Valve Health Identification Using Sensors and Machine Learning Methods

M. Atif Qureshi *[1,2] [0000-0003-4413-4476], Luis Miralles-Pechuán[1,2] [0000-0002-7565-6894], Jason Payne[3], Ronan O'Malley[3], and Brian Mac Namee[1] [0000-0003-2518-0274]

[1] Ireland's Centre for Applied AI (CeADAR), University College Dublin, Dublin, Ireland. Email: {luis.miralles, brian.macnamee}@ucd.ie
[2] Technological University Dublin, Dublin, Ireland.
muhammadatif.qureshi@tudublin.ie
[3] Wood, Galway Technology Park, Parkmore, Galway, Ireland. {jason.payne, ronan.omalley}@woodplc.com

Abstract. Predictive maintenance models attempt to identify developing issues with industrial equipment before they become critical. In this paper, we describe both supervised and unsupervised approaches to predictive maintenance for subsea valves in the oil and gas industry. The supervised approach is appropriate for valves for which a long history of operation along with manual assessments of the state of the valves exists, while the unsupervised approach is suitable to address the cold start problem when new valves, for which we do not have an operational history, come online.

For the supervised prediction problem, we attempt to distinguish between healthy and unhealthy valve actuators using sensor data measuring hydraulic pressures and flows during valve opening and closing events. Unlike previous approaches that solely rely on raw sensor data, we derive frequency and time domain features, and experiment with a range of classification algorithms and different feature subsets. The performing models for the supervised approach were discovered to be Adaboost and Random Forest ensembles.

In the unsupervised approach, the goal is to detect sudden abrupt changes in valve behaviour by comparing the sensor readings from consecutive opening or closing events. Our novel methodology doing this essentially works by comparing the sequences of sensor readings captured during these events using both raw sensor readings, as well as normalised and first derivative versions of the sequences. We evaluate the effectiveness of a number of well-known time series similarity measures and find that using discrete Frechet distance or dynamic time warping leads to the best results, with the Bray-Curtis similarity measure leading to only marginally poorer change detection but requiring considerably less computational effort.

Keywords: Time-series, Classification, Anomaly detection, Predictive maintenance models, Sensor data

* M. Atif Qureshi and Luis Miralles-Pechuán equally contributed to this work.

1 Introduction

Predictive maintenance models attempt to identify developing issues with industrial equipment before they become critical [1]. In this paper, we explore predictive maintenance tasks for subsea valves in the oil and gas industry. A valve is a key component in any industrial piping system. Valves are used to regulate the flow of fluids in one direction by opening and closing passageways. To monitor the status of valves each time they are opened or closed a suite of sensors measure volumes and pressures within the valve during the event. These measurements generate a multivariate time series that describes the behaviour of the valve during the opening or closing event.

In this paper, we present strategies to identify the state of a valve following an opening or closing event, using both supervised and unsupervised machine learning methods. In the supervised scenario, we classify valves as *healthy* or *unhealthy* following an opening or closing event based on the sensor data generated during the event. We are concerned with benchmarking the performance of different supervised machine learning algorithms and data representations for this classification task. In particular, our proposed data representation methods are able to extract frequency and time domain features from raw sensor data, increasing the accuracy, which is critical in this scenario.

In the unsupervised scenario, we propose a strategy for anomaly detection by capturing sudden or abrupt changes in valve behaviour. To achieve this, we contrast consecutive readings from a sensor for the same event (open or close), by calculating the distance between the readings. We make use of a number of popular time-series similarity measures, such as dynamic time warping, symbolic aggregate approximation and discrete Frechet distance, and evaluate their suitability for this task. The novelty of our investigation stems from the various signal transformations, such as normalisation and derivative calculations, prior to calculating distances.

The rest of the paper is organised as follows. In Section 2, we review the current state of the art in predictive maintenance, examining classification techniques, time series similarity measures, and anomaly detection. In Section 3, we discuss the dataset and derived features sets used in this paper. In Section 4 and Section 5, we present the supervised and unsupervised approaches, respectively, along with the results of experiments that were conducted to evaluate these approaches. Finally, in Section 6, conclusions and directions for future work are discussed.

2 Related Work

In an environment of lower oil prices, companies in the international energy sector are exploring new ways to reduce the cost of condition-based monitoring services for operating equipment such as subsea valves and actuators. This is being done through the development of models that can simulate the thought processes of experienced hardware engineers and automate or semi-automate the condition-based monitoring process.

Maintenance and intervention for energy assets, typically require costly downtime, deferred energy production and very expensive resources [2]. Tracking the health of equipment from design and installation through to early indicators of functional degradation is important as it enables cost-effective planning of maintenance and intervention programs. Hence, there is a pressing need to build predictive models that identify the state of equipment, determining its health and predicting possible failures as early as possible (or even before that happens).

