Global Flow Measurement Workshop
25 - 27 October 2022

Technical Paper

Detecting the Presence of Unwanted Components in a Single-Phase Flow Using Data-Driven Models

Dr Yanfeng Liang, TÜV SÜD National Engineering Laboratory
Dr Gordon Lindsay, Glasgow Caledonian University

1 INTRODUCTION

Flow meters such as ultrasonic flow meters (UFMs) and Coriolis mass flow meters are capable of outputting a large number of digital values from their electronic transmitter via fieldbus networks, which can provide information pertaining to meter health and process conditions. Device measurement error can manifest as drifts within these digital values, which are commonly referred to within industry as "diagnostics". However, different process conditions can evoke the same diagnostic value drift and therefore create ambiguity for an end-user who has been tasked with interpreting the data. A further challenge in interpretting digital data relates to the monitoring of two-phase oil and gas flow. UFM and Coriolis flow meters are predominately single phase flow meters, so problem and confusion occur when multiphase flow is present. This can happen, for example when the meters are installed in the wrong place or process conditions change such that they flash upstream. The extent of impact from multiphase flow on the measurement accuracy of flow meters is dependent on many factors such as meter operating principles, manufacturer specific variations within a metering technology category, multiphase flow conditions, and fluid properties [1][2]. As a result, modelling work involving multiphase flow is generally more complex than single-phase flow due to its multi-dimensional characteristics, where a small change in the content of one phase could alter the drift patterns and correlations in multiple parameters.

In a world where data is now output at high speeds and stored in large volumes, it is becoming ever more apparent that in order to maximise the valuable diagnostic information stored within the data, we would stand to benefit from the use of advanced modelling techniques, for example, machine learning models. Previous studies (e.g., [3][4][5][6][7][8]), using historical experimental data gathered at the UK's national standard for flow measurement (TÜV SÜD National Engineering Laboratory) have demonstrated the potential in using predictive machine learning models to overcome the above challenges, where these models can be used to classify different process conditions, based entirely on the correlations and patterns within diagnostic values. Errors such as improper installation, deposition, and the presence of a second phase can be detected using models such as condition-based monitoring which increases an operator's efficiency in the fault diagnosis process. However, while there are inherent advantages in simply owning high volumes of digital data there are further challenges to address before meaningful intelligence can be obtained from it. One such challenge is processing and analysing high dimensionality data. High dimensionality data, in statistics, refers to data sets that contain many variables, potentially leading to increased computational time for modelling as well as making it more challenging for the model to dissect layers of potentially noisy and irrelevant information and extract the most important data. This can potentially affect the accuracy of our predictive models as well as making the results less interpretable.

The results of two case studies are discussed in this paper, through which we will demonstrate the potential in using machine learning models to analyse the complex dynamics of multiphase flow and detect the presence of unwanted phases in a single phase flow. In addition, a dimensionality reduction technique which was used in [9] to overcome the challenge from high dimensionality data is discussed. This technique provided a viable method in which we were able to enhance end-users' understanding of the role of diagnostic variables output from specific flow meters.

This paper is organised as follows: in Section 2, we will discuss and compare the two main categories of machine learning models with their advantages and disadvantages. In Section 3, we focus on one of the case studies carried out on data obtained from UFMs, where supervised and unsupervised machine learning models were used to detect and quantify the concentration of an unwanted second phase (gas) in a single-phase water flow. The idea of using a dimensionality reduction technique in tackling high dimensionality data as well as its associating benefits will also be discussed in this section. In Section 4, we will expand upon our findings from Section 3 by applying supervised machine learning models to detect the presence of secondary and tertiary phases in a Coriolis meter. Lastly, we will summarise our results and key findings in Section 5. The prediction results in this paper were produced in R[1].

2 SUPERVISED MACHINE LEARNING MODELS VS. UNSUPERVISED MACHINE LEARNING MODELS

A machine learning model is a data analytics algorithm that teaches computers to perform tasks by learning from experience. Effectively there are two categories, 'supervised learning' or 'unsupervised learning', each having its advantages and disadvantages. The deciding factor on which model to use is highly dependent on the type of data and the question that is to be addressed during data analysis. A more detailed comparison between these models is given in Table 1.

In this paper, classification models are used to investigate the performance of a UFM and a Coriolis meter when exposed to different volume fractions of secondary or tertiary phases. A classification model can either be supervised or unsupervised and aims to segregate data into groups, however a supervised model requires the end-users to know what conditions the data represent so the model can "learn" from the data and then make predictions on another set of unseen data. Hence, the model is supervised and guided through the learning process.

[1] R is a well-known statistical programming language widely used by statisticians and data scientists to perform data analysis and modelling. It contains a wide variety of statistical tools and graphical techniques. (https://www.r-project.org/)

Table 1 - Supervised versus Unsupervised Learning Models

Unsupervised Machine Learning Model	Supervised Machine Learning Model
Unlabelled data (more common to find in practice). For example a list of values from a sensor that we do not understand the conditions in which it was collected.	Labelled data (less common to find in practice). For example a list of values from a sensor where we understand and know the conditions in which it was collected (e.g. the sensor values equate to a particular test condition).
Has input variables but no response (output) variables. Do not have predefined conditions or classes. For example, there is no information on the operating/process condition of the data, and no information on the target objectives.	Has input variables **and** output variables. Has predefined conditions or classes. For example, information on the operating/process condition of the data is available, interested error states are logged and contains information on target objectives.
Separate data into groups based on how similar or different the data points are. Can identify any underlying pattern. Only use input variables.	Learn the pattern and correlations using training data. Identify connections between input variables and output variables. Variable selections are available.
More suitable for real time monitoring due to its unsupervised nature.	Usually takes place offline.
Model is more complex with high computational cost.	Model is simpler with low computational cost.
Less interpretable and less accurate due to no response variables.	Very interpretable, reliable and accurate.
Example of unsupervised machine learning model: clustering model, gaussian mixture model-based model and anomaly detection model.	*Example of supervised machine learning model: tree-based model, neural network, regression model and support vector machine.*

On the other hand, an unsupervised model does not need such information and is able to segregate the data into *n* number of groups based on how similar and different the data points are from each other. The model determines which group each data point belongs to without guidance from end-users, but whilst the model is able to categorise, user intervention will still be required during post-processing to establish the nature of those categories.

To better understand and compare the type of outputs that can be extracted from supervised and unsupervised machine learning models, we will look at two case studies.

3 CASE STUDY 1 - DETECTING THE PRESENCE OF AN UNWANTED SECONDARY PHASE IN A SINGLE PHASE FLOW FOR A UFM

In order to investigate the effect of two-phase flow, different percentages of gas were injected into TÜV SÜD National Engineering Laboratory's single phase water facility. The concentration of gas present within a fluid will affect the performance of UFMs (due to scattering of the signal) and the degree of impact will vary depending on the concentration. Consequently, the gas injection tests carried out had different gas volume fractions (GVFs) ranging from 0.1 % GVF to 10 % GVF. Note that the measured GVFs had an average relative uncertainity value of ± 2 %, where the measured GVFs were then rounded to their nearest whole percentage for the purpose of modelling.

