# Probabilistic Multiple Model Neural Network Based Leak Detection System: Experimental Study

*Mohammad Burhan Abdulla*
*PhD Research Student*
*Masdar Institute for Science & Technology, UAE*
*mabdulla1@masdar@ac.ae*

*Randa Herzallah*
*Non-linearity and complexity research group, Aston University, UK*
*r.herzallah@aston.ac.uk*

## *Abstract*

*This paper presents an effective decision making system for leak detection based on multiple generalized linear models and clustering techniques. The training data for the proposed decision system is obtained by setting up an experimental pipeline fully operational distribution system. The system is also equipped with data logging for three variables; namely, inlet pressure, outlet pressure, and outlet flow. The experimental setup is designed such that multi-operational conditions of the distribution system, including multi pressure and multi flow can be obtained. We then statistically tested and showed that pressure and flow variables can be used as signature of leak under the designed multi-operational conditions. It is then shown that the detection of leakages based on the training and testing of the proposed multi model decision system with pre data clustering, under multi operational conditions produces better recognition rates in comparison to the training based on the single model approach. This decision system is then equipped with the estimation of confidence limits and a method is proposed for using these confidence limits for obtaining more robust leakage recognition results.*

***Keywords:*** *Multiple models, pipeline leak detection, uncertainty, paired t-test, negative pressure wave.*

## 1. Introduction

It is due to the extensively popular usage of pipelines in transportation of resources that pipelines possess significant interest (Prashanth et. al., 2011). Additionally, environment, economics and safety consequences due to leaks are major concerns that can be mitigated through the use of efficient leak detection systems (Abdulla et. al., 2013). Generally, when leak takes place, it generates a pressure drop wave that is propagated to both terminals of this pipe, causing the pressure to decrease at those terminals. This phenomenon is named “negative pressure wave” which can be considered a sign for leakage incidence. Nevertheless, detection is not that simple, since pressure might decrease due to other causes than leak happening, for instance, the increase in flow causes pressure drop. Additionally, inevitable noise associated with data acquisition systems makes tackling the problem analytically too complicated if not insolvable. Consequently, an efficient leak detection system must be capable of handling such conditions of variable and noisy operational conditions. The use of Neural Networks (NN) models to detect leaks in pipelines transporting liquefied gas was studied by Belsito et.al. (1998). It was found that operational conditions changes have significant negative effect as they cause high false alarms. Real-time leak monitoring system was suggested by Jian et.al. (2003). This system is based on negative pressure wave picked up via pressure transducers and signal processing techniques, mainly, wavelet analysis. It is noticed that pressure can be falsely recognized as leak in case of flow adjustments. To overcome this problem, flow is analyzed when the variables of pressure indicate leak and a decision is made accordingly. This system is implemented on the Shengli Petroleum Management Burea, where relative location detection error was estimated to be 0.45%. A leak detection system based on measuring inlet pressure, outlet pressure and inlet-outlet flow difference was developed by Jian et.al. (2004). They assumed that pressure and flow possess normal distribution and built the detection system accordingly.

Jinhai et.al. (2007) studied using a combination of Rough Sets (RS) and NN technology. RS was used to reduce the number of inputs handled by the NN. RS operated through three stages, namely, data acquisition for receiving historical data of the pipeline and arranging them in matrix form, then data is discretized, and finally, tree expression is used to reduce data attributes. As for the NN, 3-layer Multi Layer Perceptron (MLP) network was used where the hidden layer was composed of 30 neurons. The inputs to the MLP network were pressure, temperature, flow, turndown ratio of valves and pumps status. The leak detection accuracy is reported to be 97% in their work. Bicharra et.al. (2008) proposed an MLP based leak detection system. This system is based upon 3-layer MLP network with inlet flow, temperature, density, and pressure as inputs. The developed system resulted in 6.03% error for fixed operational conditions. However, when the system was tested under variable operational conditions, error rate increased to 29%. A Mixture Density Neural Network (MDNN) with 3 hidden units of Gaussian kernels to process signals from sensors was used by Khan et.al (2008). The system gave acceptable results as long as it is being operated on data from the same general operational conditions on which it has been trained. Barradas et.al (2009) studied using a 3-layer MLP network based pipeline leak detection system. Several architectures are investigated; those differ from each other by the number of hidden neurons, and number of input-delays. The latter showed to have significant effect on the performance. This is due to the fact that inlet flow was taken an input for the NN which requires considerable time to be reflected on the outlet flow.

A pipeline leak detection system based on gradient and slope turns rejection is studied by Liang (2012). This system acquires data during the transient periods of various working pressures. Then, features are extracted by the use of Wavelet packet analysis so that the dynamic behavior of pressure is defined, hence after the features space dimensionality is reduced using principle component analysis. The probability distribution for the resulted reduced-dimensional space is estimated using Gaussian Mixture Model (GMM) whose number of Gaussian functions is determined via Bayesian Information Criterion (BIC). The system was validated and proved effective experimentally. Corneliu et.al (2012) investigated the use of three layer General Fuzzy Min Max Network (GFMMNN) and graph theory in leak detection systems. This study showed that when the developed GFMMNN is trained and tested using data representing the patterns of variation of nodal consumptions, it provides better recognition capabilities compared to training based on of nodal heads patterns and pipe flows state estimators. Meribout (2011) developed a real-time pipeline leak detection system using a pipe-in-pipe configuration. It requires pipe transporting the fluid to be surrounded by another external pipe. The concept is that when a leak in the inner pipe takes place, the leaking fluid becomes present in the surrounding pipe, where an air-ultrasonic sensor is installed. This sensor then picks up the presence of the fluid and decides on leak event accordingly. Additionally, bi-directional microphones are installed for the purpose of leak location determination. The system was validated via laboratory scale tests; it was shown that the developed system accomplished 95% accurate leak rate determination. The leak in long horizontal subsea pipelines was studied by Seung (2010). Those pipelines were modeled mechanistically depending on pressure transverse calculations. It was noticed that leak produces a change in inlet pressure and outlet flow rate. Variables such as pressure gradient, flow rate, hold up and flowing fraction were studied comparatively in leak and no leak conditions. And the results were validated by the use of quasi-dynamic numerical simulation. The study shows that both inlet pressure and outlet flow can be taken as leak indicators, with the outlet flow being more reliable. Jinqiu et.al. (2011) studied leak detection in high noise environments. The developed system makes use of harmonic wavelet transforms to extract the features of negative pressure generated in leak occurrences, and then, it provides a time-frequency plot accordingly. Data from simulation based testing as well as from field experiment indicates the superior detection performance of harmonic wavelet analysis compared to other wavelet analysis such as Daubechies wavelet.

