Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 52-11 (2019) 212–217

Machine Learning approaches for Anomaly Detection in Multiphase Flow Meters

Tommaso Barbariol [*] **Enrico Feltresi** [**] **Gian Antonio Susto** [***]

[*] Department of Information Engineering, University of Padova, Italy
[**] Pietro Fiorentini S.p.A., Arcugnano (VI), Italy
[***] Department of Information Engineering and Human-Inspired Technology Center, University of Padova, Italy. E-mail: gianantonio.susto@dei.unipd.it

Abstract: Multiphase Flow Meters (MPFM) are important metering tools in the oil and gas industry. A MPFM provides real-time measurements of gas, oil and water flows of a well without the need to separate the phases, a time-consuming procedure that has been classically adopted in the industry. Evaluating the composition of the flow is fundamental for the well management and productivity prediction; therefore, procedures for measuring quality assessment are of crucial importance. In this work we propose an Anomaly Detection approach to MPFM that is effectively able to hand the complexity and variability associated with MPFM data. The proposed approach is designed for embedded implementation and it exploits unsupervised Anomaly Detection approaches like Cluster Based Local Outlier Factor and Isolation Forest.

Keywords: Anomaly Detection, Data Mining, Data Fusion, Machine Learning, Multiphase Flow Meter, Oil & Gas Industry, Self-Diagnosis.

1. INTRODUCTION

Monitoring of wells flow production and composition has become very significant in oil and gas industry, particularly as fields become economically marginal and reservoirs deplete (AL-Qutami et al., 2018). In this context, the past decade has seen an increased attention towards Multiphase Flow Meters (MPFM) (Corneliussen et al., 2005). The MPFM is a diagnostics and metering tool that provides real-time simultaneous measurements of the commingled flow of oil, water and gas. It combines the measurement of several sensors without the need to separate the phases. Recently Machine Learning (ML) approaches have been developed to enhance the MPFM capabilities (Yan et al., 2018).

An ensemble approach, based on Neural Networks, is presented in (AL-Qutami et al., 2018) to improve the flow composition estimations. With the same aim, a Long-Short Term Memory-based approach is shown in (Andrianov, 2018). Also in (AL-Qutami et al., 2017) Neural Network architectures are used for the estimation of the flow composition (Radial Basis Network), while in (Ristanto and Horne, 2018) a broader variety of ML algorithms is compared.

However, no work adopting unsupervised learning approaches to enhance MPFM capabilities has been presented in the scientific literature. This work aims at filling this gap by introducing ML approaches for Anomaly Detection (AD) in MPFM. AD is an important area of unsupervised learning that aims at detecting anomalous data that differ significantly from previously seen observations. AD approaches have been employed for Fault Detection (Meneghetti et al., 2018b), Smart Monitoring and Self-Diagnosis technologies in smart devices (Meneghetti et al., 2018a) and products (Theissler, 2017), industrial processes (Susto et al., 2017a) and productions (Susto et al., 2017b).

In particular, in this work we present a AD system to be implemented on board the MPFM and that will be used to assess the quality of produced metrology measurements. Detecting metrology anomalies in MPFM is challenging mainly because: (i) we must decouple data related to metrology anomalies from data associated to new operating conditions; (ii) MPFM generated data are strongly affected by the flow and the well conditions; (iii) MPFM data are multivariate. The proposed system aims at overcoming those challenges and provides a reliability score associated with the measurements.

The contribution of this work can be summarized as follows:

- This is one of the first Machine Learning-based technology for MPFM in literature;
- To the best of our knowledge, this is the first approach in literature that applies unsupervised ML techniques to MPFM and implements ML methodologies onboard the MPFM;
- The work employs state-of-the-art unsupervised AD approaches like Isolation Forest and Cluster-based Local Outlier Factor.

The rest of the paper is organized as follows: MPFM and its principles are described in Section 2. In Section 3 the proposed Self-Diagnosis solution is presented with a brief review of the employed unsupervised AD methods.

2405-8963 © 2019, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2019.09.143

Experimental settings and results are detailed in Section 4; finally, conclusive remarks and future works are discussed in Section 5.

2. MULTIPHASE FLOW METER

The MPFM is non-intrusive, in-line meter for measuring flow rates of oil, water and gas in the dispersed phase of the flow. MPFMs are used for onshore and offshore oil wells, both topside and subsea. In this work data have been collected using a MPFM (PietroFiorentini, 2011) which combines the input of five sensors to compute the flow rate of oil, water and gas (Falcone et al., 2009).

