Machine learning application for sensor failure detection in polymerization process

Amaro A. de Lima, Gabriel M. Araujo, Igor S. Oliveira e Bettina D. Barros

***Abstract*— This work analysed the time signals from 5 instruments distributed in an industrial polymerization facility: two temperature instruments, a water level instrument, a weight instrument and a flow instrument. The dataset is composed by a five years history with a sample rate of 1 minute. A specialist using the event related industry report labelled the signals. The results using random forest as the machine learning classifier reached significant performance in detecting failures independent of the instrument, preliminary indicating the suitability of the framework.**

***Keywords*— Machine learning, failure, instrument, polymerization.**

I. INTRODUCTION

Nowadays almost all of the industrial systems make use of some measurement instrumentation in their production process. A long with this technical evolution appears the need to ensure that these indications are reliable, avoiding loss in production or even an accident at work. An unreliable instrument can provoke a mistaken decision leading to unexpected production outages. Unplanned process halt is directly related to losses in time, raw material and production. Many industrial facilities are huge and complex making the characterization of several kinds of failures by the machine operators a very expensive and sensible to errors task.

Fault diagnosis and detection is an important tool of Process Engineering and is the main topic of abnormal events management. The early detection and diagnosis of faults, while the industrial process is still in operational conditions could avoid the escalation of anomalous events and the reduction of production loss. The estimated annual loss due to faults in industrial instrumentation is about 20 billion dollars a year. Therefore, a great interest, from the academic researchers to industry workers, came up in this area, which was not observed in the past decades.

The polymerization industry had a remarkable growth all over the world in the past decade, for instance; only in Brazil the consumption of Polyethylene terephthalate (PET) resin, the main manufacture of a polymerization industry, had an increase of 2200%. In 2005, it was estimated that 32% of this polymer world production was designed to the manufacture of food packing and bottle, and the remaining 68% was destined to the production of polyester staple fibre, which is a component widely used in making fabric, clothes, shoes, upholstery furniture as many other products [1]. Only in the state of São Paulo (Brazil) the polymerization industry raised about 2.2 billion dollars in 2005 [1]. Considering the importance of the polymerization industry in modern life products and economy the mitigation of instrumentation failures in this industry, which could also be related to the machinery associated to the process as pumps, fans, motors and many others, is of fundamental relevance and will be addressed in this work as a case study.

This work analysed the time signals from 5 instruments distributed in an industrial polymerization facility; 2 temperature instruments to monitor the water temperature in the cooling process going through the pumps; a water level instrument to check if the water reaches an electronic component; a weight instrument to evaluate the amount of manufactured products; and a flow instrument to measure the fluid flow through the pipes. The data was collected in a 5 years history of all instruments with a 1 minute sample rate simultaneously. A specialist using the event related industry report labelled the signals.

The failure detection system has four steps and the classification, performed by using random forest as the machine learning classifier, reached over than 98% of global accuracy in detecting failures independent of the instrument.

The remaining of this paper is organized as follow. Next section some related works are presented. A detailed description of the proposed system, as well as the dataset and the failure characterization is in Section III. Results and discussions about them are in Section IV. In Section V we have the conclusion.

II. RELATED WORKS

The work [2] describes a method that uses Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) to evaluate the conditions of mineral filter dispersion in a polymer matrix. Ultrasound instruments, pressure instrument, thermocouple and the electrical current of the extruder motor drive were the variables monitored by the system. The PCA algorithm were used to reduce the dataset dimensionality before applying to the ANN classifier aiming to determine the dispersion state of the filter. The system reached 10% chance of estimating dispersion index with error larger than 0.05 obtaining an efficient performance.

The approach proposed in [3] describes a technique to deal with multivariate high dimensional data in industrial applications. The idea is to use a Genetic Programming (GP) based algorithm that has built-in a mechanism to select the variables related to the problem during the simulated evolution and gradually ignore the variables that are not, which could possibly facilitate the industrial data analysis and applicability.

An extensive dataset for process monitoring focused in fault detection and identification is proposed in [4]. Although

Amaro A. de Lima, Gabriel M. Araujo e Igor S. Oliveira, Centro de Educação Tecnológica Celso Suckow da Fonseca, Nova Iguaçu - RJ, E-mails:{amaro.lima, gabriel.araujo}@cefet-rj.br, igorbraskem@gmail.com.