Various techniques have been proposed for predictive maintenance tasks in the oil and gas industry, despite it being a relatively new concept [3]. A broad categorisation of methods reveals two categories: model-based methods and pattern recognition methods [4]. Model-based methods utilise mathematical calculations and involve a manual analysis of the parameter values measured during the monitoring time and their comparison with the nominal power curve of every oil pump [5]. The more modern pattern recognition methods typically involve the use of sensor data and their working principle is based on the intuition that different system faults initiate different patterns of evolution of the interested variables [6]. These are the patterns that data-driven machine learning methods aim to capture. The model-based methods are, unfortunately, still based on the approximated statistical distribution model, and significant uncertainty is involved in the interpretation of the results [3]. This led the community to investigate artificial intelligence approaches for signal-based fault detection with commonly used techniques, including Artificial Neural Networks, regression models, and Bayesian models [7].

At the same time, a different class of fault detection utilises techniques from the domain of anomaly detection whereby patterns that do not conform to expected behaviour are detected, and extracted [8]. Anomaly detection approaches are particularly suited in scenarios where there is a lack of labelled datasets from the sensor signals, such as ongoing oil and gas operations. Among unsupervised methods, estimates on remaining useful life are also modelled as gradual change detection strategies for predictive maintenance [9].

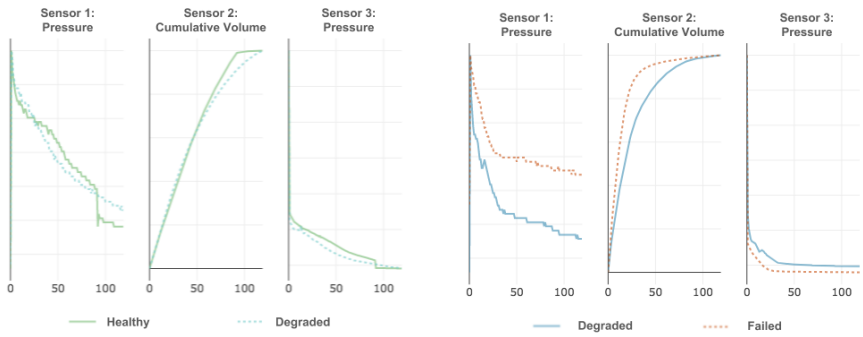
Our paper particularly concerns benchmarking the performance of different supervised machine learning algorithms with a special focus on the extraction of derived features. Additionally, we also focus on unsupervised learning and especially anomaly detection based fault identification of valve failures. Since our data is essentially time-series data, we essentially capture the anomalies by calculating the distance between signals through the application of time-series similarity metrics. The distance metrics we investigated include dynamic time warping (DTW) [10], symbolic aggregate approximation (SAX) [11], Bray-Curtis [12] and Frechet distance [13].

3 Data

In this section, we first describe the dataset used throughout this paper and how it was labelled. We then discuss the derived feature sets that were created from this original dataset.

3.1 Dataset description

The dataset used in this paper is based on monitoring 583 subsea valves over multiple years. The valves are owned by BP (www.bp.com), one of the world's leading integrated oil and gas companies, and the monitoring is supported by Wood (www.woodplc.com), an international energy services company. During the time that the valves were monitored, there was a total of 6,658 open (48.87%) and close (51.12%) events. Each time an event (a valve being opened or closed) takes place, the state of the valve is captured by three different sensors. Sensors 1 and 3 measure pressure, and sensor 2 cumulative volume. During the event, each sensor records 120 readings at regular intervals. This results in three time series (one for each sensor) for each event. Figure 1 (a) shows two examples of the sensor readings for two different closing events.



(a) *healthy* to *degraded*. **(b)** *degraded* to *failed*.

Fig. 1: Usage of distance metric as an anomaly detector.

As part of the on-going monitoring of these valves engineers review visualisations similar to Figure 1 (a) and label the current state of the valve as one of the three possible classes: *healthy*, *degraded*, or *failed*. The *healthy* class represents that the valve is performing within the optimal condition, the *degraded* class represents that the valve's performance has declined from the optimal condition but is still functioning, and the *failed* class represents that the valve has failed to perform the basic function and should be replaced. In Figure 1 (a), one set of readings represents a *healthy* valve and the other represents a *degraded* valve. The difference is most apparent from the sensor 2 readings because the abrupt change of the slope on the top of the healthy signal is transformed into a gradual curvature with smooth transitions in the degraded signal.

The total number of instances in the dataset is 6,658, where 6,232 (93.6%) are *healthy*, 122 (1.83%) are *degraded*, and 304 (4.56%) are *failed*. And each instance represents 120 captured points for sensor 1, sensor 2, and sensor 3, and the category of the output ($healthy, degraded, or failed$). Due to the fact that degraded and failed valves were quickly replaced the number of instances

belonging to these categories is very low compared to the number of healthy valves. This results in a highly imbalanced dataset, which is common in predictive maintenance scenarios where *healthy* instances tend to dominate.

3.2 Time and frequency domain features

To make the classification task more accurate, we extracted a set of time and frequency domain derived features from the raw sensor signals. Feature extraction techniques have been shown repeatedly to make classification tasks easier [14, 15]. Table 1 shows the time domain features which were extracted. Table 2 shows the frequency domain features which were calculated after applying a fast Fourier transform (FFT) [14, 15] on the raw data generated from each of the three sensors. Derived features were calculated independently for the time series arising from each of the three sensors.

Table 1: Time domain features applied over the 120 data points generated by each sensor.