In this paper, a supervised learning model (random forest model[2] [10]) and an unsupervised learning model (clustering model[3] [11]) were used to analyse the two-phase data gathered from one particular UFM, which we will refer to as Meter A. Both types of model aim to segregate the data points into relevant groups based on the drifts seen in variables. In this case, we are interested in using both models in distinguishing the percentage of gas present within the water, based entirely on the relationship between different variables. From an end-user perspective, having the ability to predict the percentage of gas present within the water and thus the degree of effect on the performance of a UFM can aid in decision-making and maintenance processes. A dimensional reduction method such as Principal Component Analysis (PCA) was also used to reduce the number of variables needed in order to simplify the data analysing process.

Recall from Section 2 that the decision on which model to use will depend on the types of data available. If we have unlabelled data (no predefined classes or conditions), where we only have input variables but do not have information on the response variable, then an unsupervised learning model would be used, where the data points will be segregated into subgroups based on the similarity between input variables. If more than one subgroup exists, then it is indicative to end-users that the data set does not represent the same condition. This approach is commonly used to find meaningful structure in data and perform data exploration, where algorithms such as clustering can automatically recognise patterns without labels. This type of analysis is known as pattern recognition.

On the other hand, if we have a labelled data where end-users are aware of the types of conditions, then a supervised model can be used to predict which condition a data point belongs to.

[2] A random forest model is an ensemble classification algorithm which consists of many decision trees with "tree branches". At each tree branch, there exists a criterion which will separate data onto different tree branches with other criteria. The data will continue to split and go through different tree branches until it reaches its final prediction outcome where the tree branches cannot split anymore.

[3] A clustering model differentiates data into groups based on how similar or different each data point is. It is an algorithm which will automatically detect trends, correlations and patterns in data and thus grouping data points which are similar together in one cluster and segregating data points which behave differently into different clusters.

To compare the different levels of insights of a supervised model versus an unsupervised model, for this case study, both models were used in an attempt to predict the presence of a second-phase in a single phase flow.

3.1 Supervised Classification Model

The raw unprocessed data from the two-phase test on Meter A consisted of 13,055 observations with 55 input variables and 10 different gas levels ranging from 0 % to 10 % of GVF with an average relative uncertainty of ± 2 %. Recall that these GVFs were rounded percentages for the purpose of modelling. Prior to using the raw data to construct the machine learning models in R, the data was cleaned where missing values[4] were removed from the data set. The final cleaned processed data consisted of 12,286 observations. The processed data were then handled as demonstrated in Fig. 1.

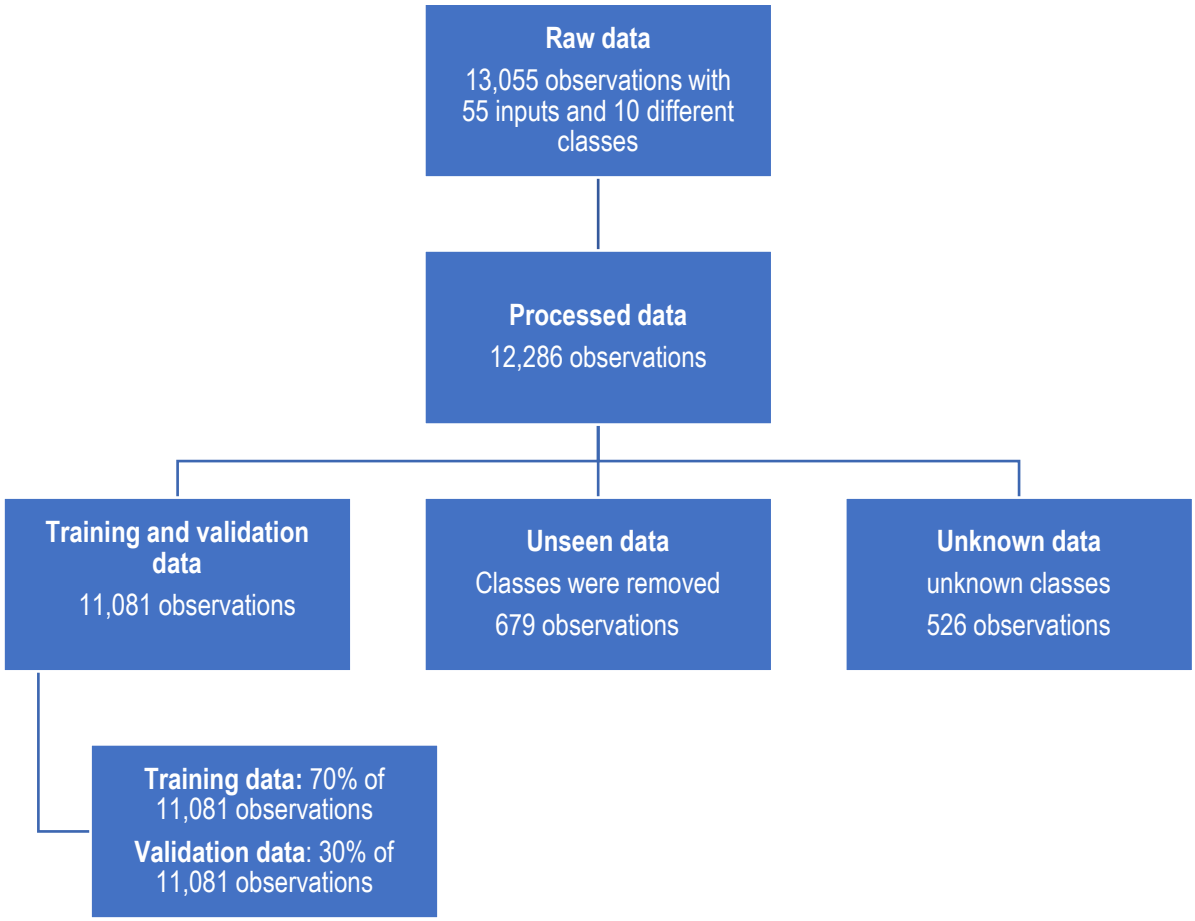
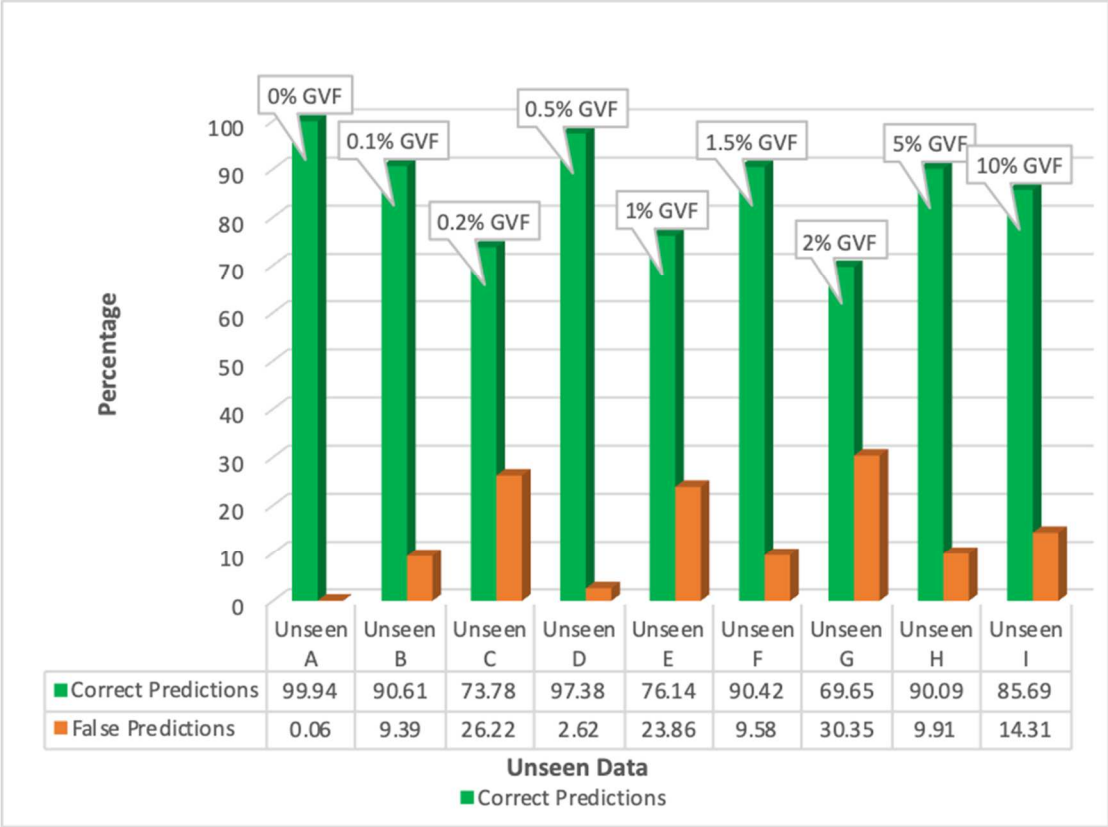