In this work, a continuous based leak detection data driven and multiple-model decision support system is developed based on inputs of inlet pressure, outlet pressure and outlet flow. The system is mainly composed of two parts; a K-Nearest-Neighbor (KNN) clustering technique and NN multiple processing models. The clustering technique maps incoming data to its appropriate processing NN model; the use of multiple models instead of a single model is proposed to improve the performance of the leak detection system under variable operational conditions in terms of increasing classification rates. Additionally, and in contrast to the currently operated systems, the proposed system is developed in a probabilistic manner after it has been optimized, which provides not only a decision concerning leak presence, but also provides a tool for evaluating how certain the made decision actually is.

To summarize, the aim of this article is to study the problem of leak detection to help mitigating the environmental, economic and safety consequences due to leak accidents. In particular, we develop a robust detection system based on using a multiple model approach which is capable of detecting leak under variable operational conditions. Compared with the existing results in the topic, this article has three distinct features that have not been reported in the literature. Firstly, we design a probabilistic based decision system, as opposed to the current existing deterministic systems. Secondly, the developed probabilistic based decision system falls under the umbrella of multiple models, hence leak can be efficiently detected under variable operational conditions. Thirdly, knowledge from the developed probabilistic model is used to provide quantification measures on the reliability of the leak decision, and hence decide further actions. This also is helpful when more than one system are used to detect leaks, that is, the reliability quantification measures will determine the system with the most confident decision.

## 2. Experimental Setup and Data Collection:

As the work is based on experimental data, the first step was to build an experimental pipeline. A schematic diagram of the experimental setup of the pipeline and additional equipment of transmitters and data logging system is given in figure (1). Data logging system is designed and implemented based on micro controller technology to capture the values of the variables to be studied and transmits them onto an accessible data base on a PC. The components of figure (1) are described in table (1).

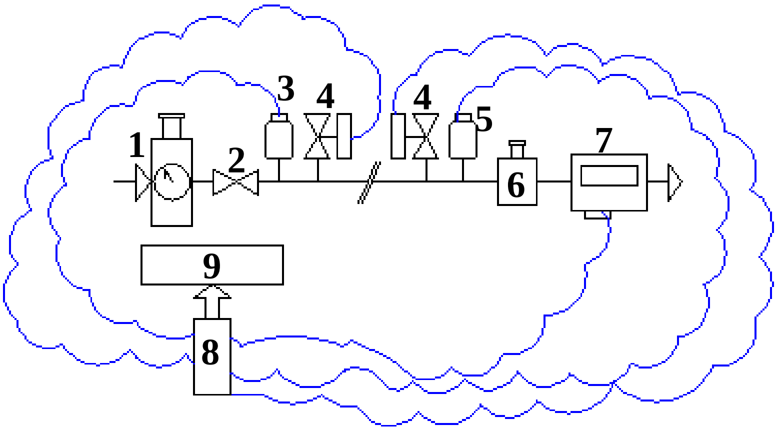


Figure (1): Schematic diagram of the experimental setup used for data collection

**Table (1): Setup Components Description**

| | | |
|---|---|---|
| 1. Pressure regulator | 2. Manual ball valve | 3. Inlet pressure transmitter |
| 4. Leak-simulating solenoid valve | 5. Outlet pressure transmitter | 6. Outlet flow control valve |
| 7. Outlet flow transmitter | 8. Data logging system | 9. Personal computer (PC) |

After the setup has been established, it was operated under multi-pressure multi-flow conditions. The purpose is to collect representative data that represents the various operating conditions of the pipeline. The pressure range for the conducted experiment is set to be 1 to 6 Bar, whereas the flow range is set to be 0 to 160 Liter Per Minute (LPM). Those ranges are forced upon the experiment due to the fact that the air compressor used in this experiment has those operational ranges. It was sufficient to collect data for the pressure starting from 1 Bar to 6 Bars with a step of 1 Bar. For each of the 6 pressure values, 5 flow rates are investigated; starting with 0 LPM and ending with 160 LPM with a step of 40 LPM. For every flow, data is collected on three cases; namely, no leak, leak at two distinct positions within the setup (locations denoted as 4 on figure (1)). For every case, about 1,000 readings are collected. This was done to capture the randomness and noise associated with the experiment due to the physical phenomenon itself, transmitters' noise, power supply noise and other inevitable noise sources.

The experiment yielded a total of 92,257 data points. Those correspond to 30 changes in the operational conditions of pressure and/or flow.

## 3. Statistical Analysis

A general realization of the collected data is established by conducting basic statistical analysis. This analysis involves the determination of the mean, range and variance for the data representing each operation condition. Table (2) gives a sample of the basic statistical analysis of data representing the 1 Bar inlet pressure.

**Table (2): Inlet pressure statistical analysis – from 1 Bar collected data**

| **Outlet flow** | $\boldsymbol{Q = 0\ LPM}$ | | $\boldsymbol{Q = 40\ LPM}$ | | $\boldsymbol{Q = 80\ LPM}$ | |
|---|---|---|---|---|---|---|
| **Leak state** | $\boldsymbol{No\ leak}$ | $\boldsymbol{Leak}$ | $\boldsymbol{No\ leak}$ | $\boldsymbol{Leak}$ | $\boldsymbol{No\ leak}$ | $\boldsymbol{Leak}$ |
| $\boldsymbol{\overline{P_{in}}}$ | **401.030** | **344.634** | **361.281** | **324.000** | **362.118** | **334.062** |
| $\boldsymbol{Range_{P_{in}}}$ | **24.000** | **27.000** | **26.000** | **26.000** | **21.000** | **25.000** |
| $\boldsymbol{S^2{}_{P_{in}}}$ | **5.121** | **4.664** | **4.201** | **5.476** | **3.641** | **5.165** |

Note that the stated numbers in table (2) correspond to the digitized values of the inlet pressure, which are converted by the Analog Digital Converter (ADC), and the whole analysis was based on these values.

As can be seen from table (2) the inlet pressure decreases in case of leak happening, also, the same behavior is noticed in case of flow increase. The analysis in table (2) was applied to all collected data of inlet pressure, outlet pressure and outlet flow yielding similar conclusions.