Features	**Description**
$\sigma(X)$	The standard deviation of the signal.
$\overline{X}$	The mean of the signal.
$\overline{\|X\|}$	The mean of the absolute of the signal.
$\left\|\tilde{X}\right\|$	The absolute value of the median of the signal.
$\|Var(X\|$	The variance of the absolute of the signal.
$\|max(X)/min(X)\|$	The absolute value of the ratio of the maximum to the minimum of the signal.
$max(X)$	The maximum value of the signal.
$min(X)$	The minimum value of the signal.
$maxInd(X)$	The index of the maximum value of the signal.
$minInd$	The index of the minimum value of the signal.
$rms(X)$	The root mean square value of the signal.
$zcr(X)$	The zero-crossing rate of the signal.
$skew(X)s$	The skewness of the signal.
$kurtosis(X)$	The kurtosis of the signal.
$P_1(X)$	The first percentile of the signal.
$P_3(X)$	The third percentile of the signal.
$IQR(X)$	The interquartile range of the signal.
$acf(X)$	The autocorrelation of the signal.

3.3 Signal Transformations

As well as calculating derived features, which are primarily used with the supervised approaches in our experiments, we also performed a set of transformations of the signals. The transformations were used to normalise the signals and take

Table 2: Frequency domain features applied over the fast Fourier transform on the 120 data points generated by each sensor.

Features	Description
$\left\|\overline{FFT(X)}\right\|$	The mean of the absolute of the FFT is calculated.
$\left\|\tilde{FFT(X)}\right\|$	The absolute of the media of the FFT is calculated.
$rms(FFT(X))$	The root mean square value of the FFT is calculated.
$S(FFT(X))$	The entropy of the FFT is calculated.
$\eta(FFT(X))$	The Shannon Entropy of the FFT is calculated.
$flatness(FFT(X))$	The spectral flatness of the FFT is calculated.
$\circ(FFT(X))$	The mode of the FFT is calculated.
$\omega_p(FFT(X))$	The peak frequency of the FFT is calculated.

their first derivative. The first derivative is calculated by taking the first-order discrete difference across the sequential 120 points in a signal. This captures the rate of change that takes place across the signals (one for each sensor). We normalise the original signals by applying range normalisation [16], where each signal is linearly scaled to the range $(0, 1)$.

4 Classifying Valve States

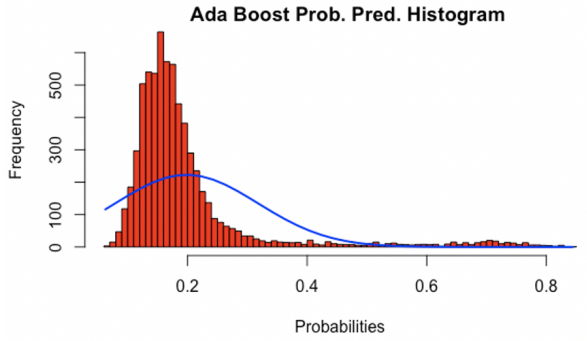


Fig. 2: The distribution of classification scores across the cross-validation experiment for the AdaBoost classifier using the 'Derived' features.

In this section, we present a supervised prediction approach to distinguish between healthy and unhealthy valves using sensor data. In this experiment we reduced the set of target values in the dataset to binary classes by combining the *degraded* and *failed* classes into a new class called *unhealthy*. This is done to addresses the issue of class imbalance within the dataset (see Section 3.1).

In a benchmark experiment we seek to find the best performing classification model and data representation for this task. We consider nine different classi-

fication algorithms: support vector machines (SVM) [17], decision trees (C5.0) [18], k-nearest neighbour algorithms (kNN) [19], random forest ensembles (RF) [20], boosted ensembles (AdaBoost) [21], a deep learning gradient boosting machines (GBM) which is a H2Os implementation of GBM using distributed trees [22, 23], gradient boosting (Xgb) [24], multi-layer perceptrons (NNet) [25], and a deep feed-forward networks (DL) [26]. We explored three different data representations: '*Raw*' representing the original raw data from three sensors, '*Derived*' representing the set of features explained in Tables 1 and 2, and '*Combined*' which is a combination of both 'Raw' and 'Derived'.

Table 3: Full benchmark with leaving-15-out strategy. The *, **, and *** show the top one, two, and three ranked in each column, respectively.