Bettina D. Barros¸ Programa de Engenharia Elétrica Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, E-mails:barrosbettina@gmail.com.

its study case is penicillin production, different from ours that is polymer production. The material draws interesting aspect on the dataset to consistently evaluate the detection and identification aspects in a system production.

A study on anomaly detection from instrument data applied to petroleum industry applications is addressed in [5]. It proposes the usage of One-Class Classifier (OCC) combined to Yet Another Segmentation Algorithm (YASA) to efficiently detect anomalies in turbomachines, which are machines, usually found in oil platforms, that are responsible for electrical power generation. The work compared the efficiency of the proposed approach with several techniques in a qualitative way, without evaluating numerically the performances differences.

A very interesting work that has some common aspects to ours is presented in [6]. It is focused in big data problem that is related to heterogeneous multi-instruments acquisition, high dimensional and large amount of data to monitor a polymer production plant. The work investigates the applicability of novelty filtering, anomaly detection and One-Class Classifier (OCC) to detect abnormal behavior in production lines that could be due to instrument fault, environmental influences, and unexpected raw material properties' deviations. The system obtained 99% accuracy using OCC for 160 instruments data collected during 3 months with 1 sample per minute.

III. Proposed System

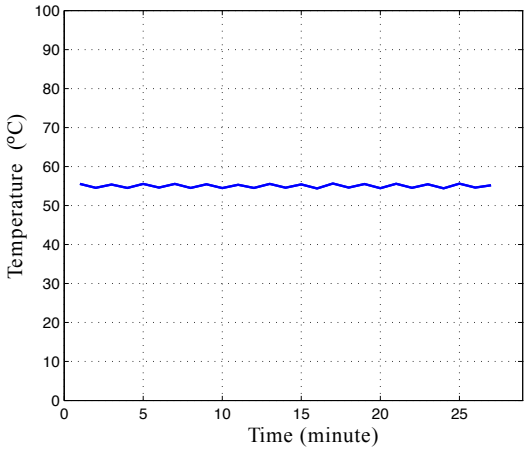
A. Dataset

The dataset used in this work was acquired from a polymerization plant through PI system from OSI®. The dataset was composed by the signal from the five instruments: FT-80001, LZT44032, TT-25867, TT23002C, WT-25005. Each signal is composed by $3,153,285$ samples from 02-Jan-2010, 02:00:00h to 31-Dec-2015, 20:44:00h. The sample rate is $1\ \mathrm{sample/min}$. So, the whole dataset is composed by two matrices with size 3153285×5. One is the data from the instruments organized in a matrix, whose each column is a signal from a instrument. The second matrix is the respective failure label, that was manually annotated by a specialist. This dataset is highly unbalanced, since a failure of any type (as described in Section III-B) is a rare event.

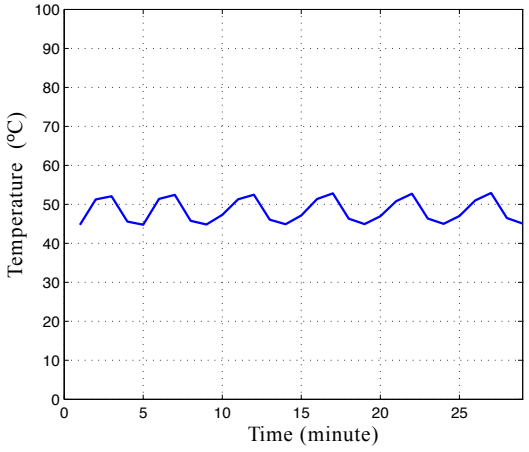
B. Failure characterization

In this paper, a Fault is defined as an undesirable event in an instrument or process which can lead the plant to instability or loss condition. It worth mention that the faults characterized here in this section are only part of the types of fault that can occur. However, the objective is to illustrate the main types of faults found in this case of study.

1) Inadvertent change in a model's parameter: When a control parameter is modified several disorders can occur, which lead to instability in the process. In Figure 1 (a), one can observe a normal behaviour in temperature indication of the water used to transport the grains formed in the process of polymer extrusion. In Figure 1 (b), there is another example of the same variable, now with an abnormal behaviour, after a change in the parameters of the PID controller by an operator.



(a)



(b)

Fig. 1. Examples of fault due to an inadvertent change in the parameters of a PID controller by an operator: (a) normal behaviour in the signal from the temperature instrument TT25867; (b) abnormal behaviour in the signal from the same instrument after the change.

