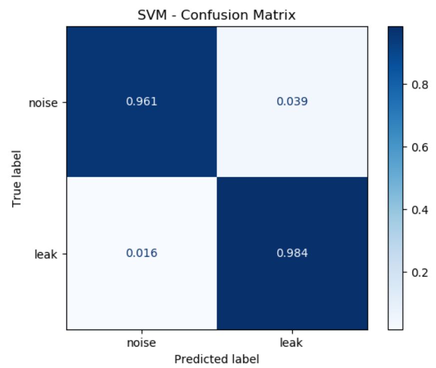

Fig. 3. Confusion matrix for the support vector machine classifier.

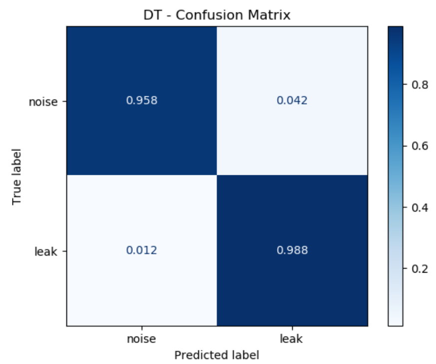

Fig. 4. Confusion matrix for the decision tree classifier.

that a noise measurement was correctly predicted as noise, whereas a false positive means that a measurement was falsely predicted as noise when it was a leak. The classification report's metrics, mentioned above, are defined by the following equations:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F1\ score = 2 * \frac{Precicion * Recall}{Precision + Recall}$$

In this study's case, it is important for the model to generate as few false alarms (false leak predictions) as possible. Considering the above equations, a high recall score for the noise class is indicative of small number of false alarms (false negatives of the noise class). As a result, the model with the highest recall score on that class will be the most suitable. In this case, the highest noise recall is achieved by the SVM classifier (97.07%) contrary to the DT (95.83%).

TABLE I. CLASSIFICATION REPORT FOR THE SVM CLASSIFIER

SVM Classification Report			
Class Labels	**Precision**	**Recall**	**F1-score**
Noise	0.9688	0.9707	0.9647
Leak	0.9797	0.9840	0.9818
Overall Accuracy:	0.9760		

TABLE II. CLASSIFICATION REPORT FOR THE DT CLASSIFIER

DT Classification Report			
Class Labels	**Precision**	**Recall**	**F1-score**
Noise	0.9769	0.9583	0.9675
Leak	0.9786	0.9883	0.9834
Overall Accuracy:	0.9780		

V. CONCLUSION

In this work, the effectiveness of two machine learning models, support vector machines and decision trees, was evaluated on a leak monitoring system. The system based on a noise reference measurement, and through adaptive filtering, records the percentage change of some time and frequency-domain features indicative of the signal's characteristics. Based on the training dataset, both models exhibit accuracy score levels over 97%, rendering them suitable for this application. However, by examining their recall score, the support vector machines seem to be slightly less prone to false alarms than the decision trees.

To improve the current configuration, the introduction of a new class to label the transient events can be studied. The main cause for a false alarm is the appearance of transient events. Training the algorithms to be able to detect them may increase the recall percentage of the noise class. While the current implementation can detect the beginning and the end of these events based on 1 sec measurements, their intermediate part is falsely predicted as leak due to their similarity in nature. A solution would be to consider measurements of higher duration (i.e. 10 sec), though, drastically reducing the system's response time and increasing the computational cost. To avoid that and keep the system's response to 1 sec or possibly reduce it, more signal features should be introduced to the existing ones. Their effect and their feature importance, as well as their impact in the system's response and computational cost is a subject for future development.

ACKNOWLEDGEMENT

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the name "Research – Create – Innovate" (project code: T1EDK-00791).

REFERENCES

[1] EPA. Drinking Water Infrastructure Needs Survey and Assessment: Third Report to Congress. EPA 816-R-05-001, 2005

[2] A. Cataldo, G. Cannazza, E. De Benedetto and N. Giaquinto, "A New Method for Detecting Leaks in Underground Water Pipelines", IEEE Sensors Journal, vol. 12, no. 6, pp. 1660-1667, 2012.

[3] J. Hu, L. Zhang and W. Liang, "Detection of small leakage from long transportation pipeline with complex noise", Journal of Loss Prevention in the Process Industries, vol. 24, no. 4, pp. 449-457, 2011.

[4] T. Quy, S. Muhammad and J. Kim, "A Reliable Acoustic EMISSION Based Technique for the Detection of a Small Leak in a Pipeline System", *Energies*, vol. 12, no. 8, p. 1472, 2019.

[5] Y. Gao, M. Brennan, P. Joseph, J. Muggleton and O. Hunaidi, "On the selection of acoustic/vibration sensors for leak detection in plastic water pipes", *Journal of Sound and Vibration*, vol. 283, no. 3-5, pp. 927-941, 2005.

[6] Q. Xu, L. Zhang and W. Liang, "Acoustic detection technology for gas pipeline leakage", *Process Safety and Environmental Protection*, vol. 91, no. 4, pp. 253-261, 2013.

[7] W. Caesarendra and T. Tjahjowidodo, "A Review of Feature Extraction Methods in Vibration-Based Condition Monitoring and Its Application for Degradation Trend Estimation of Low-Speed Slew Bearing," *Machines*, vol. 5, no. 4, p. 21, 2017.

[8] F. Wang, W. Lin, Z. Liu, S. Wu and X. Qiu, "Pipeline Leak Detection by Using Time-Domain Statistical Features", *IEEE Sensors Journal*, vol. 17, no. 19, pp. 6431-6442, 2017.

[9] A. Lay-Ekuakille, C. Pariset and A. Trotta, "Leak detection of complex pipelines based on the filter diagonalization method: robust technique for eigenvalue assessment", *Measurement Science and Technology*, vol. 21, no. 11, p. 115403, 2010.

[10] T. Mamo, I. Juran, and I. Shahrour, "Virtual DMA Municipal Water Supply Pipeline Leak Detection and Classification Using Advance Pattern Recognizer Multi-Class SVM," *Journal of Pattern Recognition Research*, vol. 9, no. 1, pp. 25–42, 2014.

[11] S. El-Zahab, E. M. Abdelkader, and T. Zayed, "An accelerometer-based leak detection system," Mechanical Systems and Signal Processing, vol. 108, pp. 276–291, 2018.

[12] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[13] National Instruments, "Datasheet NI 92323 AI, ±30 V, 24 Bit, 102.4 kS/s/ch Simultaneous", NI-9232, 2015.

[14] PCB Piezotronics, "ICP Accelerometer Installation and Operation Manual", Model 352C33, 2002.

[15] H. Jin, L. Zhang, W. Liang, and Q. Ding, "Integrated leakage detection and localization model for gas pipelines based on the acoustic wave method," *Journal of Loss Prevention in the Process Industries*, vol. 27, pp. 74–88, 2014.

[16] L. Toloşi and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, 2011.