Prior to beginning the development of NN models, we have statistically validated that the variables of inlet pressure, outlet pressure and outlet flow can be considered as leak indicators. This was achieved by testing the hypothesis of factor significance. The aim here is to test the hypothesis that the inlet pressure, outlet pressure and outlet flow are affected by leak. As the data was collected in pairs under homogenous conditions, where two data populations corresponding to leak and no leak were recorded for a particular set of pressure and flow conditions, the appropriate test is shown to be the *paired t-test*. This test investigates the following hypothesis (Montgomery, 2009), (Montgomery & Runger, 2007), (David & Gunnink, 1997), (Hedberg & Ayers, 2015):

$$\begin{matrix} H_o: \mu_D = 0 \\ H_1: \mu_D \neq 0 \end{matrix} \dots (1)$$

The null hypothesis ($H_o$) indicates that there is no strong evidence supporting the claim that leak affects inlet pressure, outlet pressure and outlet flow. In contrast, the alternative hypothesis indicates that there is strong evidence supporting this claim.

Practically, the test follows the following steps:

I. Data values representing the same operational conditions of inlet pressure, outlet pressure, outlet flow and leak status are averaged, then the difference between the two averages of leak and no leak data is calculated. After doing that, the followings are obtained (for the inlet pressure):

$$\left\{\left(\{(P_{in.no\ leak})_{Q=40\ i}\}_{i=0}^{4}\right)_{P=j}\right\}_{j=1}^{6} \ldots (2)$$

$$\left\{\left(\{(P_{in.leak})_{Q=40\ i}\}_{i=0}^{4}\right)_{P=j}\right\}_{j=1}^{6} \ldots (3)$$

The indices ($i$ and $j$) denote the values for flow and pressure respectively. It can be seen in equations (2) and (3) that the flow ranges between from $0$ to $160$ with steps of $40$ LPM; mathematically expressed as $40i$, where $i = 0,1,2,3,4$. Also, the pressure ranges between 1 Bar to 6 Bar with steps of 1 Bar; mathematically expressed as $j$, where $j = 1,2,3,4,5, and\ 6$.

This results in $60$ average values. A sample of them is shown in table (3).

Table (3): paired $t$-test data preparation for the inlet pressure.
$P_{in}$ − Inlet pressure, $Q_{out}$ − Outlet flow
$P_{in.no\ leak,Digitized}$ − digital value for inlet pressure under no leak condition
$P_{in.leak,Digitized}$ − digital value for inlet pressure under leak condition
$d_{in,Digitized}$ − digital value for the difference between inlet pressure under no leak condition and pressure under leak condition

| $P_{in}$<br>Bar | $Q_{out}$<br>LPM | $P_{in.no\ leak}$<br>Digitized | $P_{in.leak}$<br>Digitized | $d_{p_{in}}$<br>Digitized |
|---|---|---|---|---|
| 1 | 0 | 401.0 | 344.6 | 56.4 |
| | 40 | 361.3 | 339.3 | 22.0 |
| | 80 | 362.1 | 334.1 | 28.1 |
| | 120 | 354.2 | 329.4 | 24.8 |
| | 160 | 351.1 | 324.8 | 26.3 |
| 2 | 0 | 503.0 | 461.6 | 41.4 |
| | 40 | 494.8 | 457.1 | 37.7 |
| | 80 | 494.4 | 452.8 | 41.7 |
| | 120 | 487.1 | 458.7 | 28.5 |
| | 160 | 484.0 | 461.2 | 22.8 |
| 6 | 0 | 1014.2 | 923.7 | 90.5 |
| | 40 | 1000.1 | 919.6 | 80.6 |
| | 80 | 1008.1 | 904.8 | 103.3 |
| | 120 | 1004.1 | 893.3 | 110.8 |
| | 160 | 997.2 | 893.8 | 103.4 |

II. The values of the pressure difference shown in column (5) of table (3) ($d_{P_{in}}$ ) are used to evaluate equations (4) and (5):

$$\overline{d_{P_{in}}} = \frac{1}{n}\sum_{i=1}^{n}\left(d_{P_{in}}\right)_i \ldots (4)$$

$$S_{D.P_{in}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(d_{P_{in}}\right)_i - \overline{d_{P_{in}}}\right)} \ldots (5)$$

Where:

$\overline{d_{P_{in}}}$ − Differences mean

$S_{D.P_{in}} -$ Differences standard deviation

$n -$ Number of means, which is equal to 30

Substituting the values in table (3), to get:

$\overline{d_{P_{in}}} = 53.52,\ S_{D.P_{in}} = \ 25.12$

III. The test statistic $\left(t_{o_{P_{in}}}\right)$ for the *paired t-test* is evaluated using:

$$t_{o_{P_{in}}} = \frac{\overline{d_{P_{in}}}}{S_{D.P_{in}}/\sqrt{n}} = 11.67 \ldots (6)$$

IV. The significance level of $\alpha = 5\%$ is used to get a confidence interval (CI) of 95% which is then used to determine the critical values of $\mp\ t_{\frac{\alpha}{2},n-1}$ from the z-distribution table. Those values are found to be, $\mp\ 2.756$

V. A decision is made concerning the significance of the factor in study according to:

$$Decision\text{:} \begin{cases} Fail\ to\ reject\ H_o, if\ t_{opin}\ \in\ \left(-t_{\frac{\alpha}{2},n-1}, +t_{\frac{\alpha}{2},n-1}\right) \\ Reject\ H_o, Otherwise \end{cases} \ldots (7)$$

Since $t_{opin} = 11.670\ \notin\ (-2.756, +2,756)$, then, the null hypothesis is rejected, i.e. there is evidence supporting the claim that inlet pressure is affected by leak. The same analysis was followed to test the effect of leak on outlet pressure and outlet flow; and it was found that the conclusion drawn for the inlet pressure can be generalized for the outlet pressure and flow.

## 4. Neural Network Models

The collected data are used in the development of the NN based leak detection system. Generally, the NNs are to perform classification tasks; that is, they classify incoming data of inlet pressure, outlet pressure and outlet flow onto leak or no leak states. Therefore, all of the NNs are developed with three input values corresponding to inlet pressure, outlet pressure and outlet flow. Additionally, they have two output neurons; namely, leak and no leak neurons. The output neuron with the larger amount takes over and decides the leak state. If the values on both output neurons are similar to each other, the state is decided to be "unknown".

The collected data are then pre-processed prior to being used in the developed NN models. The pre-processing includes categorization and normalization of the data. The data is categorized into 3 categories, namely: training, validation and testing. The sizes and number of operational changes in each category is shown in table (4).

