CHEMICAL ENGINEERING TRANSACTIONS

VOL. 103, 2023

A publication of AIDIC

The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Petar S. Varbanov, Panos Seferlis, Yee Van Fan, Athanasios Papadopoulos
Copyright © 2023, AIDIC Servizi S.r.l.
ISBN 979-12-81206-02-1; **ISSN** 2283-9216

DOI: 10.3303/CET23103096

A Novel Approach to Recognize Hidden Machinery Malfunctions Based on Operational Data

Georgios Gravanis[b,*], Dimitrios Trigkas[c], Konstantinos Diamantaras[b], Spyros Voutetakis[a], Simira Papadopoulou[a]

[a]Chemical Process & Energy Resources Institute, Centre for Research and Technology Hellas (CPERI/CERTH), P.O. Box 60361, 57001 Thermi, Thessaloniki, Greece
[b]Department of Information and Electronic Engineering International Hellenic University, 57001 Thessaloniki, Greece
[c]Department of Industrial Engineering and Management, International Hellenic University, 57001 Thessaloniki, Greece
ggravanis@certh.gr

This work introduces a novel approach that identifies hidden operation malfunctions of complex machinery equipment. The architecture of the proposed approach is based on the $k\ Nearest\ Neighbor$ (kNN) unsupervised algorithm for anomaly detection, along with the φ_k correlation coefficient, and aims to provide information on whether systematic malfunctions exist on the equipment. The proposed approach can provide insights into the root cause of the malfunctions. For the evaluation of the proposed approach, a case of a food industry packaging machine is studied. The rejections of the packaged products can be used as a metric for defining the operational efficiency of the machine. The anomaly detection task is performed for each variable that monitors the operation of the machine. Then, the correlation coefficient between the detected anomalies and the rejections of the machine is calculated. The higher the correlation coefficient, the most probable cause for the rejections is a malfunction of the machine's subsystem that the specific measured parameters monitor. A semi-synthetic dataset based on real operational data was created to conduct experiments, and high-accuracy results were obtained.

1. Introduction

During the 4th Industrial Revolution and the blooming of industry digitisation, well-established and reliable systems for the monitoring of the machines and processes, such as Supervisory *Control and Acquisition Systems* (SCADA), *Programmable Logic Controllers* (PLC), and *Human Machine Interfaces* (HMI) are enhanced with intelligent systems. Amongst others, such intelligent systems can be used to assist in the monitoring and control of plants, or to help define the maintenance strategy for the equipment. More specifically, several works use Machine Learning methods to detect anomalies in operation of industrial assets and provide useful insights to the user to act accordingly. Monteira et. al, (2022) introduce a deep learning framework to detect bearing faults with high accuracy, while Hendrickx et al. (2020) propose a framework using clustering algorithms to monitor the operating condition of similar assets i.e. motors, inside a factory. Leveraging those insights, the maintenance strategy of the machinery can be transformed to be more efficient in both economic and work uptime terms. More specific, the Run to Failure (RtF) and the Preventive Maintenance (PM) strategies which are the most common used, can be complemented by the Predictive Maintenance (PdM) strategy as described by Zonta et al. (2020).

However, most related studies target in monitoring rotating equipment such as bearings, motors and turbines without taking account the overall condition of the machine. Pittino et al. (2020) have also recognised that despite the abundance of anomaly detection studies targeting in assets of manufacturing processes, little research has been conducted about machinery operations.

Using control statistics such as control charts, or utilising SCADA systems by setting thresholds and alarm rules, is a way to monitor the overall health of machinery. However, this approach has limitations when applied in complex systems (Kamat and Sugandhi, 2020). Another important aspect is that malfunctions in complex

Paper Received: 17 April 2023; Revised: 19 June 2023; Accepted: 23 July 2023
Please cite this article as: Gravanis G., Trigkas D., Diamantaras K., Voutetakis S., Papadopoulou S., 2023, A Novel Approach to Recognize Hidden Machinery Malfunctions Based on Operational Data, Chemical Engineering Transactions, 103, 571-576 DOI:10.3303/CET23103096

machinery do not always occur because of maintenance issues. Such machinery usually has lots of parameter micro-tuning that is done by experienced personnel. It is very common, though, for small differentiations in tuning to cause malfunctions on the machinery, usually with increased production cost.
This work proposes a framework that produces a "health" report about the machinery operation. The framework is based on unsupervised methods such as the kNN algorithm to overcome the above-mentioned limitations.
This paper is organised as follows: Section 2 presents the architecture of the proposed framework, while a short presentation of the algorithms used is presented in Section 3. Section 4 describes the rules followed for the creation of the semi-synthetic dataset. Next, Section 5 presents the detection process along with the experimental results. Finally, Section 6 discusses the results of this study.