Features	Model	Spec	Precision	$F1_{score}$	AUC	Time(Sec)
Derived	AdaBoost	*****0.9974**	*****0.9420**	*****0.8112**	***0.9900**	86.63
Derived	RF	0.9973	0.9397	***0.8191**	****0.9887**	7.64
Derived	DL Gbm	0.9974	0.9396	0.7905	*****0.9881**	81.24
Derived	C5	0.9965	0.9160	0.7655	0.9841	17.31
Combined	RF	0.9968	0.9145	0.7145	0.9823	73.54
Combined	C5	0.9968	0.9228	0.7660	0.9819	236.08
Combined	DL Gbm	0.9974	0.9396	0.7905	0.9806	474.31
Combined	AdaBoost	0.9971	0.9348	0.8050	0.9805	940.63
Raw	RF	0.9961	0.8966	0.6969	0.9779	58.67
Raw	DL Gbm	0.9947	0.8850	0.7791	0.9770	516.40
Raw	C5	0.9960	0.8884	0.6757	0.9763	234.20
Raw	AdaBoost	0.9961	0.8961	0.6946	0.9687	882.33
Raw	NNet	0.9918	0.7583	0.5556	0.9466	21.79
Combined	NNet	0.9937	0.8274	0.6328	0.9451	26.13
Raw	Gbm	0.9871	0.7173	0.6266	0.9228	5.33
Derived	NNet	0.9957	0.9069	0.8030	0.9165	25.85
Derived	Gbm	0.9919	0.6324	0.3433	0.9158	0.85
Combined	Gbm	0.9913	0.6516	0.3884	0.9034	6.00
Derived	kNN	0.9969	0.9333	****0.8185**	0.8757	3.22
Raw	kNN	0.9953	0.8953	0.7726	0.8675	17.53
Combined	kNN	0.9942	0.8800	0.7940	0.8525	8.76
Combined	Xgb	****0.9976**	****0.9436**	0.7956	0.8426	3.64
Derived	Xgb	***0.9981**	***0.9537**	0.7917	0.8374	0.53
Raw	SVM	0.0047	0.0527	0.0998	0.8142	356.01
Combined	SVM	0.0127	0.0495	0.0937	0.7732	286.40
Raw	Xgb	0.9965	0.9009	0.6814	0.7722	3.40
Raw	DL	0.9775	0.8941	0.5276	0.6759	541.61
Derived	DL	0.9742	0.9258	0.5219	0.6688	459.35
Derived	SVM	0.0023	0.0507	0.0960	0.6247	478.72
Combined	DL	0.9566	0.7565	0.3459	0.5904	43.37

The parameter values were selected after performing grid-search to tune the hyper-parameters over the 10-fold cross-validation [24]. After this process, the models were implemented in R Studio with the following configurations: SVM (method= "C-classification", kernel="linear"), C5.0 (trials = 100, winnow = TRUE, model="tree"), kNN (k=3, probability=TRUE, algorithm ="cover tree"), RF (ntree=450, norm.votes=FALSE), AdaBoost (mfinal=300, maxdepth =5, coeflearn="Zhu"), NNet (size=21, rang=0.01, decay=5e-4, maxit=500), DL Gbm (ntrees =2500, learn rate=0.001, sample rate=0.7, max depth=15, col sample rate=0.8), DL NN (activation="Tanh", balance classes=TRUE, hidden = c(100, 100, 100), epochs=3, rate=0.1, rate annealing=0.01), Xgb (booster = "gbtree", eval metric = "auc", eta = 0.02, max depth= 15, subsample=.8, colsample bytree= .87, min child weight= 1, scale pos weight=1).

The signals represent independent events when a valve was opened or closed. Therefore there are no dependencies between the instances. Additionally, there are only 120 points, which is a very small number for applying a time window. To evaluate the performance of each algorithm-data representation combination, we performed a leave-n-subjects-out cross-validation experiment [27]. We chose this evaluation strategy as each of the 583 valves represented in our dataset appears multiple times in the $6,648$ events. This means that in a standard k-fold cross-validation experiment events from the same valve would be likely to appear in both the train and test sets which could lead to an overly optimistic assessment of model performance. Specifically, we use all events from 15 valves as the test set in each fold of the cross-validation which leads to 39 folds. We measure the performance of models using macro-averaged F1 score [28] and area under the ROC curve (AUC) [29]. Table 3 shows the results of the complete benchmark ordered by AUC scores. We also show the time taken to perform the leave-n-subjects-out cross-validation experiment in each case.

As can be seen from Table 3, the Adaboost and Random Forest ensembles using the 'Derived' features outperform the other models in terms of AUC. In the operational context, in which these models are likely to be deployed, the goal is to ensure the detection of more true *unhealthy* valves even if some *healthy* valves are incorrectly predicted as *unhealthy*. There are scenarios in which reducing the errors of one class is more important than doing so in the other classes. For example, it is better to diagnose a patient with cancer when he or she is healthy than the opposite case. Likewise, for our problem, we want to avoid predicting valves as healthy if they are not. And as some authors suggest [30], a class can be prioritised by applying a threshold. In other words, we can tune the classification threshold to reduce the number of false negatives while still predicting a reasonable number of true negatives.

This issue is illustrated in Figure 2, which shows the distribution of classification scores across the cross-validation experiment for the AdaBoost classifier. Most classification scores are close to 0.2, the default classification threshold is 0.5, but better performance can probably be achieved with a different value. We use a criterion to select the best classification threshold value by maximising the sum of specificity and sensitivity as described in [31]. Table 4 shows the results

Table 4: Best cut-off parameter for AdaBoost leaving 15 valves out. SUM represents the summation of sensitivity and specificity. According to the established criteria to minimise the errors in both classes, we select the threshold where SUM gets the higher value.