**Table (4): Data categories description:**

| Category | Size | No. Operational conditions changes |
|---|---|---|
| Training | 55,634 | 18 |
| Validation | 18,311 | 6 |
| Testing | 18,312 | 6 |

The activation functions used in the NN models are of either logistic function or hyperbolic tangent function. These two functions are sensitive to inputs in the interval $[0,1]$ for the logistic, and $[-1,1]$ for the hyperbolic tangent function. Consequently, their inputs must be limited to the correspondent effective range so that saturation can be avoided. This is done by dividing all data points by their maximum value. Therefore, from this point further, when inlet pressure, outlet pressure and outlet flow are mentioned, they are referred to by their digitized normalized values.

The performance of the NN models is evaluated in terms of the classification error such that

$$e_{class.i} = \frac{n_{miss.classified_i}}{N_i} \ldots (8)$$

Where:
$e_{class.i}$ – Classification error category $i$
$n_{miss.classified} i$ – Number of miss classified data points in a category $i$
$N_i$ – Size of category $i$.

## 5. Single Model Approach

Several single model networks have been developed and inspected in (Abdulla et.al, (2013), and it was found that the best performance was obtained from a MLP model with 11 hidden neurons; yielding a validation error of 6.06%. When the performance of this model was inspected on the testing data, the error was 3.65%.

The obtained best single model network with 11 hidden neurons was then developed in a probabilistic framework and enhanced, such that it provides a quantification measure of the reliability of its predicted decision concerning the presence or absence of leak in terms of a Confidence Interval (CI) around its predicted output values. More specifically, each decision is associated with its own CI, which is established via an individual network, whose architecture is similar to the best model previously selected – 11-hidden neuron MLP. Instead of being trained to predict leak state, this network is trained to predict the variance embedded with each decision. Therefore, firstly, the variance for each decision from the training data is calculated as follows:

$$\sigma_i{}^2 = (y_i - t_i)^2 \ldots (9)$$

Where:
$\sigma_i{}^2$ – $i$th value of Variance
$y_i$, – Actual $i$th output
$t_i$ – Target $i$th output

Then, another neural network which has the same structure of the best is constructed and trained with the same training inputs but with the variance values computed in equation (9) as its targets (Herzallah et.al, 2011, Herzallah et.al, 2007). Therefore, for every input data point, two outputs are obtained through two individual networks, the first one is concerned with a decision for predicting the leak presence, and the second one is concerned with providing the predicted variance for this decision. The variance is then used as follows to construct the corresponding CI (Montgomery & Runger, 2007):

$$y_i - Z_{a/2}.\sigma_i \leq \hat{y}_i \leq y_i + Z_{a/2}.\sigma_i \ldots (10)$$

Where:
$\hat{y}_i$ – Mean predicted $i$th output, $\{0, 1\}$
$y_i$ – Actual predicted $i$th output, $[0, 1]$
$\sigma_i$ – $i$th Standard deviation
$Z_{a/2}$ – Critical value from $z$-distribution
$\alpha$ –Signifiance level

The required CI in this work is 95%, and the corresponding critical value is 1.96 from the z-distribution table.

For visual presentation of the probabilistic single model, the testing data was passed through both the leak detection network and the variance network; and the outputs were arranged in an ascending order of the leak-state output as they evaluated from the leak-output neuron as shown in figure (2). Additionally, and in order to show if the output of the network is properly predicted, each output is either marked with a green (o) or a red (x). Where the green o's stand for actual no leak, and the red x's stand for actual leak.

Two significant conclusions can be drawn from figure (2). Firstly, the range above the line of 0.5 indicates predicted leaks, whereas the range below that line indicates predicted no leaks. However, any red X's appearing beneath this line represent actual leaks falsely predicted by the system as no leak, similarly, any green O's appearing above this line represent actual no leaks falsely predicted as leaks. Secondly, as the system output approaches either 0 or 1, the CI's get narrower. This implies more confident decisions for those outputs. This can also be noticed when inspecting the constructed intervals for outputs that are far from 0 and 1.

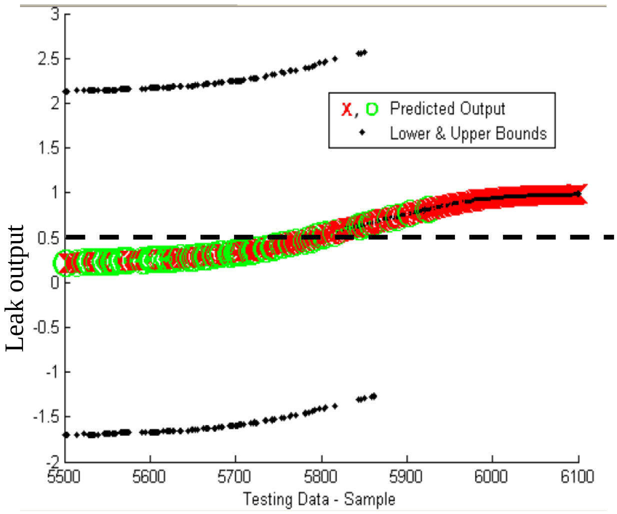


Figure (2): CI's for a sample of the testing data for the 11-hidden neuron MLP network

## 6. Multiple Model Approach

The performance of the leak detection decision support system is expected to be improved if it is designed in a multiple-model structure. That is, instead of having one network handling all incoming data points, there would be several networks, each responsible for handling a specific category. This can be interpreted as reducing the complexity of the patterns, the decision support system has to recognize. In this work, incoming data points are categorized according to the operational pressure, and then, each operational pressure category is handled by its own neural network model. In particular for the six operational pressure values; namely, 1 Bar, 2 Bar …6 Bar, we design six neural network models, each corresponds to a certain operational pressure.

For the sake of conceptualizing the overall multiple model, one shall notice the need to provide a categorizing "clustering" block that would read incoming data and pass it to the relevant processing model. Figure (3) shows a block diagram for the structure and components of the overall developed multiple model system.

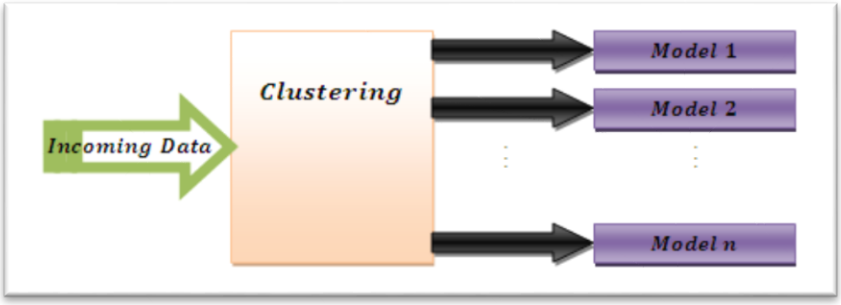


Figure (3): Multiple Model Structure

Several intelligent clustering techniques are available, in this work three of them are investigated, namely, K-means, KNN, and Probabilistic Neural Networks (PNN).