2) Structural faults: Structural faults are related to hardware elements from a process, such as: failure in the controllers hardware, broken ducts, stuck valves, leaks, water in the electronics of a field instrument etc. Figure 2 shows the moment that a level instrument fail. It was verified later the presence of rain water in the electronics compartment. One can note that the indication, as a percentage of the maximum level, rapidly goes to the end of scale.

3) Malfunctioning in instruments or actuators: Some instruments provide essential signals to plant control. A fault in one of those can potentially lead the plant to a condition beyond a acceptable, unless detected in time. The speed in detecting such a failure can severely affect the plant performance. In Figure 3 (a), there is an example of a temporary failure in the instrument measuring the temperature in the polymerization reactor. In Figure 3 (b), one can note the moment that an actuator, which controls the opening of a valve that feeds an weighing scale, stuck in open mode. At this very moment, a excessive load of product is added to the weighing scale, increasing mass indication.

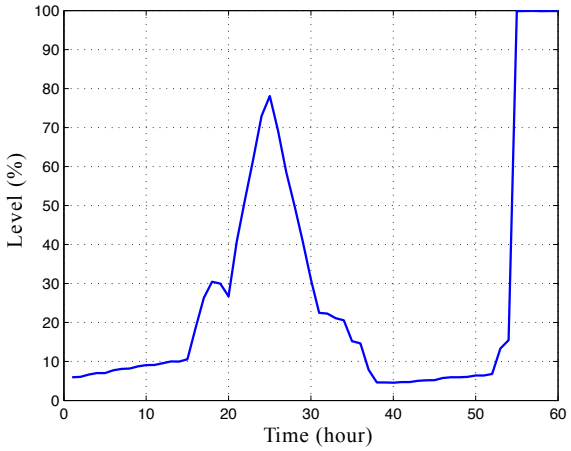
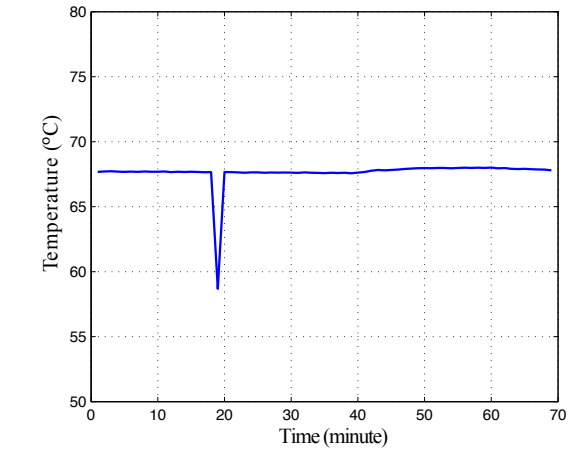
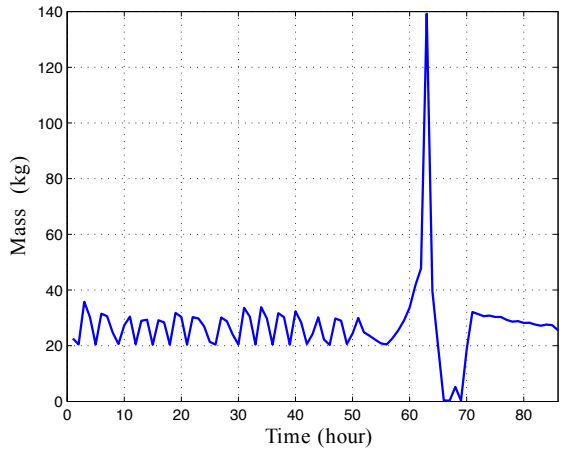


Fig. 2. Level indication from a LT-44032 (level instrument) goes to the end of scale due to rain water in the compartment containing the electronics of the instrument.



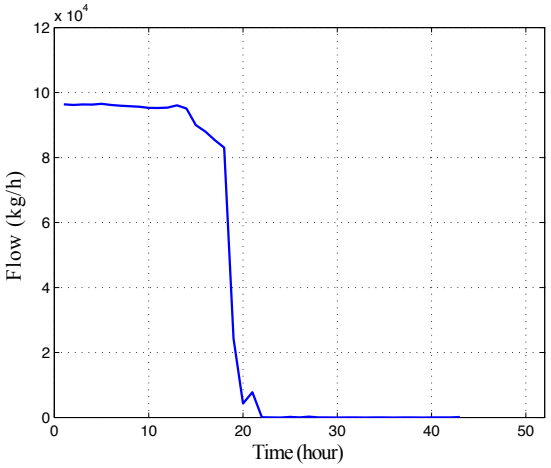
(a)