Fig. 1 – Data Structure for Case Study 1

[4] The term "missing values" in statistics refer to no data value stored for a particular variable in an observation. The presence of missing values is common in practice. Therefore, care should be taken when dealing with missing values as they can have a significant impact on the conclusion made from your data. Some models cannot handle missing values and thus sometimes it is compulsory to remove missing values prior to modelling.

For supervised classification models, it is a common practice to divide the sample data into: training data; where the model will learn the patterns and links between variables; and validation data, where the model's prediction ability will then be tested. Subsequently, unseen data is used to test the model's generalisation.

The prediction results obtained on the training and validation data from using a supervised classification model achieved an average accuracy of 98.62 % in assigning 11,081 observations correctly into the right gas classes. The model's prediction ability was further tested using unseen data. The results are shown in Fig. 2, where green bars represent correct predictions and orange bars represent false predictions.



	Unseen A	Unseen B	Unseen C	Unseen D	Unseen E	Unseen F	Unseen G	Unseen H	Unseen I
Correct Predictions	99.94	90.61	73.78	97.38	76.14	90.42	69.65	90.09	85.69
False Predictions	0.06	9.39	26.22	2.62	23.86	9.58	30.35	9.91	14.31

Fig. 2 – Prediction Results on Unseen Data from UFM Meter A Using Supervised Classification Model

For example, for unseen data D, the model predicted 97.38 % of those data correctly to represent the condition where the fluid had 0.5 % GVF, where it predicted falsely 2.62 % of those to belong to other GVF groups. Similar interpretations can be made on other unseen data. It is promising to see that the supervised classification model had classified data in different GVF classes with high accuracy by finding hidden patterns and correlations associating between variables. Results such as these would be beneficial to end-users who wish to identify how much gas is present within the fluid based on drifts experienced in certain variables.

From Fig. 1, it can be seen that there were 526 unknown observations where information on the classes of those data was not logged. In other words, we do not know what operating condition those data belong to. These unknown

observations were from two groups: Group A had 250 observations and Group B had 276 observations. Therefore, it would be interesting to use the trained supervised model built previously to predict what is the most likely GVF classes those data belong to. The prediction results for Group A and Group B are shown in Fig. 3, where, based on the drifts seen in the digital process variables, the model predicted Group A to represent having 0.10 % GVF with a mean probability of 0.9084. Similarly, Group B was predicted to represent having 0.50 % GVF with a mean probability of 0.8763. The associated probability provides additional certainty to end-users by indicating how confident the model was in predicting the most likely condition the data was collected in.

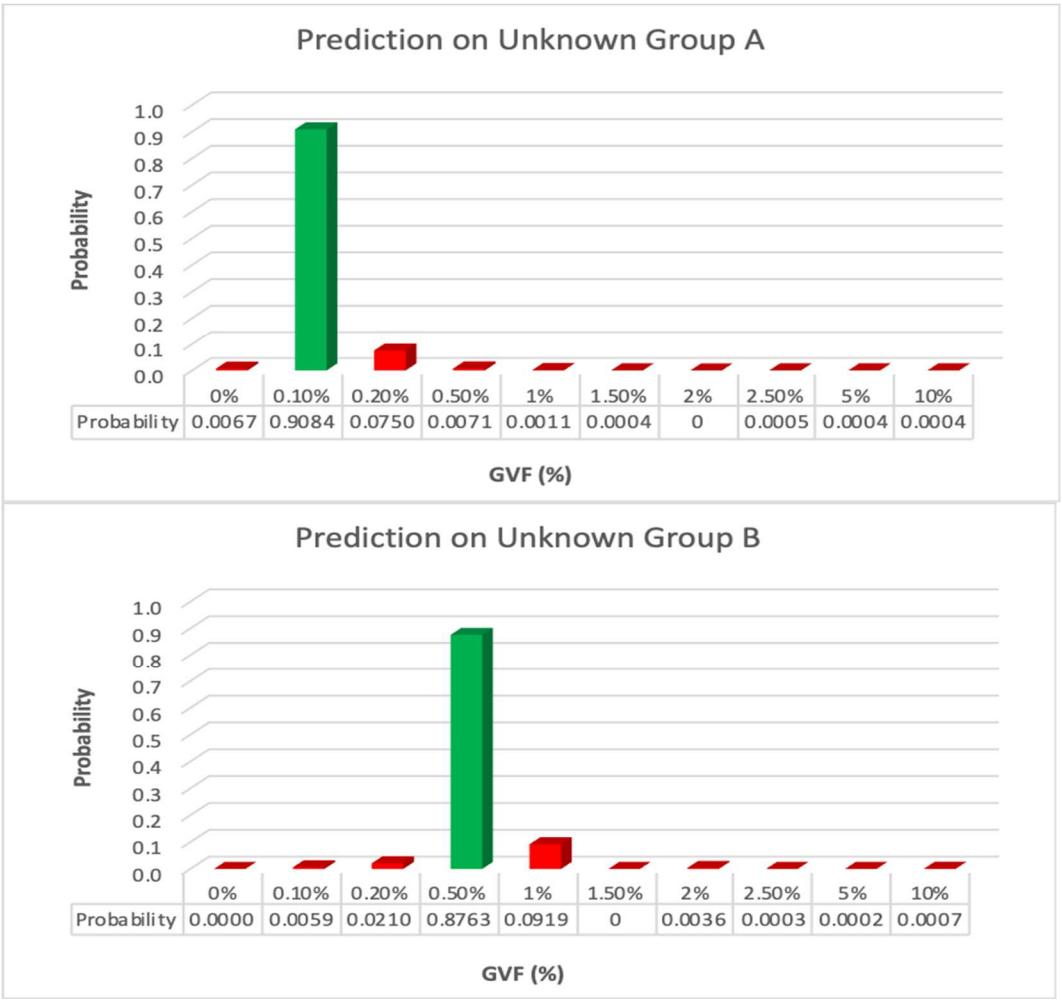


Fig. 3 – Prediction Results on Two Groups of Unknown Data from UFM Meter A Using Trained Supervised Classification Model