### 6.1. Clustering techniques:

As mentioned before, clustering is the block responsible for recognizing the category of the incoming data which is then forwarded to and processed by the appropriate model. In this work, there are six clusters corresponding to the operational pressure of 1 Bar, 2 Bar… ,6 Bar.

#### 6.1.1 K-means:

As an unsupervised clustering technique; K-means inspects solely input space. There are 6 means that correspond to 6 pressure clusters as mentioned before. An incoming data point is assigned to the cluster whose mean is the closest to the data point itself (Wang et.al, 2015), (Timmeman et.al, 2013). The means are optimized iteratively. They are initialized; in this work, data is 3-dimensional, and correspondingly, the means are 3-dimensional as well. The means are initialized via averaging the available data for each cluster. Then, a cost function is defined to be the Euclidean distance between input space and the means as follows (Bishop, 2009), (Kanungo et.al, 2002):

$$\mathcal{L} = \sum_{j=1}^{6}\sum_{i=1}^{n_j}\left\|x_i - \mu_j\right\|^2 \dots (11)$$

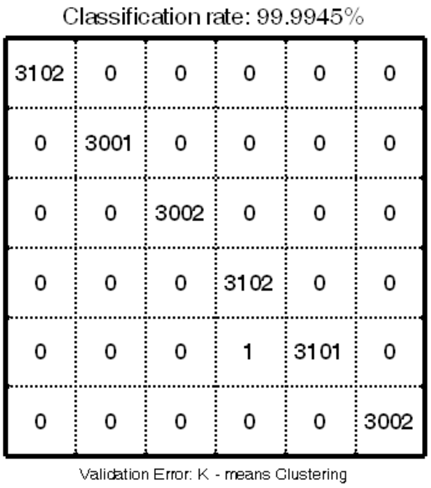
Where:
$\mathcal{L} -$ Error cost function
$x_i -$ Input data vector
$\mu_j - j^{\text{th}}$ mean
$n_j -$ Size of the $j^{\text{th}}$ cluster

Being an unconstrained non-linear optimization problem, it is iteratively solved via assigning data points to the clusters depending upon the available means at that point, and then calculating the new means as follows,

$$\mu_j = \frac{1}{n_j}\sum_{x_i \epsilon S_j} x_i \dots (12)$$

Where:
$x_i \epsilon S_j - i^{\text{th}}$ data points belonging to the $j^{\text{th}}$ cluster.

Then, data points are re-assigned in terms of the new means. This process is repeated until there is no virtual change in the means. The confusion matrix for the K-mean is shown in figure (4).



Classification rate: 99.9945%

| | | | | | |
|---|---|---|---|---|---|
| 3102 | 0 | 0 | 0 | 0 | 0 |
| 0 | 3001 | 0 | 0 | 0 | 0 |
| 0 | 0 | 3002 | 0 | 0 | 0 |
| 0 | 0 | 0 | 3102 | 0 | 0 |
| 0 | 0 | 0 | 1 | 3101 | 0 |
| 0 | 0 | 0 | 0 | 0 | 3002 |

Validation Error: K - means Clustering

Figure (4): Confusion matrix of k-mean method

It can be seen from figure (4) that among the 18,311 cases in the validation data, only 1 case was improperly clustered which gives an error of about 0.01%.

### 6.1.2 K-Nearest Neighbors (KNN):

Unlike K-means, KNN is a supervised clustering technique; where a database of data points with known clusters is stored (Bishop, 2009), (Yang et.al, 2007), (Yingguan et.al, 2002), (Dashiell et.al, 2014). Whenever a new data point is to be assigned for a certain cluster, a hyper sphere is constructed whose centre is the new data point itself. This sphere is then enlarged until it encompasses $K$ number of clustered-points among the stored database, then a majority vote is taken, and the new data point is assigned according to the voting process (Bishop, 2007). Mathematically, if the stored database is of size $N$ in which, there are $n_j$ points belonging to cluster $(j)$, and a new data point is to be clustered, a sphere of volume $V$ is constructed whose centre is the new data point. Inside this sphere there are $K$ data points with known cluster, among which there are $K_j$ data points belonging to cluster $(j)$. The cluster conditional density is given by:

$$p(\mathrm{X}|\, S_j) = \frac{K_j/n_j}{V} \ldots (13)$$

And the unconditional density is then:

$$p(\mathrm{X}) = \frac{K/N}{V} \ldots (14)$$

Equation (15) gives the prior:

$$P(S_j) = \frac{n_j}{N} \ldots (15)$$

Correspondingly, Bayes' theorem is used to give equation (16).

$$p(S_j|\, \mathrm{X}) = \frac{p(\mathrm{X}|\, S_j)P(S_j)}{p(\mathrm{X})} = \frac{K_j}{K} \ldots (16)$$

Equation (16) shows that any new data point shall be assigned to the cluster such that the ratio resulted in equation (16) is maximized. In KNN, the $K$ parameter shall be determined, which then determines the number of clustered data points that are involved in the clustering process. In this work, this parameter is determined via cross validation (Bax, 2012). That is, 12 values are investigated on the validation data starting with 1,000 till 12,000 with 1,000 step size. The results are shown in figure (5).

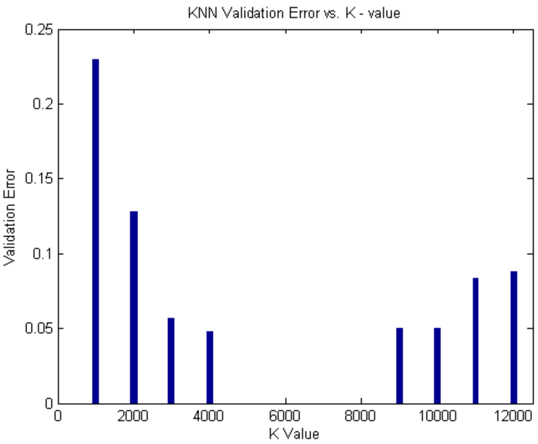


Figure (5): validation error vs. K-values of KNN clustering

All validation data points were clustered properly using *K* equal to 5000, 6000, 7000 and 8000. The selected value of K in this work is 5000, the minimum of the best values. For this value of *K*, figure (6) gives the probability of data in the validation category belonging to cluster 6 Bar.