MPFMs exploit the combination of the following measurement principles:

- Venturi differential pressure;
- Electrical impedance;
- cross-correlation of impedance measurements (capacitance/conductance);
- gamma ray densitometry;
- absolute pressure and temperature.

Oil, water and gas flow rates are calculated based on the measurements obtained by the electrodes and the measurement of the differential pressure (DP) across a Venturi throat. The capacitance or conductance of the mixture flowing through the meter is measured in the impedance section. Velocity is measured by cross-correlating the high resolution time signals from the electrodes. The MPFM is also equipped with a gamma densitometer. In the following the various metrology modules are briefly described.

Venturi Module - The DP is measured at the inlet and the throat of the venturi and is proportional to the velocity of the fluids passing by and the density of the mixture. In the same module both the absolute temperature and pressure are measured.

Impedance Module - Impedance module is made up of a series of electrods. Depending on flow regime, it works between capacitive and conductive modes, measuring respectively the permittivity and conductivity of the flow. Since multiple electrode couples are used, linear velocity of the flow is determined through cross-correlation of the measurements of each couple of electrodes.

Figure 1. Example of a Subsea MPFM

Gamma Module - This technology is based on the property of gamma rays to interact with matter and to exponentially attenuate as they pass through it. The attenuation is proportional to the amount of the passed material, therefore the module detects the density variations in the flow. The source of gamma rays is the radioactive isotope of Cesium ^{137}Cs, it has a half-life of almost 30 years.

3. ML-BASED MPFM SELF-DIAGNOSIS APPROACH

To cope with the challenges described in the previous section, we resort to modern unsupervised learning AD approaches like Isolation Forest and Histogram-based Outlier Score. Moreover, the self-diagnosis system (as shown on Figure 2) is composed of different procedures to adapt to the peculiarity of the MPFM structure. This instrument is made up of different metrology modules to allow redundancy and robustness in the flow estimation. In particular, a windowing is in place to allow continuous monitoring, and an ad-hoc feature extraction (FE) procedure designs quantities that capture the inter-module differences.

The motivations that pushed the authors to develop a new kind of FE will be analyzed in the first part of this section, while the implemented method will be described in the second part.

As explained in Section 2, the MPFM continuously measures different properties of the flow. The flow conditions (namely the composition and mass discharge) can vary widely over time and can be very different among different wells. According to literature (Brennen, 2005) the flow can be classified as *homogeneous*, *bubble*, *anular* or *slug flow*. When the type of flow evolves, also the relation between the measured quantities changes: data form new and different manifolds (Figure 3). For this reason the task of detecting anomalous patterns in the *measuring instrumentation* can be particularly challenging. The aim of the following method is to filter out the complexity given by the flow dynamics, leaving to the AD algorithm only the task of detecting anomalies in the instrumentation. In this way the algorithm works independently on the operating point.

The flowchart of the algorithm is presented in Figure 2. As previously stated in Section 2, the signals are continuously collected from different modules $\mathcal{X}^{(i)}, i \in \{1, \dots, M\}$. Later the signals are preprocessed in order to generate features that will be analyzed by an AD. The output is a stream of labels that tags every observation as outlier or inlier depending on the anomaly score.

3.1 Preprocessing

Since the algorithm works on fixed size batches, the signals need to be windowed. The windowing box splits the original signals in small overlapping intervals of size τ and time-overlap α.