Cut-off	AUC	Acc	Sens	Spec	Prec	$F1_{score}$	TP	FP	FN	TN	SUM
0.300	0.9879	0.9675	0.9351	0.9697	0.6730	0.7827	389	189	27	6043	1.9048
0.295	0.9879	0.9657	0.9351	0.9677	0.6593	0.7733	389	201	27	6031	1.9028
0.290	0.9879	0.9636	0.9447	0.9649	0.6422	0.7646	393	219	23	6013	1.9096
0.285	**0.9879**	**0.9607**	**0.9519**	**0.9613**	**0.6217**	**0.7522**	**396**	**241**	**20**	**5991**	**1.9132**
0.280	0.9879	0.9571	0.9519	0.9575	0.5991	0.7354	396	265	20	5967	1.9094
0.275	0.9879	0.9528	0.9567	0.9525	0.5735	0.7171	398	296	18	5936	1.9092
0.270	0.9879	0.9495	0.9591	0.9488	0.5557	0.7037	399	319	17	5913	1.9079

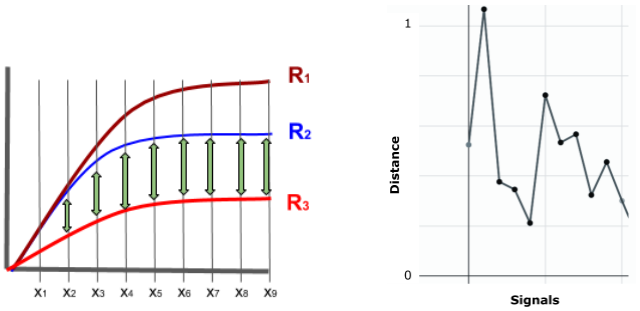
of setting different cut-off values for AdaBoost, where the area-under-the-curve, accuracy, sensitivity, specificity, precision, F1-score, true positives, false positives, false negatives, and true negatives are shown. As can be seen from the results, we are able to tune the classification threshold value while minimising false positives and yet maximising true negatives. According to the established criteria to minimise the errors in both classes, we select the threshold where *SUM* gets the higher value.

The overall results of this experiment show that it is possible to classify valve states to a high level of accuracy and balance false alarms with detecting actual defects. This approach should work well in operational contexts in which valves in operation are of a similar model and operate under similar conditions to those in the labelled training set. This is the case in many scenarios. There is, however, a *cold start* problem in other scenarios in which valves that are new or that will operate under unique conditions are deployed. The supervised learning approach will not work in these scenarios. The next section describes an unsupervised anomaly detection approach that addresses this scenario.

5 Detecting Anomalous Valve Behaviour

If new valve types come online or valves are put into operation in contexts very different from what has been seen before, the supervised approach to recognising valve health will not work as the data generated by these new valves will be so different to what has been seen before. This is what we refer to as the *cold start problem*. Instead, an unsupervised approach is more appropriate. To detect anomalous behaviour in valves we calculate the distance between signals over consecutive events from the same valve and flag an anomaly when this distance is sufficiently large. Performing this anomaly detection, therefore, requires selecting an appropriate distance metric to compare consecutive signals, and then thresholding this signal to flag anomalies. Figure 3(a) shows an example with the distance between readings R_3 and R_2 highlighted by the bi-directional arrows. Figure 3(b) shows a typical plot of distances between consecutive signals over time.

In this section, we present two sets of experiments. First, we conduct an experiment to select the top-performing distance metrics for the anomaly detection task. Then we use the winning distance metrics from the first experiment to evaluate the feasibility of using it to perform anomaly detection. We refer to this as a feasibility study, as we currently do not implement an approach to set the threshold algorithmically but rather choose the best possible threshold for a given test dataset.



(a) The distance between two sample readings is highlighted by bi-directional arrows.

(b) The distance between consecutive signals can be used to detect anomalies.

Fig. 3: Usage of distance metric as an anomaly detector.

5.1 Effectiveness of distance metrics

To evaluate the suitability of different distance metrics to the kind of data we are studying, we performed a classification experiment using a 1-nearest-neighbour-classifier [19], which is known to be a robust approach for time series problems [32]. To conduct this experiment, we selected a subset of the valves in the dataset described in Section 3 for which at least 7 events labelled as *healthy* and one event labelled either *degraded* or *failed* existed. 11 valves matched this criterion and for these 11 valves, 389 events were present in the dataset. We randomly choose 201 of these as training examples (128 *healthy*, 61 is *degraded*, and 12 *failed*) and 188 as test examples (107 *healthy*, 68 *degraded*, and 13 *failed*). We then performed a simple experiment in which we trained 1-nn classifiers based on different distance measures on the training set and evaluated them on the test set. We measured classifier performance using F1 score. We perform these experiments independently for the time series that come from the three different sensors.

We experiment with twelve well-known distance measures that are used with time series data: Euclidean, Minkowski, Manhattan, Chebyshev, discrete Frechet[13], Bray-Curtis [12], dynamic time warping (DTW)[10], Hausdorff, Levenshtein, Canberra, SAX[11], and SAX+DTW (better explanations on these

measures can be found in [32, 16]). Table 5 shows the performance of different distance measures for all sensors, an average of the performance of all sensors, and the running time.

Table 5: The performance of distance metrics using raw sensor data along with running time. The *, **, and *** show the top one, two, and three ranked $F1scores$ in the column, respectively.