2. Architecture

Let us assume that we have a production machine, as depicted in Figure 1. To have the final product in the output of the machine, a list of actions is required to be followed. Those actions are implemented by the relevant subsystems where several parameters have been set and monitored through sensors. Finally, before releasing the product to the market, several tests are performed to ensure that the product meets certain quality standards.

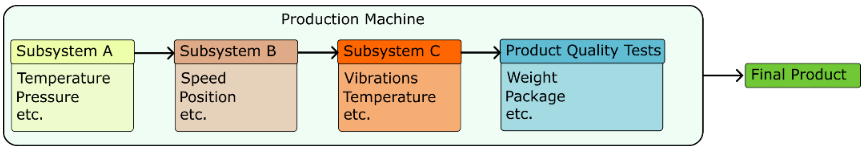
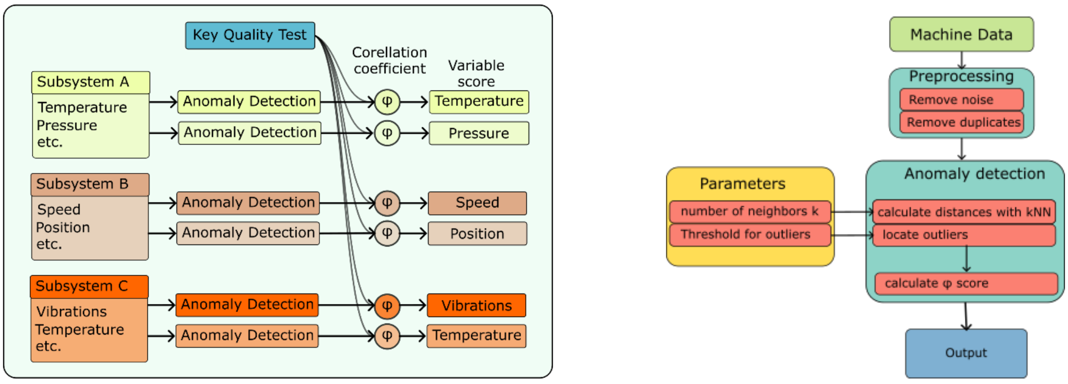


Figure 1: Production machine example case

The proposed framework (Figure 2) aims to identify any hidden malfunctions that cause rejections of the ready-to-market product. A very important aspect of this approach is to define the key quality test that best represents the operational health of the machine. The proposed framework detects any anomalies in the subsystem's operation. As soon as the key is defined and the anomalies are detected, a correlation coefficient is applied to reveal the reason for the malfunction. If there is a high correlation score between the anomalies detected in a parameter and the key quality test, then it is probable that the rejections are caused by the subsystem that the parameter belongs.



(a) (b)

Figure 2: (a) The architecture of the proposed framework. (b) The data pipeline followed inside the framework to obtain insights about the machine operation

The algorithm utilised for the anomaly detection is the *kNN,* while the correlation coefficient used in this study is the φ_k. The *kNN* is used due to its simplicity as it is categorised as a non-parametric machine learning algorithm, and in our case, it can also capture the dimension of time. Concerning the algorithm's complexity, the simplest but computationally expensive method used to search the neighbors, is the *brute force* that scales as *O(ndk)* where O(.) is the algorithm's complexity, *d* is the dimension of the dataset, *n* is the number of samples and *k* is the number of nearest neighbors. In the proposed framework, the kNN is applied in time series data with dimension one (d=1). With that consideration, even the brute force method can produce the results effectively with low computational cost. Subsequently, the implementation can be easily scaled up in order to monitor in parallel more subsystems of a single machine, or even more machines in a plant. In the example case of this study the key quality test is a pressure switch that produces binary results. Thus the φ_κ correlation coefficient is used because it is capable to calculate the correlation between binary vectors. Both algorithms are briefly explained in the next Section. Concerning the parameters of the proposed framework, as depicted in Figure 2b, those are the number of the neighbours and the threshold for the outlier definition. The combination of those parameters is used to detect the outliers of the monitored variables. Both parameters are dependent on the nature of the problem and should be defined after a sensitivity analysis of the problem studied.

3. Algorithms

3.1 k-Nearest Neighbours (kNN)

Since the proposed framework targets to provide insights of complex equipment operation status, where problems in operation may occur by numerous faults and situations, the basic architecture should rely on an algorithm that can detect outliers without having previous knowledge about operational data distribution. A well-established algorithm with great performance for this task is the kNN algorithm.