The results shown in this section demonstrated the capability and potential of using a supervised machine learning model to detect the presence of an unwanted second phase in a UFM as well as the level of insights that can be obtained. In this case study, by building a machine learning model using labelled GVFs data, not only was the model capable of distinguishing between single-phase flow and two-phase flow, it also had the capability of quantifying the most likely percentage of GVFs based

entirely on the patterns and correlations observed in data. These additional insights will improve end-users' fault diagnosis and rectification.

3.2 Unsupervised Classification Model - Clustering Analysis

In Section 3.1, a supervised classification model was used to sort data into the appropriate conditions based on the learnt correlations and trends observed between variables; as expected, high accuracy predictions were made as a result. However, it is more common in practice to have unlabelled data where the response variable is not necessarily known. This can sometimes be due to a knowledge gap in a particular process or site installation where domain expertise is limited or indeed due to the fact that a given site may have experienced practioners with extensive domain knowledge, however patterns in the data are being detected that do not align with the previous experiences or expectations of the site operators. So in other words, the questions that are to be asked of the model are what conditions do the data represent and how many conditions do we have? By using unsupervised classification models to better explore and investigate the types of information our data hold, these questions can be answered. In statistics, this procedure is known as exploratory data analysis (EDA).

In this section, an unsupervised classification model known as the centroid-based clustering was used to segregate our data into different clusters based on their similarity and difference to identify how many possible conditions are present within one data set. Our prior knowledge of this data already tells us that the data set should contain at least 2 clusters representing single-phase data (water only) and two-phase data (gas and water). This type of modelling technique will be beneficial to end-users as a form of anomaly monitoring, where the presence of a second group could indicate the presence of an error. Note that the data set used in this section is the same data set used in Section 3.1, but now with the predefined classes column removed to represent unlabelled data.

Centroid-based clustering is an iterative clustering process where k number of starting points are selected at random and act as the centre of each cluster. The data is then assigned to their nearest centre and form into *n* number of clusters. A new centre will get selected again and the process repeats itself until the algorithm converges. The results of centroid-based clustering are therefore highly dependent on the selected initial starting points. Therefore, it is best to run this process multiple times with different starting conditions. In this case, our model had been run with 50 randomly generated starting points.

Prior to feeding the processed data into the clustering algorithm built in R, the predefined classes column was removed from the data set to mimic an unlabelled data. The processed data was then normalised to have a mean of 0 and a standard deviation of 1.

Before going through the clustering process, Hopskin statistic *(H)* can be calculated to quickly indicate whether our data belongs to a uniform distribution. In other words, does the data contain any meaningful clusters? If *H* is less than the critical threshold value of 0.5, then we can conclude that the data does contain a meaningful cluster and therefore the clustering method can be used to extract significant information. However, if *H* is greater than 0.5, then clustering analysis is not an appropriate technique. For this data set, H=0.029 indicates that the data contains meaningful clusters and thus clustering analysis can be meaningfully used, as expected.

In centroid-based clustering, an initial guess of the number of clusters needs to be stated in the algorithm. As we are dealing with an unlabelled data, this information is not known. Therefore, the well-known silhouette method was used in R to determine the optimal number of clusters which will result in the highest average silhouette width. The average silhouette width measured the average within-cluster distances, *φ(i)*, as well as the average between-cluster distances, *θ(i)*. In other words, how close are the points within the same group and how far apart are the points from different groups. A high average silhouette width, with a small within-cluster distance and a high between-cluster distance would indicate a good cluster. The silhouette width *s(i)* can be calculated using the following formula:

$$s(i) = \begin{cases} \dfrac{\varphi(i) - \theta(i)}{\max\{\theta(i), \varphi(i)\}}, & if\ |C_i| > 1 \\ 0, & if\ |C_i| = 1. \end{cases} \quad (1)$$

For any data point $i \in C_i$ *and* $j \in C_i$, the within-cluster distance is defined as:

$$\varphi(i) = \frac{1}{|C_i| - 1} \sum_{i,j \in C_i, i \neq j} d(i,j), \quad (2)$$

where *d(i,j)* represents the distance between data point *i* and data point *j* in the same cluster C_i.

Similarly, for any data point $i \in C_i$ and $j \in C_k$, in other words, data points from different clusters, the between-cluster distance is defined as:

$$\theta(i) = min_{k \neq i} \frac{1}{|C_k|} \sum_{i \in C_i, j \in C_k} d(i,j), \quad (3)$$

where *θ(i)* is considered as the dissimilarity between data point *i* and its neighbour cluster of which *i* does not belong. If *s(i)* is close to 1, it indicates the observations are well clustered, whereas if *s(i)* is close to 0, then observations are overlapping two clusters. If *s(i)* is less than 0, then it is indicating that observations are being placed in the wrong clusters.

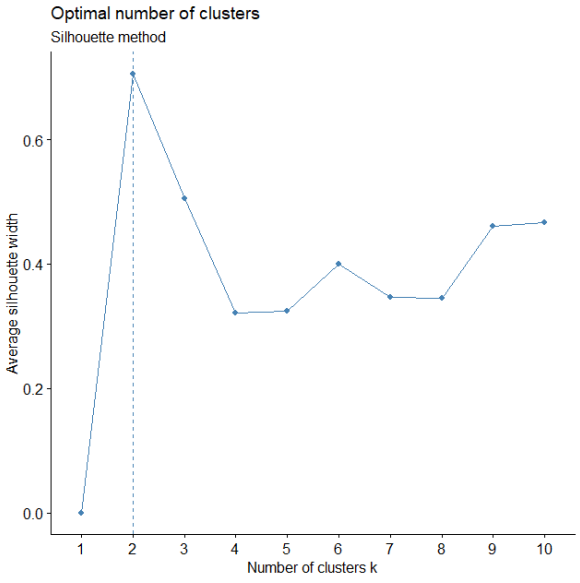


Fig. 4 – Optimal Numbers of Cluster

From Fig. 4, it was clear that the optimal number of clusters should be 2 (with an average silhouette score of 0.71), followed by 3 clusters and then 10 clusters ranked by the average silhouette width. Therefore, it is clear that the unsupervised model managed to segregate between single-phase data and multi-phase data. However, as expected, it was challenging for the model to segregate between different percentages of gas, thus indicating potential data overlapping in classes.