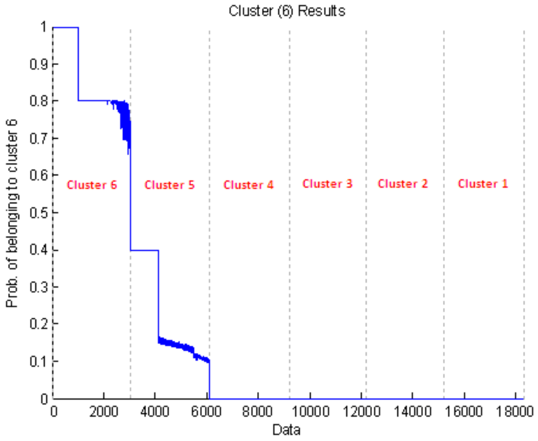


Figure (6): validation data probability of belonging to cluster 6 Bar

### 6.1.3 Probabilistic Neural Networks (PNN):

Being a supervised clustering technique; PNN requires data points with known clusters. PNN is composed of 4 layers; namely, *input*, *pattern*, *summation*, and *output layers*. A schematic diagram of the PNN is shown in figure (7) (Specht, 1990), (I-Cheng & Kuan-Cheng (2011).

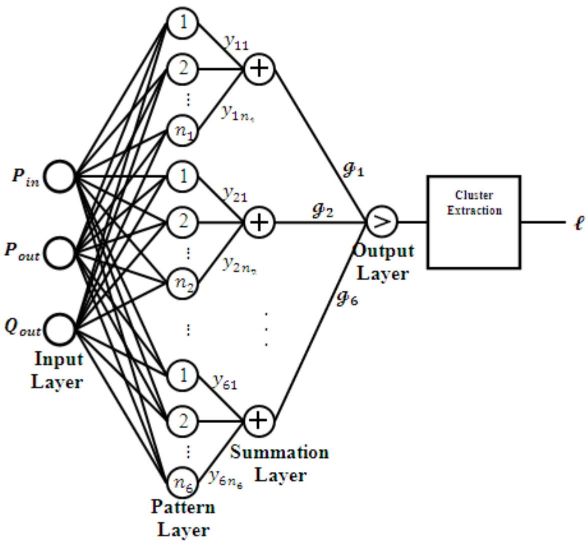


Figure (7): PNN Schematic diagram

As shown in figure (7), the input layer consists of 3 units, namely, inlet pressure, outlet pressure and outlet flow, these are fully connected to the units in the pattern layer. Additionally, the pattern layer consists of 6 blocks that correspond to 6 clusters; where each block consists of processing units (neurons) whose size is similar to the size of the portion of the training data points belonging to that cluster. In each neuron within the pattern layer, there exists a Gaussian kernel to process the summed input. The output of the neurons within the same block are summed and averaged in the related neuron in the summation layer, after which, the *cluster extraction block* inspects the 6 results and determines the cluster according to the larger output. This is mathematically illustrated through the following equations:

$$y_{jk} = \exp\left\{-\left(\frac{\left(P_{in}-P_{in_{jk}}\right)^2+\left(P_{out}-P_{out_{jk}}\right)^2+\left(Q_{out}-Q_{out_{jk}}\right)^2}{2\sigma^2}\right)\right\}\ldots.\ (17)$$

Where:
$y_{jk}$ – Output from neuron $k$ of block $j, j = 1,2 \ldots 6, k = 1,2 \ldots n_j$
$n_j$ – Size of training data points for cluster $j$
$P_{in}, P_{out}, Q_{out}$ – Input vector elements to be clustered
$P_{in_{jk}}, P_{out_{jk}}, Q_{out_{jk}}$ – Training data point $k$ of cluster $j, k = 1,2 \ldots n_j, j = 1,2 \ldots 6$

$$g_j = \frac{1}{n_j}\sum_{k=1}^{n_j} y_{jk} \ldots (18)$$

Where:
$g_j$ – Output from the $jth$ neuron in the summation node, that corresponds to cluster $j, j = 1,2, \ldots 6$

$$g = \max\{g_1, g_2, g_3, g_4, g_5, g_6\} \ldots (19)$$

Where:
$g$ – Result of output layer

$$\ell = \begin{cases} 1, g = g_1 \\ 2, g = g_2 \\ 3, g = g_3 \\ 4, g = g_4 \\ 5, g = g_5 \\ 6, g = g_6 \end{cases} \ldots (20)$$

Where:
$\ell$ – Cluster of the input data point

The parameter $\sigma$ in equation (18) refers to the spread of the Gaussian kernel, and it is selected via cross validation. Several values for this parameter have been inspected as shown in figure (8). Zero validation error was obtained for $\sigma = 0.15$.

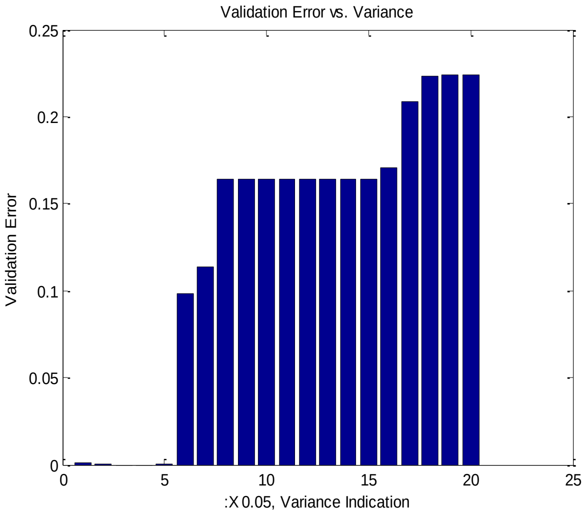


Figure (8): Validation error vs. values of spread parameter

### 6.1.4 Best Clustering Technique:

As expected, the supervised clustering – KNN and PNN, resulted in lower classification error with respect to the unsupervised K-means technique. Since the number of data point required for training the KNN was significantly less than that for the PNN (5000 in the KNN as opposed to 55000 in the PNN) we decided to use the KNN clustering technique with $K = 5000$ for further analysis and development.

## 6.2. Overall Deterministic Multiple Model System:

Following the clustering of the data into 6 clusters, six models as shown in figure (9) are developed to handle the data belonging to each of the six clusters; namely clusters: 1 Bar, 2 Bar…to 6 Bar. Here it is found sufficient to use Generalized Linear Model (GLM) as a classification model for each cluster.