Metric	$\mathbf{F1_{score}}$ S.1	$\mathbf{F1_{score}}$ S.2	$\mathbf{F1_{score}}$ S.3	Avg-$\mathbf{F1_{score}}$	Time(Sec)
DTW	***0.8677**	***0.9101**	***0.8307**	***0.8695**	4795.43
Hausdorff	0.7884	*****0.9048**	0.8148	****0.836**	173.31
Manhattan	0.8095	0.873	****0.8254**	****0.836**	51.87
Bray-Curtis	0.8095	0.873	*****0.8148**	0.8324	55.96
Frechet-DISC	****0.8148**	0.8466	*****0.8148**	0.8254	4310.39
Euclidean	0.7831	*****0.9048**	0.7725	0.8201	54.38
Minkowski	0.7831	*****0.9048**	0.7725	0.8201	53.31
SAX	0.7302	***0.9101**	0.7196	0.7866	196.13
Canberra	0.7249	0.8783	0.7566	0.7866	53.31
Chebyshev	0.7778	0.8466	0.7249	0.7831	50.71
SAX+DTW	0.7725	0.8201	0.6402	0.7443	1172.69
Levenshtein	****0.8148**	0.7143	0.6825	0.7372	1169.29

We can see that DTW outperforms all other distance metrics, however, this is at the expense of significant running time. For use as the basis of an anomaly detector, for each sensor and the average we selected the 3 best performing distance measures. These are: DTW, Frechet-DISC, Levenshtein, SAX, Hausdorff, Euclidean, Minkowski, Manhattan, and Bray-Curtis.

5.2 Anomaly detection using distance metrics

In this experiment, we evaluate the suitability of the distance measures, selected from the previous experiment, as the basis for an anomaly detector that captures an abrupt change in the state of a valve. For this experiment, we use the same subset of 11 valves used in the previous section. We attempt to classify each opening or closing event by each valve as *anomalous* or *normal*. To use as a gold standard for measuring performance in our experiments we mark the transitions between event labels as *anomalous* and all other events as *normal*. So, for example, when an event labelled as *degraded* follows an event labelled as *healthy* we mark that event as *anomalous*. If, however, subsequent events by the same valve are also labelled as *degraded* then they are labelled in this experiment as *normal* as the approach is designed to recognise abrupt changes in behaviour. This means that for each valve we have a series of distance measures similar to that shown in Figure 3(b) with each point marked as *anomalous* or *normal* and measure how well this signal allows anomalous events to be separated from

normal ones when it is based on different distance measures. We also experiment with applying the distance measures to the original raw sensor signals, the first derivative of the sensor signals, and range normalised versions of the sensors signals (see Section 3.3). In all cases, we measure the ability of an approach to detect anomalies using F1 score.

In addition to measuring the ability of our approach to capturing anomalies using just one of the three sensors we also investigate an approach that allows voting among sensors with following strategies: *(i)* If the signal from any sensor identifies an event as anomalous, it is flagged as an anomaly. *(ii)* If the signals from the majority of sensors identify an event as an anomaly, it is flagged as an anomaly. *(iii)* If the signals from all of the sensors identify an event as anomalous, it is flagged as an anomaly.

In this experiment, we do not attempt to determine thresholds algorithmically. Instead, for the signal arising from each valve, we examine all possible thresholds and report the one that leads to the highest F1 score. This indicates the best possible performance that could be achieved using a particular data representation and distance measure and is sufficient to compare the feasibility of using this approach for anomaly detection. We leave the algorithmic selection of thresholds for future work.

Table 6: The performance of anomaly detectors based on different distance metrics for valve opening and closing events when comparing the distance between the last two signals. (s), (δs), and $(norm)$ indicate that the named measure was applied to the original signal, the derivative of the original signal, or the normalised signal, respectively. The subscripts All, Maj, and S.x indicate that the anomaly detection decisions were made using all sensor agreement voting, majority voting, or only with sensor x, respectively.

Opening Events		Closing Events	
Metric	$\mathbf{F1_{score}}$	**Metric**	$\mathbf{F1_{score}}$
Bray-Curtis$(s)_{\text{All}}$	0.7745	Frechet-DISC$(norm)_{\text{All}}$	0.8847
Euclidean$(s)_{\text{Maj}}$	0.7603	DTW$(\delta s)_{\text{All}}$	0.8762
Minkowski$(s)_{\text{Maj}}$	0.7603	Bray-Curtis$(\delta s)_{\text{All}}$	0.8578
Manhattan$(s)_{\text{Maj}}$	0.7571	Euclidean$(norm)_{\text{All}}$	0.8451
Hausdorff$(s)_{\text{Maj}}$	0.7521	Hausdorff$(norm)_{\text{All}}$	0.8451
Frechet-DISC$(s)_{\text{Maj}}$	0.7285	Minkowski$(norm)_{\text{All}}$	0.8451
DTW$(s)_{\text{All}}$	0.7213	SAX$((\delta s)_{\text{S.2}}$	0.8335
SAX$(\delta s)_{\text{All}}$	0.6654	Manhattan$(\delta s)_{\text{All}}$	0.8249
Levenshtein$(\delta s)_{\text{All}}$	0.6516	Levenshtein$(s)_{\text{All}}$	0.7555

Table 6 shows the best combination of signal and sensor voting strategy for each distance measure. From this table, we can see a reasonably good ability to recognise anomalies using several distance measures.