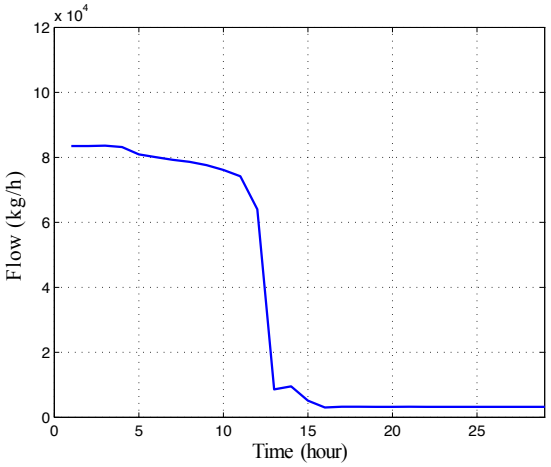
(b)

Fig. 3. Instrument failure examples: (a) temporary failure in temperature instrument TT-23002C; (b) product overweight due an actuator failure in instrument WT25005.

4) Zero shift error: There are faults related to a malfunctioning in some instrument. In Figure 4 (a) there is a signal from a flow meter in normal condition. After stopping the product flow, it indicates a zero mass flow. In Figure 4 (b), a bad contact in inner connections of a flow meter lead to an elevation in instantaneous mass flow indication. As a consequence, it caused an inventory error in a liquid gas sphere. It could illustrated by the discharge from a ship, in which is necessary to assess the inventory coming from it through the total of mass flow measurements. After the end of the operation the flow should stabilize in zero. This is the expected behaviour whenever there is an interruption of product flow from the ship to the plant. However, when the flow is interrupted and the instrument holds an instantaneous measurement different from zero. It causes a difference in the total of the inventory of the ship in the next discharge, since the difference to zero of the flow indicated in the last one is summed to the instantaneous measurements in the next discharges.



(a)



(b)

Fig. 4. Failure example in mass flow instrument FT80001A: (a) mass flow meter in normal condition; (b) mass flow meter with zero shift error.

5) External failure: There are some cases in which the failure is related to external factors, since those ones are not in the plant domain. In Figure 5, one can see that a transient in the supplied energy caused a failure in the frequency inverter that feeds a recycle pump in a polymerization reactor. This

failure could be detected through the variation in the current indication from a pump motor.

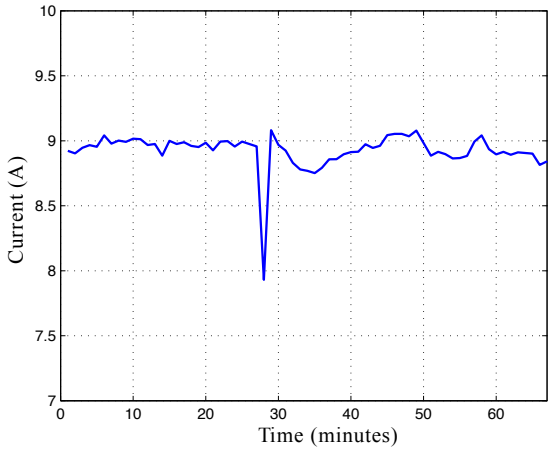


Fig. 5. Variation in instantaneous current in electrical feed of a pump motor using instrument IT23004.

C. Proposed system

The proposed system is depicted in Figure 6. Each part of the system is detailed in the next subsections.

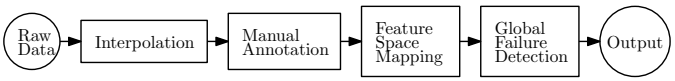


Fig. 6. Block diagram of the proposed system.

1) Interpolation: One of the characteristics of the raw data extracted from PI System® is that each instrument has its own sample rate. As a matter of fact, in order to save storage, this sample rate is not constant. Since each sample has a time-stamp related to it, we employed a linear interpolation to have all instrument with the same constant sample rate.

2) Manual annotation: Once all interpolated signals from each instrument has the same constant sample rate with aligned time-stamp, an specialist in the plant analysed in this work manually annotated each failure according to the types of failures characterized in Section III-B. Each instrument has a binary vector with the same size of the signal. All the elements of these vectors are 0 but in the places in which the indexes are related failures in its respective instrument. In these places of the vector label there is a 1.