Under the assumption that this is an unsupervised classification model, with no information on the real number of clusters, we used the estimated optimal number suggested and continue with the clustering method. The cluster plot is shown in Fig. 5 where the data has been transformed into a two-dimension space for easy interpretation, an approach known as dimensionality reduction.

Fig. 5 – Cluster Plot for Centroid-Based Clustering

The advantages of this approach, and the insights that can be extracted from using dimensionality reduction, are discussed in more detail in Section 3.2. For now, in this section, we focus on the results obtained from using clustering algorithms. Note that the percentages shown on the x and y-axis of Fig. 5 represent how much information was explained based on the first dimension and second dimension. In other words, the variables used to construct the first dimension illustrated as the x-axis on Fig. 5 explained 46.2 % of the variances shown within the original data, while the second dimension, given as the y-axis, explained 13.2 % of the original data. More information with regards to the formation of the new dimension space is given in Section 3.2. In summary, the percentages shown on the x and y-axis of Fig. 5 described 59.4 % of the total data variance.

Although the unsupervised model had successfully segregated the data into two clusters, no information was provided to indicate the nature/condition of these

clusters. Furthermore, the model was not able to segregate the data further into smaller clusters denoting different percentages of GVFs. This is why, as stated previously, an unsupervised learning model is often less accurate and precise than a supervised learning model. One of the potential reasons as to why the model failed to identify 10 clusters might have been due to the high dimensionality of the data (11,081 x 55) thus making it harder for the model to distinguish between. It was also observed in previous experiments [4] that some of the variables experienced the same drifts despite being collected under different conditions. This would also increase the challenge presented to the unsupervised learning model. To overcome this, variable selection can be performed to eliminate and remove irrelevant and confusing variables. This can either be done by getting experts' input or performing additional variable selection modelling or performing dimensionality reduction. Note that variable selection is beneficial regardless of what types of machine learning models we are using. However, for an unsupervised learning model, removing irrelevant and confusing variables will significantly improve its prediction accuracy as well as its modelling speed, due to the fact that it will have less trends, correlations and patterns to recognise. In the next section, we will look at the benefits of using dimensionality reduction technique in unsupervised machine learning models.

3.2 Unsupervised Classification Model - PCA

PCA is an example of a dimensionality reduction technique based on concepts from linear algebra in mathematics. The purpose of using PCA is to reduce the dimension of a given data set by minimising the number of variables needed to retain the maximum information in the original data. Dimensionality reduction is carried out by using linear transformation where a set of possibly correlated variables are transformed into a new set of linearly uncorrelated variables known as principal components (PCs). Each principal component is orthogonal to its subsequent principal component and independent of each other. As a result, PCA is extremely useful when it comes to variables which are highly correlated with each other and thus resolving the problems of multicollinearity as well as transforming data with high dimensionality to a lower dimension, whilst retaining the maximum amount of information. In addition, PCA can be used to help end-users better understand the relationships in variables which can help extract useful insights when working with unlabelled data.

To demonstrate the advantages of using PCA and the type of insights that can be extracted from variables, let us recall that the UFM data collected consisted of 11,081 observations with 55 variables, which is an example of a high dimensional data count which can benefit from dimensionality reduction. Prior to carrying out PCA, it is important to check that our data does in fact consist of highly correlated variables. A simple correlation plot was produced in R where it was observed that our data did indeed contain highly correlated variables. As a result, PCA can be used in this data set. After obtaining the required eigenvalues and eigenvectors, the PCA algorithm then reduced the dimensions of our data into *k* dimensions where the directions with the largest variances are considered to be the most important.

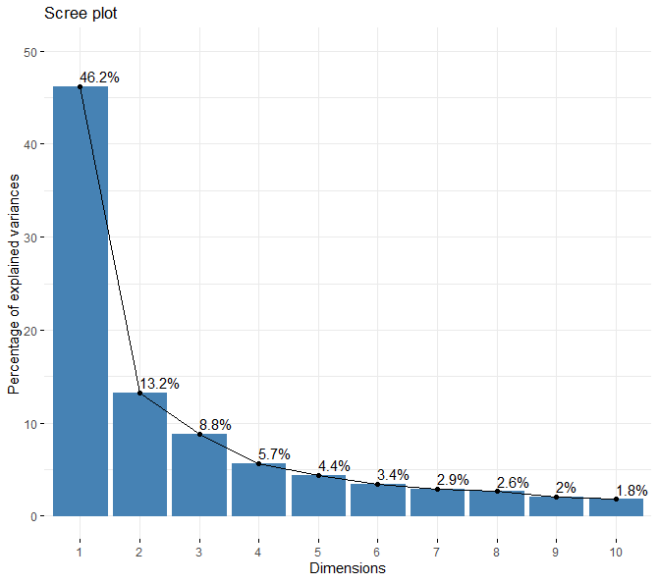


Fig. 6 – Scree Plot

A scree plot (Fig. 6) was obtained in R to identify the number of principal components (shown as dimensions in the plot) we should keep so as to retain the maximum amount of information from the original data. From Fig. 6, we can see that by using the first two dimensions, (first two principal components), we are able to explain around 59.4 % of the variance that occurred in the original data. In other words, by reducing the original dimensions from 55 to only 9, the transformed data can already replicate 89.2 % of variability shown in the original data. Although, we are ultimately losing some information by taking this approach, we have significantly reduced the number of dimensions while still retaining a high percentage of useful information. The percentage of explained variance decreases as the dimension increases. Therefore, variables which contributed and correlated to the first and second dimensions are considered to be the most important. On the other hand, variables which are not correlated with any of the principle components or correlate with the last dimensions are variables with low contributions and therefore they might be removed to simplify the data analysing process.

For end-users who are not familiar with the input variables and the role they play, PCA can also be used to help better understand their dynamic behaviour. A variable correlation plot was produced in R and is shown in Fig. 7, where different types of information can be extracted from such a plot. The correlation plot had grouped variables which shared the same traits and behaviour within the same cluster. The number of clusters was determined using the previously mentioned silhouette technique. Therefore, generally speaking, the input variables portrayed ten different types of behaviour.