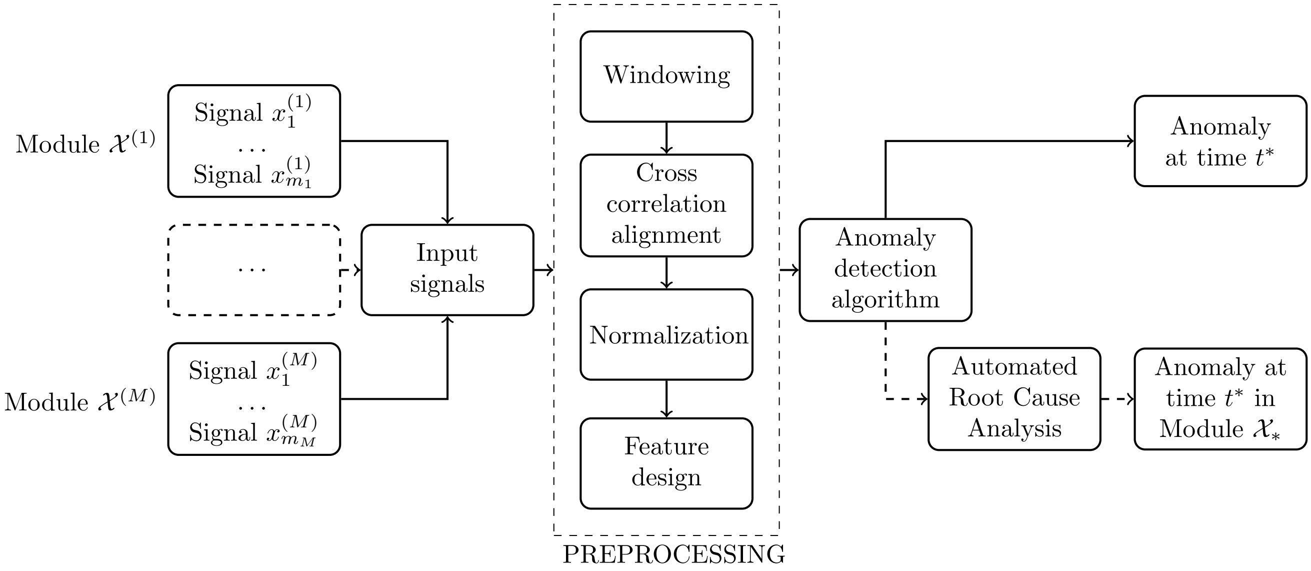

Module $\mathcal{X}^{(1)}$
Signal $x_1^{(1)}$
...
Signal $x_{m_1}^{(1)}$
...
Module $\mathcal{X}^{(M)}$
Signal $x_1^{(M)}$
...
Signal $x_{m_M}^{(M)}$
Input signals
Windowing
Cross correlation alignment
Normalization
Feature design
PREPROCESSING
Anomaly detection algorithm
Anomaly at time t^*
Automated Root Cause Analysis
Anomaly at time t^* in Module $\mathcal{X}_*$

Figure 2. Flowchart of the proposed algorithm. The Root Cause Analysis box is not discussed in this paper.

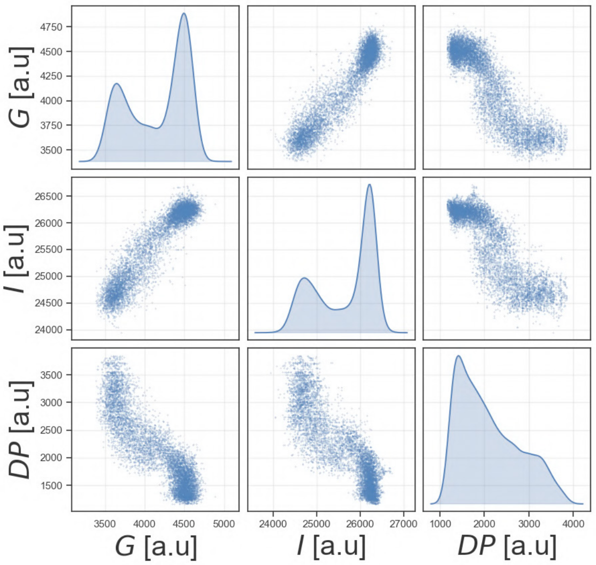


Figure 3. Example of scatter-distribution plot during slug flow conditions. Manifolds and clusters are typical of this condition, in this example there are *not* instrumental anomalies. G (Gamma), I (Impedance), DP (Pressure Difference) *time-aligned* signals. Cross correlation $K_{G,I} = 0.94$, $K_{G,DP} = -0.88$. The acronym *[a.u]* means *arbitrary units*.

The insight behind this novel feature extraction is that all the correlated modules see the same physical process from different points of view. An anomaly is likely to be present when the modules disagree: it occurs when the informative content of one signal differs from all the others. This intuition motivates the employed feature design approach that is based on two basic assumptions:

- all sensors at any given moment measure a different property of the same underlying physical phenomenon;
- the measurements from two different sensors are linearly (or approx. linearly) correlated.