The kNN algorithm is a non-parametric method for classification problems and was introduced by Cover T. and Hart P. (1967). The algorithm processes the k nearest points of a dataset *i.e. k-neighbours*. When used in classification problems, the algorithm labels the instances of the dataset according to the distance between the k neighbours. Several distance metrics such as the *Manhattan,* the *Minkowski* and the *Euclidean* (Eq(1)) can be utilized in the algorithm, with the latest to be the one used in this work.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

Except for classification and regression tasks, the *kNN* algorithm has been proposed also for outlier detection tasks. Knox and Reymond (1998) proposed a definition for outliers that overcomes the need to have a human expert define several parameters and evaluate the results. Later, Ramaswamy et al. (2020) proposed a definition for outliers given that a user is generally interested in the top n outliers of a dataset. Their definition is based only on the distance of the k^{th} nearest neighbour of a point p which is denoted as $D^k(p)$. Given the parameters k and n, a point p is defined as an outlier, if no more than $n-1$ other points in the data set have a higher value for D^k than p. To make it simpler, the top n points with the maximum D^k values are considered as outliers and are denoted as D_n^k.

3.2 φ_k correlation

As described in Section 2, to identify any hidden and systematic malfunction of the machine, a coefficient that measures the correlation between the detected outliers in each monitored parameter with a key quality test should be used. A standard approach to measure the correlation between two variables is Pearson's correlation coefficient. However, the Pearson correlation coefficient usually is applied between continuous variables. In this framework, both the outliers and the key quality test are categorical (binary) vectors. A more suitable correlation coefficient should be found to reveal any systematic malfunction of the machine. Baak et al. (2020) proposed the φ_k correlation coefficient that is based on *Pearson's x^2 test* and is able to handle categorical variables. According to the authors, φ_k can capture non-linear dependencies, and it reverts to the Pearson correlation coefficient in the case of a bi-variate normal input distribution.

4. Dataset

The proposed framework is developed to be used in a real-case scenario of a large dairy production unit in Greece. The target is to minimise the rejections of a fully automatic yoghurt packaging machine. Operational data were collected from the HMI and the PLC of the machine to perform the necessary analysis. Table 1 presents the operation parameters monitored and collected from the machine. The collection of the data is performed with a sampling rate of one second while the machine has a production step of eight cups. After the data analysis, it is shown that the sampling rate is much faster than the production speed, which leads to oversampling. For that reason, the collected data were transformed according to the production step. This allows the framework to monitor each production step separately.

As described previously the machine has a production step of eight cups. Practically, this means that the filling, the sealing and the sealing test is performed in batches of eight in the relevant subsystem of the machine.

For this work, the key production quality test is defined as the cRejected parameter which is cumulative of the cRejectHead(1-8). The latest parameter monitors whether the aluminium foil is sealed with the cup. Concerning the subsystem parameters that the framework examines, those are the eight operation temperatures, one for each sealing head. The expected result is to reveal any hidden malfunctions, i.e. sealing issues caused by temperature anomalies that lead to rejections of ready products. Additionally, the insights produced by the proposed framework can be used to optimise production planning and schedule machine maintenance. Unexpected breakdowns that cause production downtime could be avoided.

Table 1: Machine parameters monitored and stored for the dataset creation

Parameter	Description	Unit
cRejected	Number of total rejected cups	Integer
cRejectHead (1-8)	Number of rejected cups per head	Integer
cProduced	Number of total produced cups	Integer
mPosition	Machine position	Degrees
sTemp (1-8)	Temperature setpoint in heat sealing head 1-8	ºC
oTemp (1-8)	Temperature operation in heat sealing head 1-8	ºC
sSpeed	Machine speed setpoint	%
iSpeed	Machine operation setpoint	%
sWeight	Quantity setpoint for head	g

4.1 Dataset enhancement

During the data collection period, no serious issues were presented in the machine. To validate the proposed method, a semi-synthetic dataset was created. This dataset is based on the production data collected and enhanced with anomalies and rejections. The dataset enhancement is performed following a certain procedure, as described next.