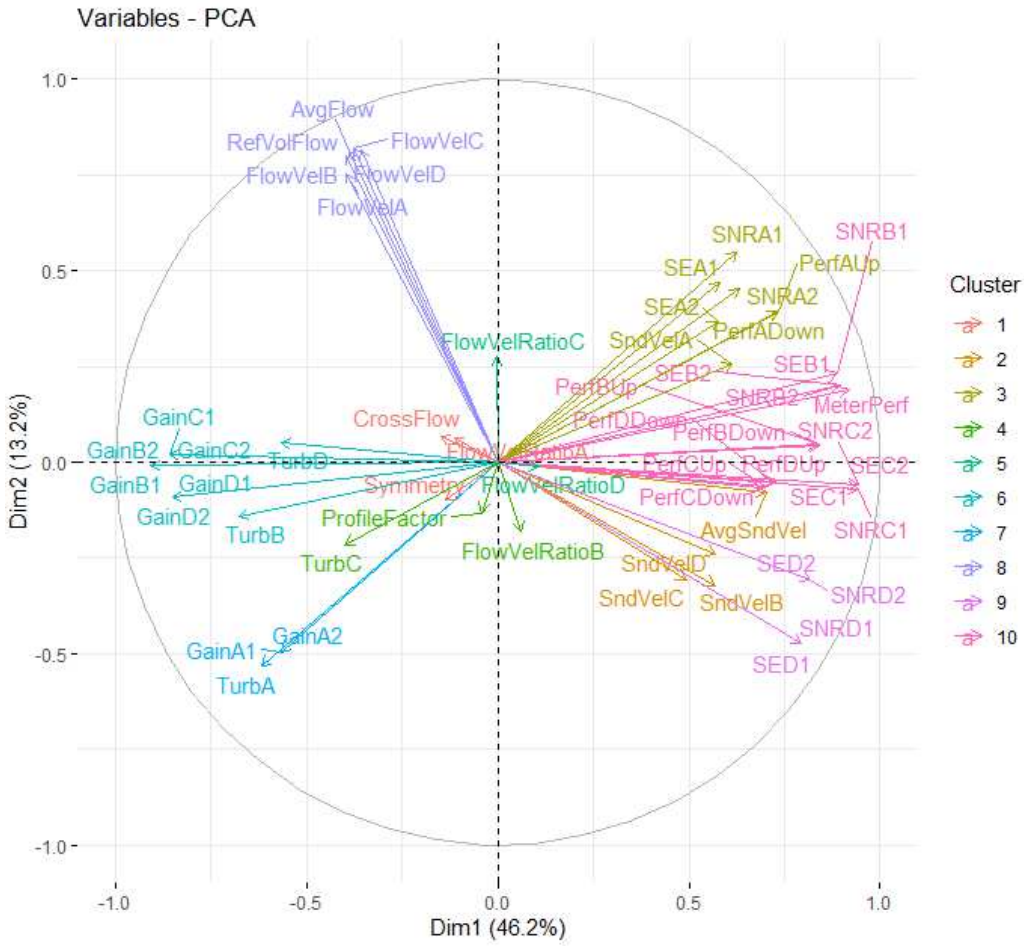


Fig. 7 – Variable Correlation Plot Grouped By Groups

Visualisation such as Fig. 7 can quickly enable end-users to identify which variables share the same traits as others. Variables which are positively correlated are grouped together (and should be colour coded the same, due to being clustered together) while variables which are negatively correlated are positioned on the opposite quadrant. Clusters such as cluster 8 (at the 11 o'clock position) showed that all variables in the group were tightly correlated, whereas clusters such as cluster 10 (at the 3 o'clock position) were more dispersed and hence less tightly correlated. The length of the arrows indicates the predictive strength of each variable, where the further away the variables are from the origin, the more important they are to the first component. In this case, we can see that variables such as cross flow and symmetry were not as important as other variables for predicting the presence of a gas-phase. Thus, making them less indicative in distinguishing between different percentages of GVF.

In this section, PCA was used to reduce the dimensionality of the data, where we have reduced the requirement for 55 variables to explain 100 % of the variability of the data to only needing 9 variables to achieve 89 % of the explained variance. Furthermore, PCA can also be used to understand the dynamical behaviour of different variables and how they interact with each other. This type of information

will be useful to end-users who wish to better understand their variables and determine which variables interact with each other. However, it is important to note that it is not easy to interpret the outputs from PCA as input variables have been transformed into another subspace in the form of principal components. In situations where we have unlabelled data, there is a limited amount of insights that can be extracted, consequently the use of data exploratory techniques such as PCA could provide additional insights into the interactions between different variables when exposed to different process conditions.

Although there are modelling techniques that can be used to handle unlabelled data, it is clear from Section 3.1 and Section 3.2 that supervised machine learning models are more reliable, where insights are more meaningful and interpretable. Due to the nature of unlabelled data, there is a limited amount of actionable insights that can be extracted and expert's inputs are often required to further "translate" the results output by unsupervised models.

4 CASE STUDY 2 - DETECTING THE PRESENCE OF AN UNWANTED MULTIPHASE FLOW IN A CORIOLIS METER

In this section, we look at another case study that was carried out on a Coriolis meter with multiphase flows, to investigate the possibilities of applying similar modelling techniques as already discussed to enhance our understandings on the impact of multiphase flow on the performance of a Coriolis meter.

The data used in this section was obtained from another experiment conducted at TÜV SÜD National Engineering Laboratory, where a Coriolis meter which was installed in the multiphase flow loop while being operated under different multiphase flow conditions which consisted of different GVFs and water cut (WC) percentages ranging at various flow rates. As a result, the data can be categorised into 4 groups – "oil", "oil and gas", "oil and water" and "oil, gas and water".

The ability to detect and predict the amount of water in oil is extremely useful, especially in the oil and gas industry. Having the ability to predict and detect the level of water cut can help end-users indicate when to shut a well in order to optimise its efficiency. Motivated by this, two different machine learning models were constructed to perform the following predictions:

- Predict single-phase flow, multiphase flow and multicomponent flow.
- Predict the percentage of GVF in oil.

Similar to Section 3.1, the models were trained with 70 % of the data to learn the patterns, correlations and interrelationship in variables when exposed to different operating conditions, whereby the models' performance and prediction capability were tested using the remaining unseen 30 % of the data to mimic how well they would have performed in predicting the said condition in the real situation. The prediction results are presented in the following subsections.

4.1 Model 1 - Detect The Presence of a Multiphase Flow

The machine learning model was built to segregate and distinguish the differences in data when the meter was exposed to the four conditions, namely "oil", "oil and gas", "oil and water" and "oil, gas and water". The trained model had an accuracy rate of 94.72 % and a kappa[5] value of 0.92, where the model's parameters were tuned and selected through maximising these measuring metrics. The model was then used to predict the operating condition on the remaining 30 % of the data where the model was required to detect the presence of a multicomponent flow. If a multicomponent flow is detected, the model is then required to further specify and predict the most likely phase condition amongst the choices of "oil and water", "oil and gas" and "oil, gas and water". In other words, the model had to segregate the data into subgroups to indicate which data was collected in which specific condition. The prediction results are summarised in Table 2, where the balanced accuracy[6] for each phase and the overall accuracy are presented in Fig. 8. Note that balanced accuracy and accuracy are not the same: balanced accuracy focused on the prediction capability of each group whilst the overall accuracy considered the general performance of the whole model when applied to the testing data.

Table 2 - Prediction Results on Phase Condition

	Oil (single phase)	Oil and Water	Oil and Gas	Oil, Water and Gas
Sensitivity [7] (true positive)	0.75	1	1	1
Specificity [8] (true negative)	1	0.94	1	1

[5] In statistics, Cohen's kappa [12] measures the interrater reliability. A value of 1 indicates a perfect agreement between each rater. A kappa value of 0 indicates no agreement between each rater other than what would be expected by chance.