Our assumptions are supported by visual inspection of the designed features (see Fig. 3 and Fig. 6) and by the effectiveness of monitoring such quantities with AD as proved by the experimental results reported in Section 4.

The efficacy of such feature design is obtained only when the modules measure the same event at the same time, namely when the signals are well aligned in time and the correlation between them is sufficiently high. To ensure such alignment we resort to a cross correlation procedure; such choice over more sophisticated time alignment algorithms like Dynamic Time Warping (Keogh and Ratanamahatana, 2005), was motivated by the computational resources constrains imposed by the MPFM.

3.2 *Feature Extraction*

The following step is the normalization of the windowed and aligned signals. In order to compare the features from different signal windows, we have defined a *reference* batch, whose median and median absolute deviation have been used to normalize the other batches. The reference batch is collected in controlled conditions and, thus, it does not contain anomalies.

In order to define features that encode the explained insights, we consider the difference between the normalized time series of all the modules. Therefore, the features z_{ij} are defined as:

$$z_{ij} = x_i - \mathrm{sign}(K_{ij})x_j \qquad z_{ij} = -z_{ji}$$

where x_i is a generic normalized windowed signal and K_{ij} is the correlation matrix between all the selected signals. The presence of the $\mathrm{sign}(\cdot)$ is necessary when the considered data are negatively correlated.

3.3 Anomaly detection algorithms

The MPFM self-diagnosis system is finally equipped with an AD algorithm that provides flags for anomalous conditions and an Anomaly Score, a quantitative index that defines the degree of anomaly of the considered data. The MPFM will provide the users and the maintenance operators, with useful information on the equipment health status.

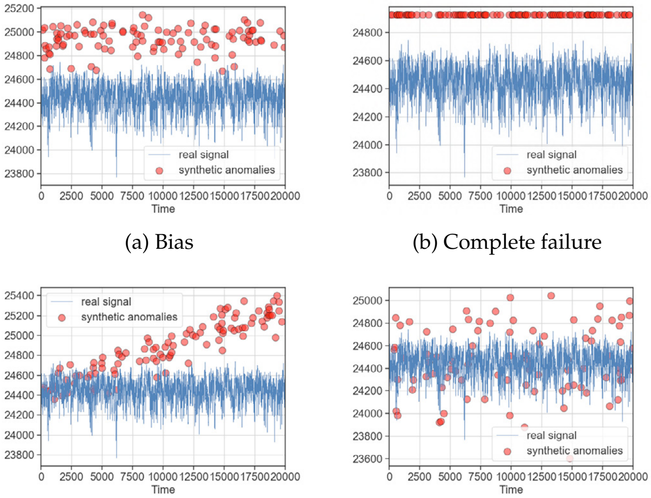
A typical approach in Fault Detection is to resort to classification methods, that rely on data tagged with faulty and un-faulty conditions. However, labelling data in real-world scenario is typically unfeasible, therefore we resort to unsupervised AD algorithms.

Four different AD algorithms have been considered in this work: (a) Cluster Based Local Outlier Factor (CBLOF), (b) Principal Component Analysis (PCA), (c) Histogram-based Outlier Score (HBOS) and (d) Isolation Forest (IF). A brief introduction to the aforementioned AD methods is reported in the following. Some AD approaches that require long evaluation time, like k-Nearest Neighbour, have not been considered since they are not suitable for the application at hand.

Unsupervised AD approaches can be divided into different families. One of these is represented by the so-called *density-based* AD approaches; the most famous approach is the Local Outlier Factor (LOF). The degree of anomaly in data is determined in LOF by taking into account the clustering structure in a bounded neighborhood of the observation (He et al., 2003). In this work we resort to an optimized extension of LOF called Cluster-based Local Outlier Factor (CBLOF) (He et al., 2003). A simpler density approach considered in this work is Histogram-based Outlier Score (HBOS) (Goldstein and Dengel, 2012): for each considered feature, an univariate histogram is constructed first. The frequency (relative amount) of samples falling into each bin is used as an estimate of the density (height of the bins).

Another family of AD approaches is represented by Isolation-based methods (Liu et al., 2012), whose Anomaly Score is defined by space partitioning. In Isolation Forest (IF) (Liu et al., 2012), the most famous Isolation approach, a procedure to partition the space based on random choices of variables and splitting points is in place; the procedure iterates until the observation under exam is isolated, i.e. no other observations are contained in the remaining space. The underlying idea of this approach is that outliers, being different from normal data, are easy to be isolated, while inliers will require a more complicated and longer procedure.