First, a parameter of the machine is randomly selected, and the distribution and the basic statistics, such as mean value and standard deviation are calculated. Then, a random quantity of outliers between 1 % and 3 % of the production is generated. Since the area between 3σ and 4σ is commonly used to define the outliers in a distribution, the same area is used to obtain the values of the outliers, while the distribution side is randomly selected. Additionally, the outlier is noted as a rejection in the *cRejected* parameter vector. Figure 3 depicts an example case with the distributions of the selected parameter before and after the enhancement. More specific, Figure 3a depicts the distributions of the temperature measurements during the production process, while Figure 3b shows the same distribution after the outlier insertion. Figure 4a shows how the temperature of a sealing head variates during production, while Figure 4b depicts the position and the value range of the outliers inserted.

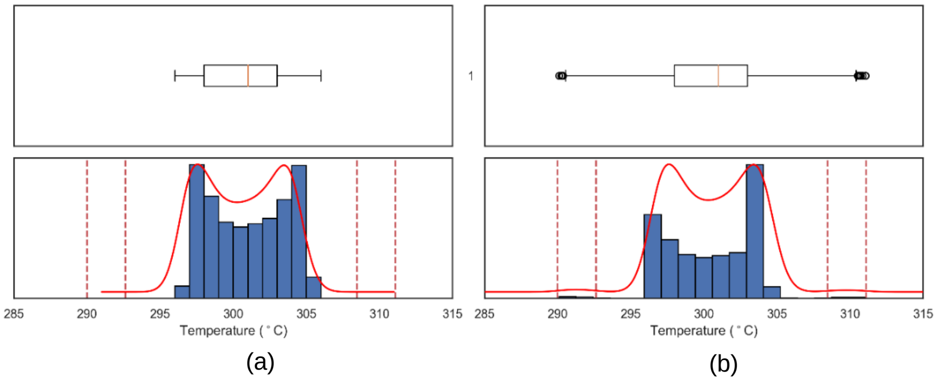


Figure 3: Example case for dataset enhancement with outliers. The vertical lines indicate the area [3σ,4σ] where the outliers will be inserted. (a) shows the original data distribution before. (b) shows the data distribution after inserting the outliers

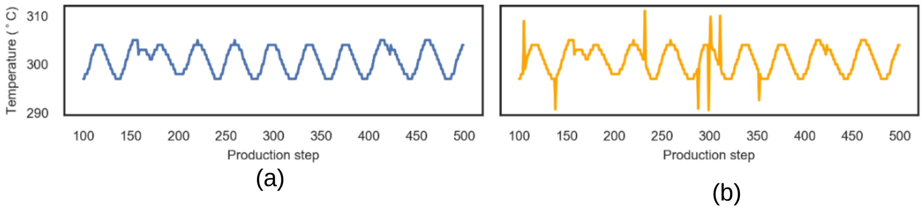


Figure 4: Outlier insertion example in Temperature Head 3. Time series plot before (a) and after (b) outlier insertions

5. Experiments

The proposed framework has been tested with the semi-synthetic dataset created. For each machine parameter the *kNN* algorithm is applied. The k value can be different according to the target parameter to which the algorithm is applied. For this study, $k = 50$ was set for all parameters. Since data are in a time series format, the value of k depicts that the algorithm will take under consideration the previous k neighbours. Next, a statistical analysis on the obtained distances from the *kNN* algorithm is performed, and by applying the statistical processing principle for defining outliers, the values above 3σ are recorded. Finally, the φ_κ correlation between the detected outliers and the rejections is calculated. The higher the correlation the most probable that a hidden systematic malfunction exists.

Figure 5 shows the detected outliers in the Temperature of head 3 by the kNN algorithm, while in Table 2 the φ_k correlation score of the example case is presented. It is clearly depicted that the framework managed to indicate the root cause of the problem, i.e. the Temperature in head 3.

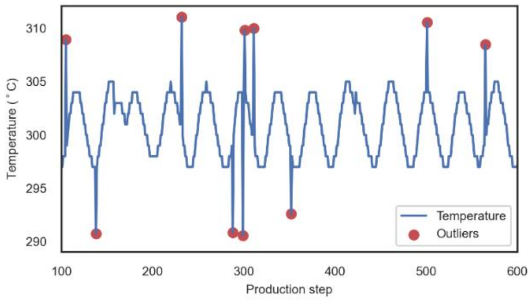


Figure 5: Detected outliers by the kNN algorithm in the example case

Table 2: $\boldsymbol{\varphi_\kappa}$ scores for the example case

Parameter	Outliers detected	φ_k
Temperature 1	20	0.0
Temperature 2	37	0.064
Temperature 3	**108**	**0.988**
Temperature 4	24	0.0
Temperature 5	51	0.0
Temperature 6	5	0.017
Temperature 7	29	0.0
Temperature 8	13	0.0

The proposed framework is iteratively tested with random root cause selection and a random number of outliers inserted. As displayed in Table 3, each run produces different results concerning the φ_k score. However, in all cases, the recognition of the subsystem that causes the rejections is successfully achieved. Finally, while the results is indicated only one as a systematic malfunction, with the proposed approach, it is possible to identify more than one. That is because the correlation score is calculated independently for each subsystem.