3) Feature space mapping: The interpolated data is windowed by a N-hour rectangular window with 50% of overlap. In this scheme, each sample is $N \times 5$ matrix, since our dataset has 5 instruments. If a window has more than 50% of its samples labelled as failures, then it is considered as belonging to fault classes. Otherwise it is considered as belonging to no-fault class. The number of samples depends on the size of the window N. The greater the window, the smaller the number of samples.

4) Failure detection: The classification step is performed by an ensemble learning method known as random forest [7] [8]. In this method, a set of decision trees are learned by using an ensemble-meta algorithm known as bootstrap aggregating (also known as Bagging [9]). In Bagging, random subsets of the training set are used to train several instances of a classifier which are combined to provide an improved classification with reduced variance and overfitting. In random forest, each split in each Decision Tree growing algorithm uses only a random subset of the features. In other words, a Bagging of features is performed in learning process of each tree of the classifier. In doing so, the most important features for the desired class are going to be selected in many of the B trees. In this work, we used the default parameters of Matlab implementation of random forest with 100 trees. The number of trees was empirically obtained.

IV. Results and Discussion

In order to assess the performance of the proposed method we employed a k-fold cross-validation with 10 folds. In other words, the data was partitioned in 10 parts, in which 9 were used in training and 1 in test. This process was repeated up to all 10 parts were used as test set.

A. Experiment 1

In this experiment, the time label vector, which is a vector comprising the labels of fault (F) and no-fault (NF) for the whole time period of 3,153,285 minutes, is windowed by a 24-hour rectangular window with 50% of overlap. If a specific window sum more than 50% of the samples labelled as fault, the window is also labelled as fault. The same is applied to no-fault labelling. This settings generated 590 fault and 3787 no-fault windows. Each feature vector consists of the concatenation of the 5 instruments data related to the time of the corresponding window making a 7200 dimension vector, which is a 24 hour (1440 minutes) multiplied by 5 instruments.

Table I shows the confusion matrix of the experiment 1 using window of 24 hours achieving the mean accuracy over the folds of 98.9% and 91% in detecting the no-fault and fault elements, respectively. It represents an overall system accuracy of 97.8%.

TABLE I
EXPERIMENT 1 - CONFUSION MATRIX OF THE NO FAULT × FAULT DETECTION PRESENTING THE CLASS ACCURACY AND ITS RESPECTIVE STANDARD DEVIATION USING A 24-HOUR WINDOW.

Target	Output classes (%)	
	No Fault	Fault
No Fault	$\mathbf{98.9 \pm 0.56}$	1.1 ± 4.66
Fault	9 ± 0.56	$\mathbf{91 \pm 4.66}$

B. Experiment 2

In this experiment, the time label vector is windowed by a 12-hour rectangular window with 50% of overlap following the same rationale as applied to experiment 1. These settings generated 1187 fault and 7570 no-fault windows and vector size of 3600 dimensions.

Table II presents the mean accuracy and standard deviation over the folds in the format of confusion matrix for the experiment 2, which uses a 12-hour window. The accuracy

of no-fault, fault classes and overall system are 99%, 91.7% and 98%, respectively. Compared to the previous experiment the individual classes and the overall accuracies had some improvement. Furthermore, it presented a reduction in the standard deviation indicating a more precise experiment, i.e., the fold accuracies vary not so largely from the overall mean.

TABLE II
EXPERIMENT 2 - CONFUSION MATRIX OF THE NO FAULT × FAULT DETECTION PRESENTING THE CLASS ACCURACY AND ITS RESPECTIVE STANDARD DEVIATION USING A 12-HOUR WINDOW.

Target	Output classes (%)	
	No Fault	Fault
No Fault	$\mathbf{99 \pm 0.47}$	1 ± 2.49
Fault	8.3 ± 0.47	$\mathbf{91.7 \pm 2.49}$

C. Experiment 3

The experiment 3 applies a 6-hour rectangular window with 50% of overlap to the time label vector generating 2400 fault and 15166 no-fault vectors with dimensionality 1800.

The mean accuracy and standard deviation over the folds in the format of confusion matrix for the experiment 3, which uses a 6-hour window, are detailed in Table III achieving 99.3%, 92.8% and 98.4% of accuracy for no-fault, fault and overall classes, respectively. Once again, comparing to experiments 1 and 2, the individual classes and the overall accuracies got some improvement, while the fault class had a slightly reduction in standard deviation. However no precision change was observed in no-fault class.