Finally a Principal Component Classifier-based approach (PCA) is considered; PCA is a well know procedure used for defying a set of uncorrelated variables, ordered by explained variability, from a dataset. In the approach proposed by (Shyu et al., 2003), an anomaly predictive model is constructed from the major and minor principal components of inliers. A measure of the difference of an anomaly from the normal instance is the distance in the principal component space.



(a) Bias

(b) Complete failure

(c) Drifting

(d) Precision degradation

Figure 4. Type of synthetic anomalies used to test the algorithm

The implementation employed in this work for the AD algorithms detailed in this Section is the one provided by (Zhao et al., 2019).

4. EXPERIMENTAL RESULTS

4.1 Synthetic data generation

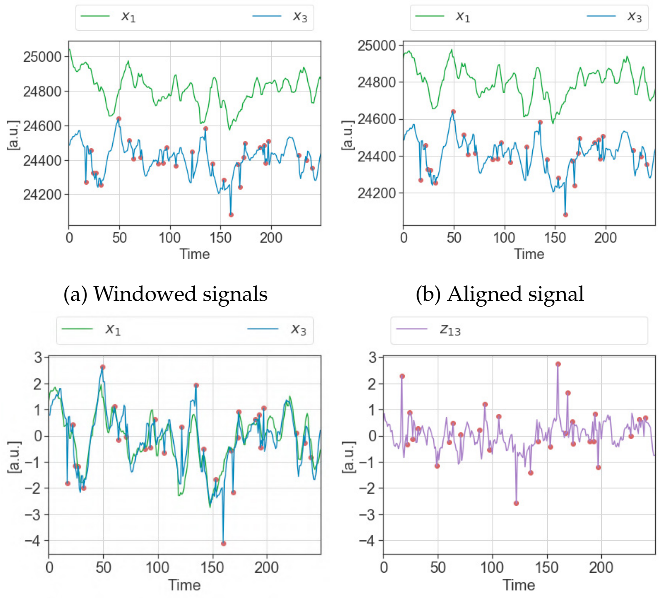
The algorithm explained in section 3 has been applied to a semi-synthetic dataset. It is composed of three sets of data collected in dedicated facilities for testing MPFM. A set contains 150 operating points of 3 million samples each. Even if we run unsupervised approaches, we need labeled data for evaluating the performances. Therefore we create a semi-synthetic dataset injecting anomalies inside the real data. In this way the added noise is used as a ’sure’ anomaly label, although ’natural’ anomalies may still be present in the data.

Four types of sensor failures are considered in the literature (Narasimhan and Jordache, 2000), namely bias, complete failure, drifting and precision degradation (see Fig. 4 where examples of synthetic faults are depicted). In the following, only the drifting case will be discussed due to the fact that this type of anomaly is the most frequent and relevant in the application at hand.

4.2 Feature extraction

Figure 5 shows the application of the FE procedure described in 3.1. We stress the effectiveness of the alignment procedure employed in the proposed MPFM self-diagnosis scheme: normal operating conditions of the equipment are in fact associated with high cross correlation, as it can be appreciated in the data reported in Figure 5a.

As said, one of the main benefits of the proposed preprocessing phase is the decoupling of (a) different operating conditions (associated with normal process physical dynamics), that the measuring system aims at detect-



(a) Windowed signals

(b) Aligned signal

(c) Normalized signal

(d) Feature extracted

Figure 5. Example of preprocessing on a synthetic dataset. Precision degradation of the blue signal. $z_{13} = x_1 - x_3$.