Table 3: $\boldsymbol{\varphi_\kappa}$ scores for several random runs

Parameter	Run1	Run2	Run3	Run4	Run5	Run6	Run7	Run8	Run9	Run10
Temperature 1	0.018	0.021	0.000	0.000	0.000	0.000	0.000	0.031	0.000	0.000
Temperature 2	0.000	0.000	**0.978**	0.015	0.024	0.000	0.070	0.000	0.016	0.000
Temperature 3	0.000	0.000	0.000	**0.978**	0.000	0.000	**0.964**	**0.986**	0.000	0.000
Temperature 4	0.000	0.000	0.010	0.000	0.000	**0.994**	0.000	0.000	**0.890**	0.000
Temperature 5	0.000	0.000	0.000	0.000	0.009	0.022	0.052	0.000	0.000	0.000
Temperature 6	0.075	**0.977**	0.006	0.024	0.030	0.016	0.001	0.018	0.000	0.002
Temperature 7	**0.826**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.028	0.000
Temperature 8	0.035	0.038	0.000	0.000	**0.983**	0.000	0.000	0.000	0.000	**0.954**

6. Conclusions

A framework to provide insights and reveal hidden malfunctions that can lead to production losses was presented. The proposed approach utilised the kNN anomaly detection algorithm along with statistical process methods and the φ_k correlation coefficient to produce insights. To validate the efficacy of the method, experiments were performed using a semi-synthetic dataset of a real, fully automatic yoghurt packaging machine that is installed in a large diary. The results indicate that the proposed framework can efficiently provide insights by correlating any anomalies in measured parameters with a key quality test. Since the framework is based on unsupervised methods, this data-driven approach can be easily adapted with minimum effort in similar standalone automated machines that perform a specific task in a production line. To implement the proposed framework, the product quality should be directly monitored. Examples of such machinery can be found in the food industry sector, e.g., bottle labelers, fillers, sealers, in manufacturing e.g., mills, lathes and others. The main limitation of the proposed approach is that the framework cannot produce insights when combined or propagating malfunctions occur. Another challenge, is to apply the proposed framework in systems of higher complexity, where computational bottlenecks could appear. In such cases, decentralised architectures could be utilised to tackle the computational complexity. Those challenges will be further investigated in our future works.

Acknowledgements

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project quote: T2EΔK-01658). Georgios Gravanis and Dimitrios Trigkas have a second affiliation in CERTH. Professor Simira Papadopoulou has a second affiliation in International Hellenic University.

References

Baak M., Koopman R., Snoek H., Klous S., 2020, A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics, Computational Statistics Data Analysis, 152, 107043.

Cover T., Hart P., 1967, Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 13(1), 21-27.

de Paula Monteiro R., Lozada M.C., Mendieta D.R.C., Loja R.V.S., Bastos Filho C.J.A., 2022, A hybrid prototype selection-based deep learning approach for anomaly detection in industrial machines, Expert Systems with Applications, 204, 117528.

Hendrickx K., Meert W., Mollet Y., Gyselinck J., Cornelis B., Gryllias K., Davis J., 2020, A general anomaly detection framework for fleet-based condition monitoring of machines, Mechanical Systems and Signal Processing, 139, 106585.

Kamat P., Sugandhi R., 2020, Anomaly detection for predictive maintenance in industry 4.0-A survey, In E3S web of conferences (Vol. 170, p. 02007). EDP Sciences.

Knox, E.M., Raymond T. Ng., 1998, Algorithms for mining distance-based outliers in large datasets, In Proceedings of 24th international conference on very large databases (VLDB’98), New York, USA, 392–403.

Pittino F., Puggl M., Moldaschl T., Hirschl C., 2020, Automatic anomaly detection on in-production manufacturing machines using statistical learning methods, Sensors, 20(8), 2344.

Ramaswamy S., Rastogi R., Shim K., 2000, Efficient algorithms for mining outliers from large data sets, In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, New York, NY, USA, 427–438.

Zonta T., Da Costa C.A., da Rosa Righi R., de Lima M.J., da Trindade E.S., Li G.P., 2020, Predictive maintenance in the Industry 4.0: A systematic literature review, Computers & Industrial Engineering, 150, 106889.