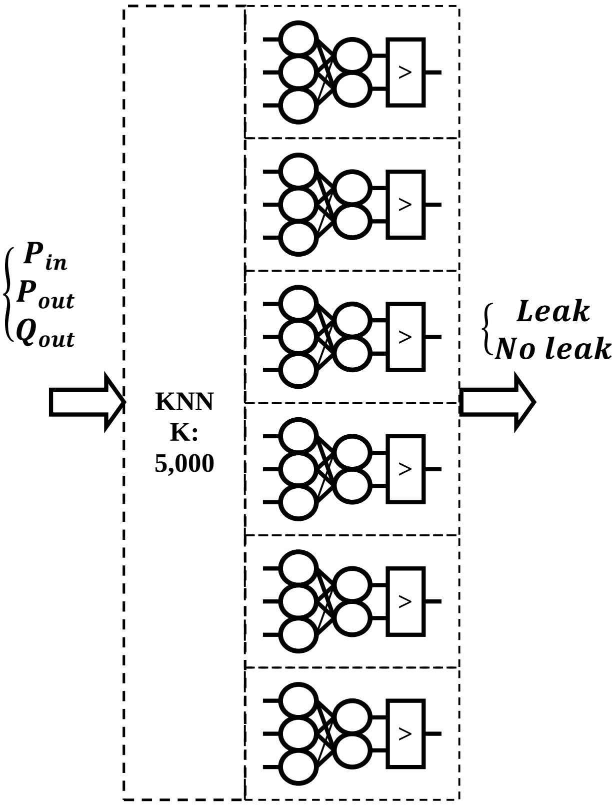

$\begin{cases} P_{in} \\ P_{out} \\ Q_{out} \end{cases}$
KNN
K:
5,000
>
>
>
>
>
>
$\begin{cases} Leak \\ No\ leak \end{cases}$

Figure (9): Overall deterministic multiple model system – schematic diagram

This system is firstly trained on the training data where zero training error was achieved and validated on validation data resulting in zero validation error as well. This indicates the successful application of the multiple GLM for leak detection under different operating conditions. As can be noted, the validation error as a result of the best single model structure (11 hidden neuron MLP model) was reduced from 6.06% to 0% when the proposed clustering and multiple models method is used.

For investigating the performance of the overall system, the testing data was passed through the overall system; consequently, the testing error was found to be 1.01%.Compared to the 11-hidden neuron MLP NN, the testing error was reduced from 3.5% to 1.1%.

Similar to the plotting procedure in figure (2), the outputs from the multiple model system are plotted and shown in figure (10).

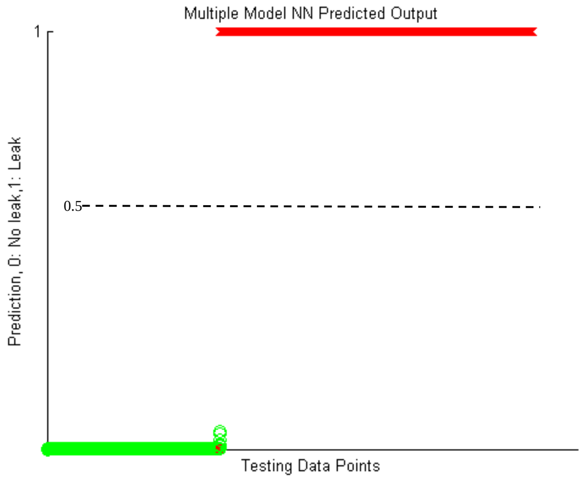


Figure (10): Multiple model output from leak output neuron

Actual leak points are presented with red x's, whereas actual no leak points are presented with green o's. The abrupt transition between 1 and 0 indicates more accurate performance compared to what the single model achieved in figure (2).

### 6.3. Overall Probabilistic Multiple Model System:

The analysis followed in the probabilistic single model approach was similarly followed here, however, here it is repeated 6 times. This means that, for every GLM NN in figure (9), an additional GLM was developed to predict the expected variance associated with the leak decision as given in figure (11). The CIs for a portion of the testing data are shown in figure (12). It can be clearly seen that the performance of the multiple model approach is superior to the single model approach not only in terms of high classification rate, but also in terms of narrower CI, indicating more precise and confident decisions.

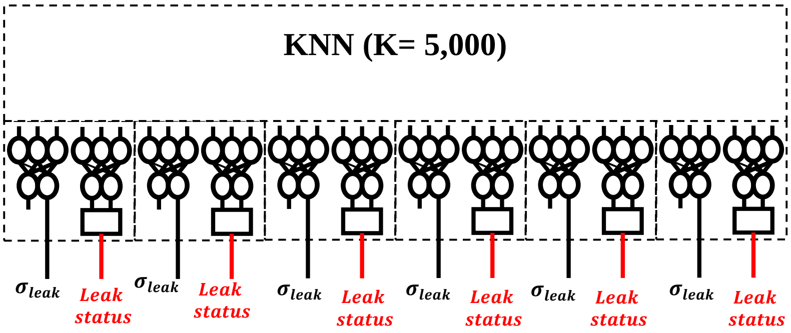


Figure (11): Overall probabilistic Multiple Model System

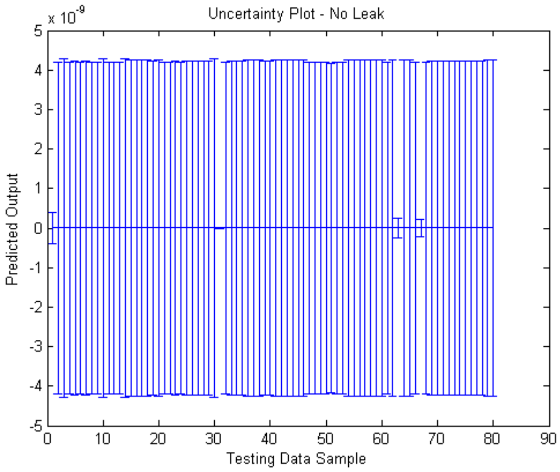


Figure (12): CI’s for a sample of the testing data for the multiple model system

## 7. Conclusions

Noise immunity provide the NN-based with a crucial advantage for overcoming inevitable sources of noise in practice, such as Electromagnetic interference, and ADC related noise. Also, the variations in the operational conditions significantly affect the performance of the NN-based system as it leads to more complicated patterns. This was made explicit in the validation error reduction from 6.06% to 0%, and testing error from 3.5% to 1.1% when the performances of the single model and multiple model systems are compared. Also the corresponding confidence intervals were narrower as demonstrated in figure (12).