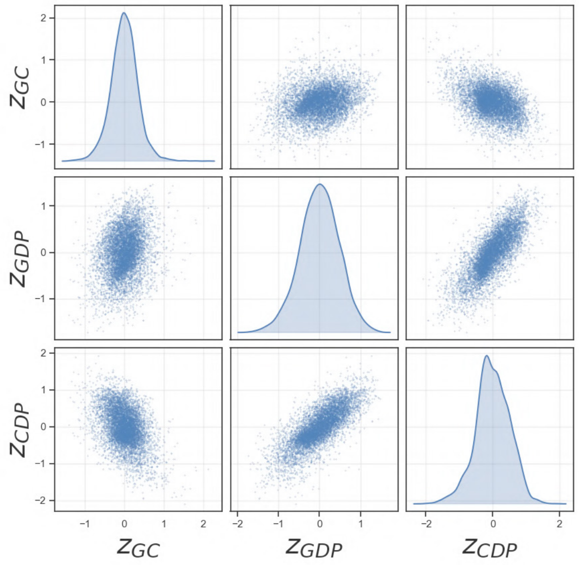


Figure 6. Feature extraction applied to the slug flow in Figure 3. The flow dynamics has been removed from data and the AD problem has become a lot simpler.

ing, and (b) anomalous measures, the target of our self-diagnosis system.

In fact, the effectiveness of the proposed pre-processing phase is demonstrated by Fig. 6 and Fig. 3. In presence of different operating conditions (for example, slugs or bubble flows), original data exhibit many clusters and complex manifolds but, exploiting the proposed features, the detection of instrumentation anomalies is greatly simplified and the influence of process physical dynamics disappears. By having a unique cluster, the work of AD algorithms is greatly facilitated; in particular IF will benefit from the proposed procedure, since it suffers for the presence of multiple clusters (Hariri et al., 2018).

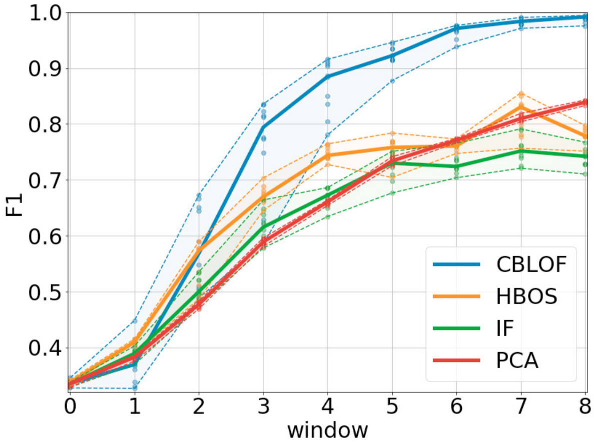


Figure 7. F1 score obtained by the different AD algorithms at the change of the evaluation window. The coloured dots show the score computed for every experiment repetition while the solid line represents the median. Maximum and minimum values are connected with a dashed line.

4.3 *Comparison of anomaly detection algorithms*

The classification ability of an AD algorithm can be summarized by the F-metric (Saito and Rehmsmeier, 2015). Figure 7 shows the $F1$-score computed for every window on a synthetic dataset where a drifting has been added to real experimental data (Fig. 4c): in our settings, the 20% of the data have an added degradation, while the rest are retained as the original data. The experiment has been repeated 10 times with the same settings. As expected, we notice an increasing evolution of $F1$ along the windows: the outliers are much more difficult to detect in the first window than in the last one since the drifting values increase over time.

On average HBOS and IF perform better in the first windows, but starting from window-2 CBLOF grows at much higher rate. The drawback of CBLOF is the high variance associated with its $F1$-score. PCA is always outperformed by the other methods until window-6. It's interesting to notice that IF trend stabilizes around window-5.

Table 1 shows all the metrics used to test the algorithms. They refer to the threshold that maximises the $F1$-score. Moreover, given the importance for the application at hand, *time complexity* has been reported, it was measured in seconds on a 2.7 GHz Intel Core 5 Processor; w.r.t. this metric it can be appreciated there is no significant difference between CBLOF and IF, while PCA and HBOS are much faster, with a reduced time complexity of 3 orders of magnitude.

Table 1. Performance metrics for the tested AD algorithms. These results pertain to one repetition done in the window-4 (see Fig. 7).

		CBLOF	HBOS	IF	PCA
Time complexity	[sec]	284.2	**0.7**	192.8	2.3
F1 score	[%]	**86.8**	74.7	66.8	66.0
Precision	[%]	**84.6**	61.7	51.6	59.1
Recall	[%]	89.2	94.8	**95.0**	74.9

5. CONCLUSIONS AND FUTURE WORK

In this work we have presented a Machine Learning-based approach to Anomaly Detection in Multiphase flow meters. The system allows the end user to promptly detect anomalous measures and to have an indication of measure reliability on historical data.