[6] Balanced accuracy is calculated using 0.5*(True Positive Rate + True Negative Rate).

[7] Sensitivity = True Positive / (True Positive + False Negative).

[8] Specificity = True Negative / (True Negative + False Positive). Note that 1 – Specificity will yield the false positive rate.

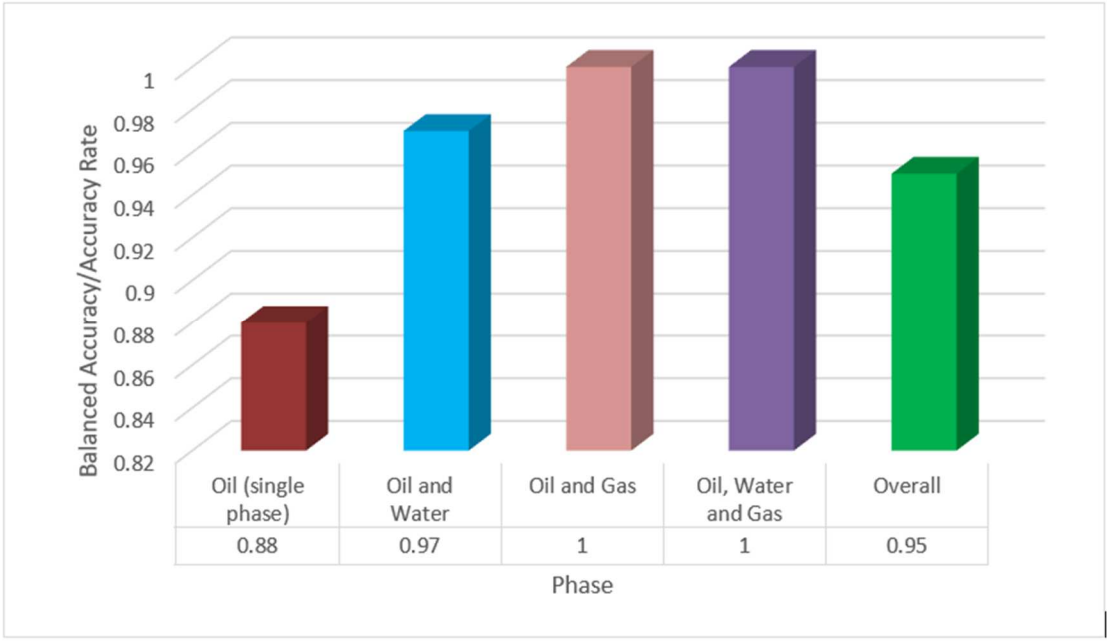


Fig. 8 – Balanced Accuracy for Each of the Four Phase Conditions and the Overall Accuracy for Model 1

Note that different measuring metrics were used in this case study to help us better evaluate and compare the performance of the model. As evident from the results in Table 2, the model has successfully predicted all data within the oil and gas condition and the three-component flow conditions. Some mispredictions were made on the single-phase data, where the model had predicted those to be from the oil and water phase, hence the 0.75 sensitivity rate.

Although it is important for the model to be able to detect all positive cases (faulty conditions), it is equally crucial that it not be overly sensitive, resulting in a high false positive rate, as this will unnecessarily increase the operating costs for the business. The false positive rate can be calculated as "1- Specificity rate" where, in this case, the oil and water phase had a false positive rate of 0.06 – those data were in fact single-phase data.

4.2 Model 2 - Detect the Amount of Gas in Oil

There were 11 different GVFs conditions logged by the data, whereby the model required to learn and distinguish between each of them. Note that similar to Section 3, the measured GVFs (with an average relative uncertainty of around ± 2 %) were rounded to their nearest whole percentage for the purpose of modelling.

This is significantly more challenging than the previous case studies as the more classes/conditions the model needs to learn, the more likely for the model to get confused as some variables could have the same drift patterns under different conditions. Therefore, when handling large number of classes/conditions, more data is needed to ensure any subtle changes are detected and learnt by the models. However, in this case there was only a very limited amount of data. Consequently,

the trained model based on the 70 % training data only had an accuracy rate of 87.79 % and a kappa value of 0.86, which were lower than the previous model. The results from the testing data (unseen data) are given in Table 3 and Fig. 9, where the model's overall accuracy rate was around 0.93. This is also lower than the previous model. Note that the red highlighted areas represent when the model achieved a lower prediction result. In particular, the model only achieved an accuracy rate of 0.75 when predicting data that was output by the meter when exposed to 10 % GVF, followed by a low sensitivity rate of 0.50.

Table 3 - Prediction Results on GVF Percentages

GVF (%)	0	1	1.5	2	2.5	4	6	8	10	15	20
Sensitivity (true positive)	1	0.67	1	1	1	1	1	1	0.5	1	1
Specificity (true negative)	1	1	0.97	1	1	1	1	1	1	0.95	1

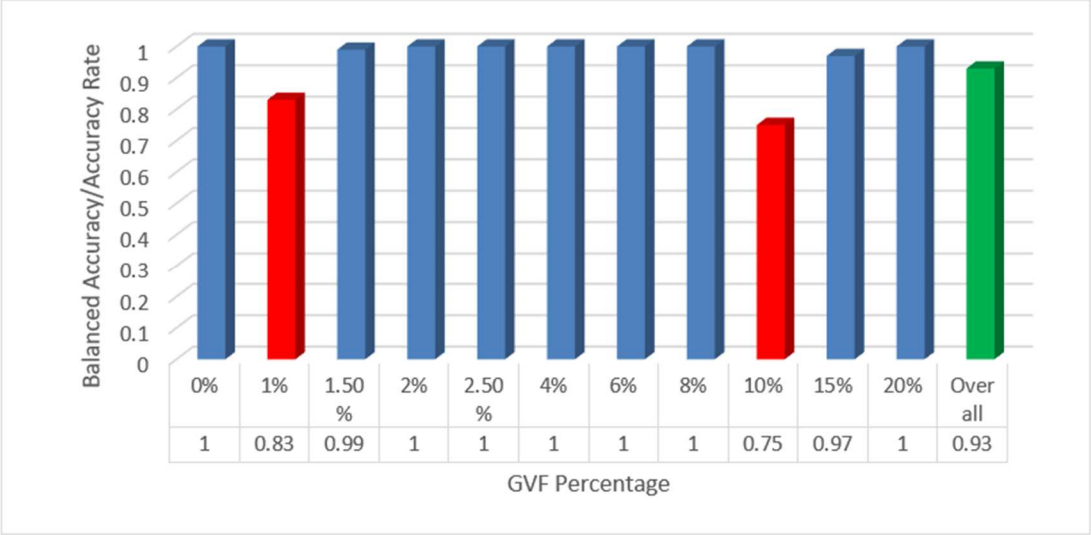


Fig. 9 – Balanced Accuracy for Each of the GVF Percentages and the Overall Accuracy for Model 2

The poorer performance in Model 2 when predicting the GVFs is mostly due to the model having to learn and distinguish more classes, but with limited observations. The performance for Model 2 should improve if more data points can be collected under each GVF to ensure that all the subtle variations in variables can be learned when operating under different GVFs.