Rather than basing the anomaly detection signal on the distance between just a pair of signals we also experiment with comparing the current signal to the average of the preceding three signals, as this could help with smoothing noise from the signal. We refer to this approach as $avg_{step\text{-}3}$ to distinguish it from the previous approach, referred to as $abs_{step\text{-}1}$. Table 7 shows the performance of the anomaly detectors based on the $avg_{step\text{-}3}$ approach. This addition, however, did not improve the performance of the anomaly detectors.

Table 7: The performance of distance metrics as an anomaly detector for the open and close event of the valve when comparing distance between the average of last three signal and the most recent signal. Where, (s), $(norm)$, show the original signal and normalised signal, respectively. Furthermore, $_{\text{All}}$ and $_{\text{Maj}}$ show all sensor agreement and majority agreement, respectively.

Open Event		Close Event	
Metric	**F1$_{\text{score}}$**	**Metric**	**F1$_{\text{score}}$**
Frechet-DISC$(s)_{\text{All}}$	0.6667	Manhattan$(s)_{\text{Maj}}$	0.8710
Bray-Curtis$(s)_{\text{All}}$	0.6269	DTW$(norm)_{\text{Maj}}$	0.8684
DTW$(s)_{\text{All}}$	0.6269	Bray-Curtis$(s)_{\text{Maj}}$	0.8436
Manhattan$(s)_{\text{All}}$	0.6269	Euclidean$(norm)_{\text{All}}$	0.8344
Euclidean$(s)_{\text{All}}$	0.6197	Hausdorff$(norm)_{\text{All}}$	0.8344
Hausdorff$(s)_{\text{All}}$	0.6197	Minkowski$(norm)_{\text{All}}$	0.8344
Minkowski$(s)_{\text{All}}$	0.6197	Frechet-DISC$(s)_{\text{All}}$	0.8318
SAX$(s)_{\text{All}}$	0.4823	SAX$(s)_{\text{Maj}}$	0.7105
Levenshtein$(s)_{\text{All}}$	0.3169	Levenshtein$(\delta s)_{\text{All}}$	0.4086

Table 8 shows a summary of the results from Tables 6 and 7 (average F1-scores are based on micro averaging). Based on the average scores Frechet-DISC performs the overall best, with DTW close behind. Both Frechet-DISC and DTW are computationally very expensive (see Table 5), however, so the third-best measure, Bray-Curtis, is interesting as it is at least 77 times faster than the other two and only marginally more poorly performing. Therefore, if computational speed is a consideration (as may be the case in real-time field deployments), our findings demonstrate Bray-Curtis as a good choice, and if computational speed is a not a crucial requirement then Frechet-DISC or DTW are good choices.

Table 8: Shows the overall best performing metric using micro averaging of $F1_{score}$. There are 115 and 189 open and close events respectively, for $abs_{step\text{-}1}$, and 77 and 154 open and close events respectively for $avg_{step\text{-}3}$. These instances report those valves where at least an anomalous behaviour was once observed. The top ranked ones for in each column is represented in bold with the (*).

	$abs_{step\text{-}1}$			$avg_{step\text{-}3}$			**Overall**
Metric	**Open**	**Close**	**Avg.**	**Open**	**Close**	**Avg.**	**Avg.**
Frechet-DISC	0.7285	***0.8847**	0.8256	***0.6667**	0.8319	0.7694	0.8013
DTW	0.7213	0.8762	0.8176	0.6269	0.8685	0.7771	0.8001
Bray-Curtis	***0.7745**	0.8578	***0.8263**	0.6269	0.8437	0.7617	0.7984
Manhattan	0.7571	0.8249	0.7993	0.6269	***0.871**	***0.7787**	0.7904
Euclidean	0.7603	0.8451	0.813	0.6197	0.8344	0.7532	0.7872
Minkowski	0.7603	0.8451	0.813	0.6197	0.8344	0.7532	0.7872
Hausdorff	0.7521	0.8451	0.8099	0.6197	0.8344	0.7532	0.7854
SAX	0.6654	0.8335	0.7699	0.4823	0.7105	0.6242	0.707
Levenshtein	0.6516	0.7555	0.7162	0.3169	0.4087	0.374	0.5685

6 Conclusion and Future Directions

In this contribution, we presented predictive maintenance models for the identification of developing issues in the subsea valves. We discussed supervised and unsupervised approaches to aid in the assessment of the health of the valve using the sensor data. For the supervised approach, we modelled the problem as a binary classification problem between healthy and unhealthy valves due to operational needs. AdaBoost was discovered to be the best performing model and was able to gain 1.21% in terms of AUC using derived frequency and time domain features (0.99 with derived features) compared to original raw data features from sensors (0.9725 with raw features). Random Forest performed comparable to the AdaBoost but required considerably less computational effort (11.33 times faster than AdaBoost). Furthermore, in order to address an acceptable trade-off between the number of true negatives (TN) and the number of false negatives (FN), we adjusted the cut-off parameter to maximise TN (predicting *healthy* as *healthy*) while minimising FN (predicting unhealthy as healthy). We use a criterion to justify the trade-off and reported 0.961 of accuracy, with as little as 20 FN and yet a high value of 5991 TN. For the unsupervised approach, we found discrete Frechet and dynamic time warping as the best performing distance metric for anomaly detection and Bray-Curtis under-performed the best by a small fraction of 0.03 F1-score, but performed 77 times faster.