In particular the system exploits a preprocessing procedure that is very effective for distinguishing metrology anomalies within the typical physical dynamic of the underlying process. This model has been tested over semi-synthetic datasets, based on real data, with multiple AD algorithms and type of anomalies.

This study is the first step towards Multiphase Flow Meters able to self-diagnose their metrology modules. Our approach could be applied to other appliances that have many different but correlated modules, electric cars, batteries, redundant systems. The proposed approach has been designed for Plug & Play implementations, without the need of tuning the module for the well that hosts the MPFM.

As a future research direction, we will investigate different types of anomalies (ie. bias, complete failure) and different types of AD algorithms. In particular, we are currently investigating approaches for Root Cause Analysis that will allow the users and maintenance operator to move from Anomaly Detection to Fault Classification.

REFERENCES

AL-Qutami, T.A., Ibrahim, R., Ismail, I., and Ishak, M.A. (2017). Radial basis function network to predict gas flow rate in multiphase flow. In *Proceedings of the 9th International Conference on Machine Learning and Computing*, 141–146. ACM.

AL-Qutami, T.A., Ibrahim, R., Ismail, I., and Ishak, M.A. (2018). Virtual multiphase flow metering using diverse neural network ensemble and adaptive simulated annealing. *Expert Systems with Applications*, 93, 72–85.

Andrianov, N. (2018). A machine learning approach for virtual flow metering and forecasting. *IFAC-PapersOnLine*, 51(8), 191–196.

Brennen, C.E. (2005). *Fundamentals of Multiphase Flow*. Cambridge University Press. doi:10.1017/CBO9780511807169.

Corneliussen, S., Couput, J.P., Dahl, E., Dykesteen, E., Frøysa, K.E., Malde, E., Moestue, H., Moksnes, P.O., Scheers, L., and Tunheim, H. (2005). Handbook of multiphase flow metering. *Norwegian Society for Oil and Gas Measurement (NFOGM), Revision*, 2.

Falcone, G., Hewitt, G., and Alimonti, C. (2009). *Multiphase flow metering: principles and applications*, volume 54. Elsevier.

Goldstein, M. and Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, 59–63.

Hariri, S., Kind, M.C., and Brunner, R.J. (2018). Extended Isolation Forest. 1–11. URL `http://arxiv.org/abs/1811.02141`.

He, Z., Xu, X., and Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10), 1641–1650.

Keogh, E. and Ratanamahatana, C.A. (2005). Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3), 358–386.

Liu, F.T., Ting, K.M., and Zhou, Z.H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1), 3.

Meneghetti, L., Terzi, M., Del Favero, S., Susto, G.A., and Cobelli, C. (2018a). Data-driven anomaly recognition for unsupervised model-free fault detection in artificial pancreas. *IEEE Transactions on Control Systems Technology*.

Meneghetti, L., Terzi, M., Susto, G.A., Del Favero, S., and Cobelli, C. (2018b). Fault detection in artificial pancreas: A model-free approach. In *2018 IEEE Conference on Decision and Control (CDC)*, 303–308. IEEE.

Narasimhan, S. and Jordache, C. (2000). Data Reconciliation & Gross Error Detection. 1–57.

PietroFiorentini (2011). (access february 27, 2019), mpfm flowatch hs datasheet. URL `https://www.fiorentini.com/ww/en/product/components/mpfm_eng/flowatchhs/`.

Ristanto, T. and Horne, R. (2018). *Machine Learning Applied to Multiphase Production Problems*. Ph.D. thesis, MS Thesis. Stanford University.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), e0118432.

Shyu, M.L., Chen, S.C., Sarinnapakorn, K., and Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. Technical report, MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING.

Susto, G.A., Beghi, A., and McLoone, S. (2017a). Anomaly detection through on-line isolation forest: An application to plasma etching. In *2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 89–94. IEEE.

Susto, G.A., Terzi, M., and Beghi, A. (2017b). Anomaly detection approaches for semiconductor manufacturing. *Procedia Manufacturing*, 11, 2018–2024.

Theissler, A. (2017). Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. *Knowledge-Based Systems*, 123, 163–173.

Yan, Y., Wang, L., Wang, T., Wang, X., Hu, Y., and Duan, Q. (2018). Application of soft computing techniques to multiphase flow measurement: A review. *Flow Measurement and Instrumentation*, 60, 30–43.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*.