TABLE III
EXPERIMENT 3 - CONFUSION MATRIX OF THE NO FAULT × FAULT DETECTION PRESENTING THE CLASS ACCURACY AND ITS RESPECTIVE STANDARD DEVIATION USING A 6-HOUR WINDOW.

Target	Output classes (%)	
	No Fault	Fault
No Fault	$\mathbf{99.3 \pm 0.48}$	0.7 ± 2.20
Fault	7.2 ± 0.48	$\mathbf{92.8 \pm 2.20}$

D. Experiment 4

A 1-hour rectangular window with 50% of overlap is used in the experiment 4 and is applied to the time label vector generating 16066 fault and 89041 no-fault feature vectors with dimensionality 300.

The confusion matrix for the experiment 4 representing the mean accuracy and standard deviation over the folds using a window of 1 hour are shown in Table IV achieving 100%, 87% and 98% of accuracy for no-fault, fault and overall classes, respectively. Comparing this performance to the ones in experiments 1, 2 and 3, only the no-fault class presented an improvement in accuracy. The fault and overall classes had a significant accuracy reduction. The system precision was ever better than the ones presented in previous experiments, with lower variation around the mean accuracy.

The experiments 1, 2, 3 and 4 show the performances in detecting the fault and no-fault classification cases with different time length windows. Comparing the results, performance and precision improvement are observed while the window size is reduced, a partial exception is the experiment 4, where only the fault and overall accuracy do not demonstrate it. This search for the appropriate window size is intimately related to the physical statistics associated to the industrial process indicating that no-fault and fault are better statistically represented with smaller windows. However, the interval of windows from 6 to 1 hour should be better investigated, because windows smaller than 1 hour seem to affect negatively the fault class accuracy.

TABLE IV
EXPERIMENT 4 - CONFUSION MATRIX OF THE NO FAULT × FAULT DETECTION PRESENTING THE CLASS ACCURACY AND ITS RESPECTIVE STANDARD DEVIATION USING A 1-HOUR WINDOW.

Target	Output classes (%)	
	No Fault	Fault
No Fault	$\mathbf{100 \pm 0}$	0 ± 0.66
Fault	13 ± 0	$\mathbf{87 \pm 0.66}$

V. CONCLUSION

In this paper we propose describe a failure detection system. A polymerization plant was used a case of study. The failures was detected by analysing signals from 5 instruments in this industrial plant. Each signal is composed by $3{,}153{,}285$ samples from five years, which gives a $1\ \mathrm{sample/min}$ sample rate. A specialist characterized the types of failures and manually annotated them. In the classifications step of the proposed system, we employed random forest and obtained a global accuracy of 98.4% in detecting failures independent of the instrument. As future work, other types of classifiers are going to be employed and the problem of failure classifications (in order to determine which failure just occurred) is going to be addressed as well.

REFERENCES

[1] Evangelista, V. F.: Modelagem e simulação do processo industrial de polimerização em estado sólido do poli(terftalato de etileno) e do nylon 66. PhD Thesis. Chemical Engineering Program, Federal University of Rio de Janeiro (COPPE/UFRJ) (2010).

[2] Sun, Z., Yan, J., Jen, C., Chen, M.: Application of Principal Component Analysis and Neural Networks in the Determination of Filler Dispersion during Polymer Extrusion Processes. In: 4th IEEE International Conference on Control and Automation (2003).

[3] Smits, G., Kordon, A., Vladislavleva, K., Jordaan, E., Kotanchek, M.: Variable Selection in Industrial Datasets Using Pareto Genetics Programming. Genetic Programming Theory and Practice III. Genetic Programming, vol 9, 79–92 (2006).

[4] Impe, J.V., Gins, G.: An Extensive Reference Dataset for Fault Detection and Identification in Batch Processes. Chemometrics and Intelligent Laboratory Systems. 148, 20–31 (2015).

[5] Marti, L., Sanchez-Pi, N., Molina, J.M., Garcia, A.C.B.: Anomaly Detection Based on Sensor Data in Petroleum Industry Applications, Sensors. 15, 2774–2797 (2015).

[6] Kohlert, M., Konig, A.: Large, high-dimensional, heterogeneous multi-instrument data analysis approach for process yield optimization in polymer film industry. Neural Computing and Applications. 26, 581–588 (2015).

[7] Tin Kam, H.: Random Decision Forests. In: 3rd International Conference on Document Analysis and Recognition (2015).

[8] Tin Kam, H.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8), 832–844 (1998).

[9] Breiman, L.: Bagging predictors. Machine Learning. 24 (2), 123–140 (1996).