As a future direction for the supervised approach, we intend to investigate the application of two cut-off parameters instead of one to predict three classes: *healthy* (higher than the first cut-off), *unhealthy* (lower than the second cut-off) and *warning* (between the first and second cut-off parameters). This will aid in the reduction of the FN along with the FP and can be used to notify the maintenance team of the valves which are likely to fail. Consequently, this idea can be transformed into a regression approach informing the maintenance team about

the number of days after which the valve is likely to fail. For the unsupervised anomaly detection approach, we intend to investigate strategies to establish the optimal threshold, and also, we intend to investigate the performance of derived frequency and time domain features as a metric for anomaly detection.

Acknowledgements. This publication has emanated from research conducted with the support of Enterprise Ireland (EI), under Grant Number IP20160496 and TC20130013. The data was kindly supplied by BP, supported by Wood.

References

1. A Delmas, Mohamed Sallak, Walter Schön, and L Zhao. Remaining useful life estimation methods for predictive maintenance models: defining intervals and strategies for incomplete data. *in Industrial Maintenance and Reliability Manchester, UK 12-15 June, 2018*, page 48, 2018.
2. Marta Fernandes, Alda Canito, Verónica Bolón-Canedo, Luís Conceição, Isabel Praça, and Goreti Marreiros. Data analysis and feature selection for predictive maintenance: a case-study in the metallurgic industry. *International Journal of Information Management*, 46:252–262, 2019.
3. Sze-jung Wu, Nagi Gebraeel, Mark A Lawley, and Yuehwern Yih. A neural network integrated decision support system for condition-based optimal predictive maintenance policy. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(2):226–236, 2007.
4. Francesco Di Maio, Jinfei Hu, P Tse, M Pecht, K Tsui, and Enrico Zio. Ensemble-approaches for clustering health status of oil sand pumps. *Expert Systems with Applications*, 39(5):4847–4859, 2012.
5. Jingwen Tian, Meijuan Gao, Kai Li, and Hao Zhou. Fault detection of oil pump based on classify support vector machine. In *Control and Automation, 2007. ICCA 2007. IEEE International Conference on*, pages 549–553. IEEE, 2007.
6. HQ Wang and P Chen. Fault diagnosis of centrifugal pump using symptom parameters in frequency domain. *Agricultural Engineering International: CIGR Journal*, 2007.
7. Isaac Animah and Mahmood Shafiee. Condition assessment, remaining useful life prediction and life extension decision making for offshore oil and gas assets. *Journal of Loss Prevention in the Process Industries*, 2017.
8. Luis Martí, Nayat Sanchez-Pi, José Manuel Molina, and Ana Cristina Bicharra Garcia. Anomaly detection based on sensor data in petroleum industry applications. *Sensors*, 15(2):2774–2797, 2015.
9. Qiyao Wang, Shuai Zheng, Ahmed Farahat, Susumu Serita, and Chetan Gupta. Remaining useful life estimation using functional data analysis. In *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–8. IEEE, 2019.
10. Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.
11. Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2):107–144, 2007.
12. J Roger Bray and John T Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349, 1957.

13. Thomas Eiter and Heikki Mannila. Computing discrete fréchet distance. Technical report, Citeseer, 1994.
14. Hiram Ponce, Luis Miralles-Pechuán, and María de Lourdes Martínez-Villaseñor. A flexible approach for human activity recognition using artificial hydrocarbon networks. *Sensors*, 16(11):1715, 2016.
15. Asanka Sayakkara, Luis Miralles-Pechuán, Nhien-An Le-Khac, and Mark Scanlon. Cutting through the emissions: Feature selection from electromagnetic side-channel data for activity detection. *Forensic Science International: Digital Investigation*, 32:300927, 2020.
16. John D Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press, 2015.
17. Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.
18. J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.
19. Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.
20. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
21. Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.
22. Cliff Click. Gradient boosted machines with h2o, 2015.
23. Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
24. Tianqi Chen, Tong He, Michael Benesty, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4, 2015.
25. Simon Haykin and Neural Network. A comprehensive foundation. *Neural networks*, 2(2004):41, 2004.
26. Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.
27. Ganggang Xu, Jianhua Z Huang, et al. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *The Annals of Statistics*, 40(6):3003–3030, 2012.
28. Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer, 2005.
29. Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.
30. Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.
31. Farrokh Habibzadeh, Parham Habibzadeh, and Mahboobeh Yadollahie. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia medica: Biochemia medica*, 26(3):297–307, 2016.
32. Rafael Giusti and Gustavo EAPA Batista. An empirical comparison of dissimilarity measures for time series classification. In *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*, pages 82–88. IEEE, 2013.