The probabilistic nature of the developed multiple model NN leak detection system gives the decision makers the benefit of analyzing risks associated with their decisions, and correspondingly take more cautious decisions. Also, the probabilistic nature provides a frame work for hybrid decision making algorithm. Such systems encompasses several leak detection systems that collectively make decisions; where the system with the higher confidence possesses more weight in the overall decision. This is expected to reduce the false alarm rates even further, and therefore reduce the unnecessary shut down events, which leads to a reduction in the associated financial loss.

## 8. References

H. Prashanth, Shankar Narasimhan, S. Murty Bhallamudi, S. Bairagi. (2011). Leak Detection in Gas Pipeline Networks Using Efficient State Estimator. Part I: Theory and Simulation. International Journal of computer application in chemical engineering, 34 (4), 651 – 661

Abdulla M.B., Herzallah R.O., Hammad M.A. (2013). Pipeline Leak Detection Using Artificial Neural Network: Experimental Study. Proceedings of the international conference on Modelling, Identification and Control, Cairo, Egypt, August 21 – September 2

Belsito S., Lombari P., Andreussi P., Banerjee S. (2004). Leak Detection in Liquefied Gas Pipeline by Artificial Neural Networks. AiChE journal, 44 (12)

Jian L., Li-kin W., Yan Z., Shi-jiu J. (2003), Study on Detection Technique for Pipeline Leakage, transaction of Tianjin University 9 (2), China, 112 – 114

Jian F, Huaguang Z. (2004). Oil Pipeline Leak Detection and Location Using Double Sensors Pressure Gradient Method. Proceeding of the fifth international world congress on intelligent control and automation 4, 3134 – 3137

Jinhai L., Huaguang Z., Jian F., Heng Y. (2007). A New Fault Detection and Diagnosis Method for Oil Pipeline Based on Rough Set and Neural Networks. Proceeding of the fourth international symposium on neural networks, Nanjing, China, June 3 – 7

Bicharra A., Ferra I., Bermardini F. (2008). Artificial Neural Networks Ensemble Used for Pipeline Leak Detection Systems. Proceeding of the seventh international pipeline conference, Alberta, Canada, Sept. 29 – Oct. 3

Khan A., Widdop P., Day A., Wood A., Mounce R., Machell J. (2008). Artificial Neural Network Model for Low Cost Failure Sensor: Performance assessment in pipeline distribution. Proceeding of the world academy of science, engineering and technology 2 (9), 44 – 50

Barradas I., Garza L., Menendez R., Martinez V. (2009). Leaks Detection in Pipeline Using Artificial Neural Networks, proceeding of the Iberoamercian congress on patter recognition conference, Guadalajara, Jalisco, Maxico, Nov. 15 – 18
Wei Liang, Laibin Zhang. (2012). A Wave Change Analysis (WCA) Method for Pipeline Leak Detection Using Gaussian Mixture Model. Journal of loss prevention in the process industries 25, 60 – 69
Corneliu T.C. Arsene, Bogdan Gabrys, David Al-Dabass. (2012). Decision Support System for Water Distribution Systems Based on Neural Networks and Graphs Theory for Leakage Detection. journal of expert systems with applications 39 (18), 13214 – 13224
Mahmoud Meribout. (2011). A Wireless Sensor Network-Based Infrastructure for Real-Time and Online Pipeline Inspection. IEEE sensors journal 11 (11), 2966 – 2972
Seung Ihl Kam. (2010). Mechanistic Modeling of Pipeline Leak Detection at Fixed Inlet Rate. journal of petroleum science and engineering 70 (3 – 4), 145 – 156
Jinqiu Hu, Laibin Zhang, Wei Liang. (2011). Detection of Small Leakage from Long Transportation Pipeline with Complex Noise. journal of loss prevention in the process industries 24 (4), 449 – 457
Montgomery D. (2009), Design and Analysis of Experiments, John Wiley and Sons Inc
Montgomery D., Runger G. (2007), Applied Statistics and Probability for Engineers, John Wiley and Sons Inc.
H. A. David, Jason L. Gunnink. (1997). The Paired t Test Under Artificial Pairing. Journal of the American statistician 51 (1), 9 – 12
Hedberg E. C. Ayers Stephanie. (2015). The Power of a Paired t-test with Covariance. Journal of social science research 50, 277 – 291
R. Herzallah and Miroslav Karny. (2011). Fully probabilistic control design in an adaptive critic framework. Neural Networks, 24. doi:10.1016/j.neunet.2011.06.006
R. Herzallah and D. Lowe. (2007). Distribution modelling of nonlinear inverse controller under a Bayesian framework. IEEE Transactions on Neural Networks, 18 (1), 107–114
Wang Jianfeng, Wang Jingdong, Song Jingkuan, Xu Xin-Shun, Shen Heng Tao, Li Shipeng. (2015). Optimized Cartesian K – Means. IEEE transaction on knowledge and data engineering 27 (1), 180 – 192
Marieke E. Timmeman, Eva Ceulemans, Kim De Roover, Karla Van Leeueven. (2013). Subsurface K means Clustering. Journal of behavior research methods 45 (1)
Bishop C. (2009). Pattern Recognition and Machine Learning, Springer Science and Business Media, USA
Kanungo T., Mount D., Netanyahu N., Piatko C., Silverman R. (2002). An efficient K-Means Clustering Algorithm: Analysis and Implementation. IEEE transaction on pattern analysis and machine intelligence 24 (7), 881 - 892
Yang S., Huang J., Zhou D., Zha H., Giles C. (2007). IKNN: Informative K-Nearest Neighbor Pattern Classification, proceeding of the eleventh European conference on principles and practice of knowledge discovery in databases, Warsaw, Poland, Sept. 17 – 21
Wu Yingguan, Lanakiev Krassimir, Govindaraju. (2002). Improved K-nearest neighbor Classification, journal of pattern recognition society 35 (10), 2311 – 2318
Kolbe Dashiell, Zhu Qiang, Pramanik Sakti. (2014). K-Nearest Neighbor Searching in Hybrid Spaces, journal of information system 43, 55 – 64
Bax E. (2012). Validation of K-Nearest Neighbor Classifier. IEEE transaction on information technology 58 (5), 3225 – 3234
Specht D. (1990). Probabilistic Neural Networks. Journal of Neural networks 3 (1), 109 – 118
Yeh I-Cheng, Lin Kuan-Cheng. (2011). Supervised Learning Probabilistic Neural Networks. Journal of neural processing letters 34 (2), 193 – 208