5 CONCLUSIONS AND DISCUSSIONS

This paper discussed the use of different machine learning models to detect the presence of an unwanted two-component flow in a UFM and a two-component and three-component flows in a Coriolis meter. Two case studies were discussed in this paper, detailing the modelling results obtained on each of the meter type. Two main types of machine learning models, namely supervised and unsupervised models, were discussed and compared.

A supervised learning model is often more accurate and reliable than an unsupervised learning model due to the fact that we have information on the response variable and thus we can "teach" the model which types of data belong to what kind of condition before it needs to carry out a prediction. However, it is not always possible to have labelled data, especially in real world scenarios, where the data may in fact be passed through multiple logging and data storage systems before it is presented to the end-user. Therefore, how do we expect the model to learn and predict an error when we do not even know what the error is or even how many errors we have? An unsupervised learning model can be used to extract potentially useful information from the data despite not having information on the response variables. This type of model will learn the correlations, patterns and trends by itself and segregate the data into *n* number of conditions based on what is observed. The computational cost associated with an unsupervised learning model is therefore higher than a supervised model. The information obtained from an unsupervised learning model is often used for data exploration purpose to help end-users better understand the dynamics of the data as well as indicating potential anomalies within the data. However, the prediction accuracy from an unsupervised learning model is often lower compared to a supervised learning model as we do not have information on the response variables, which also makes it harder for end-users to interpret the meaning of each cluster.

Case Study 1 looked at the two-phase fluid flow historical data gathered from a UFM, where different percentages of gas were deliberately injected into the fluid in an effort to investigate the subsequent effects on meter performance. Two types of machine learning model were used in this case study where we aimed to predict the percentage of gas present within the fluid based entirely on the correlations and interactions between variables. The models mentioned here will be beneficial to end-users in monitoring and determining whether a second-component is present within the fluid, where such a scenario could severely affect the accuracy of the flow measurement data produced by the flow meter.

The original two-component data obtained from the UFM was cleaned and scaled where rows and columns with missing values were removed. The data was then split into three different parts: training set, validation set and unseen data. The supervised learning model learnt the patterns, trends and correlations from the training data before being validated and tested against several sets of unseen data. An average prediction accuracy of 86 % was achieved for predicting the correct percentage of gas present within the fluid, where one data set had a prediction

accuracy of 99 %. This type of prediction can help end-users in decision-making processes relating to determining the severity of gas effects on meter output data.

The response variables column was then removed to mimic the situation where we have unlabelled data. This set of data was then fed into an unsupervised learning model constructed in R, where the model had successfully identified more than one group of conditions exist within the data. If this was used in practice as a monitoring and fault detection procedure, then it would be a clear indication to end-users that there are some anomalies within the data especially if the data should only represent one group, such as a normal operating condition. Although, in this case, the unsupervised model was not able to identify exactly what each cluster represents, it still provides helpful information which can aid the decision-making process.

The two-phase data consisted of 55 variables which can be challenging for models to digest and sieve out noisy information. As a result, a dimensionality reduction technique, namely PCA was used to reduce the dimension of the data by transforming them into a new subset represented as principal components. As a result, the transformed data only consisted of 9 dimensions whilst retaining 89 % of information shown in the original data.

In Case Study 2, similar modelling techniques were used and extended to a Coriolis meter, where two different models were trained to classify different phase conditions, namely "oil and water", "oil and gas" and "oil, gas and water" as well as predicting the percentages of GVFs the unseen data sets were operating in. Each model had an overall prediction accuracy rate of 95 %, and 93 %, accompanied by a high sensitivity rate and a low false positive rate.

In this paper, machine learning models were used to enhance our understanding of multiphase flow and its effect on the performance of flow meters. They were used due to their ability to digest and dissect complex interrelationships in multiple variables as well as being able to extract hidden patterns, correlations and subtle changes when changing operating conditions which would have been hard to pick up by human observation techniques. For example, it would not have been possible to perform manual trend analysis via supervisory control and data acquisition (SCADA) screens for the amount of data that was produced during the experiments discussed in the case studies. A well-trained machine learning model can apply what it has learned to detect targeted conditions with high certainty. In addition, machine learning models are not affected by a large number of variables, as they have the ability to rank variables in the order of importance and therefore neglect redundant variables automatically.

Results from Case Study 2 demonstrated the capability and the potential of using machine learning models to detect, with high confidence, the presence of a multiphase/multicomponent flow as well as predicting the amount of water and/or gas present within the oil.

The dynamics of multiphase flow and its impact on flow meters is a complicated process which is often dependent on multiple factors. In this paper, data-driven models and advanced modelling techniques have demonstrated the potential capabilities of helping end-users better understand the digital 'diagnostic' data output by flow meters with respect to their bespoke installations and therefore this is a further step in enabling condition-based monitoring on a larger and more agnostic scale. Future research within TÜV SÜD National Engineering Laboratory will continue to explore the potential and capability of generalising data-driven models to apply in wider applications.

6 REFERENCES

[1] H. Manus, M. Tombs, F. Zhou and M. Zamora. "New Applications for Coriolis Meter-Based Multi-Phase Flow Metering in the Oil and Gas Industries". The 10th International Symposium on Measurement Technology Intelligent Instruments, (2011).

[2] J. Weinstein. Multiphase Flow in Coriolis Mass Flow Meters – Error Sources and Best Practices. 28th International North Sea Flow Measurement Workshop 26th – 29th October 2010, (2010), 19 pages.

[3] Y. Liang. Predicting the Remaining Useful Life of Flow Meters During Erosive Flow Conditions. NEL Report No. 2020/435, (2020).

[4] Y. Liang. Monitoring the Performance of Flow Meters Through Advanced Modelling Techniques. 38th International North Sea Flow Measurement Workshop 26-29 October 2020, (2020), 17 pages.

[5] Y. Liang, B. Nobakht and G. Lindsay. The Application of Synthetic Data Generation and Data-Driven Modelling in the Development of a Fraud Detection System for Fuel Bunkering. *Measurement: Sensors*, 18 (2021), p100225.

[6] Y. Liang. Using Data to Unlock Plant Potential. Process and Control, October, (2020).

[7] Y. Liang. Predictive Model Development. The British Measurement and Testing Association Newsletter Spring 2020, (2020), pp. 10-11. Access from: https://www.bmta.co.uk/news-events/news/213-bmta-newsletter-spring-2020.html.

[8] Y. Liang. Optimising Flow Meter Management with Big Data. Oil Review Middle East, 23 (5), (2020), pp. 26-27.

[9] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374, (2016), 16 pages. Accessed from: http://dx.doi.org/10.1098/rsta.2015.0202.

[10] A. Liaw. and M. Wiener. Classification and regression by Randomforest. R News, 2(3), (2002), pp. 18-22.

[11] E. Diday, and J. C. Simon. Clustering analysis. In Digital pattern recognition, 1976, pp. 47-94. Springer, Berlin, Heidelberg.

[12] M. L. McHugh. "Interrater Reliability: the kappa Statistic". *Biochemia medica*, *22*(3), pp. 276–282, (2012).