Università
degli Studi
di Ferrara

DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA

CICLO XXXIII

COORDINATORE Prof. Stefano Trillo

DETECTION AND CLASSIFICATION OF FAULTS AND ANOMALIES IN GAS TURBINE SENSORS BY MEANS OF STATISTICAL FILTERS AND MACHINE LEARNING MODELS

Settore Scientifico Disciplinare ING-IND/09

Dottoranda
Dott.ssa Manservigi Lucrezia

Tutore
Prof. Venturini Mauro

Anni 2017/2020

INDEX

OBIETTIVO

Il monitoraggio e la diagnosi delle turbine a gas sono essenziali e possono essere efficacemente effettuati solo se i sensori installati forniscono una misura attendibile del funzionamento della macchina.

Perciò, l’affidabilità dei sensori è un prerequisito indispensabile ai fini di valutare l’effettivo stato di salute della macchina. Infatti, un sensore guasto potrebbe fornire informazioni inesatte, causando perciò l’interruzione della produzione e un incremento dei costi di manutenzione.

Per questo motivo, questa tesi sviluppa, calibra e valida metodologie finalizzate ad individuare e classificare i guasti e le anomalie dei sensori installati nelle turbine a gas.

La tesi documenta due attività di ricerca con cui è stato raggiunto l’obiettivo prefissato.

In primo luogo, è stato sviluppato lo strumento diagnostico denominato “Improved Detection, Classification and Integrated Diagnostics of Gas Turbine Sensors” (I-DCIDS). Tale strumento è costituito dal Fault Detection Tool e Sensor Overall Health State Analysis (SOHSA). Il Fault Detection Tool individua e classifica le categorie di guasto più frequenti. Invece, SOHSA valuta lo stato di salute complessivo del sensore.

I-DCIDS può essere utilizzato per valutare lo stato di salute sia di sensori singoli sia di sensori ridondanti/correlati, utilizzando equazioni matematiche che richiedono il settaggio di alcuni parametri di configurazione. A tal fine, viene effettuata un’analisi di sensibilità mediante quattro set di dati eterogenei per definire il valore ottimale di tali parametri. Successivamente, I-DCIDS viene validato su un ulteriore set di dati.

Inoltre, I-DCIDS viene anche utilizzato per valutare lo stato di salute di numerosi sensori, analizzando un elevato numero di dati, rappresentativi di sei grandezze fisiche. Queste analisi sono volte ad individuare regole generali con l’obiettivo di determinare la magnitudo del guasto del sensore e l’istante di tempo in cui si verifica.

I risultati ottenuti testimoniano la capacità diagnostica di I-DCIDS sul campo sperimentale. Inoltre, si dimostra che la nuova metodologia può analizzare qualsiasi tipo di dataset e grandezza fisica; infatti, grazie al suo settaggio ottimale, I-DCIDS può anche individuare l’esatto istante di tempo in cui il guasto si è verificato.

Un altro studio condotto in questa tesi riguarda la valutazione dell’affidabilità dei dati acquisiti, che può essere compromessa a causa di anomalie di processo. Questa tipologia di anomalie, raramente investigata in letteratura, può causare errori tali per cui l’unità di misura di un sensore viene erroneamente assegnata. In questa tesi, tale situazione è denominata “Unit Of Measure Inconsistency” (UMI).

Quindi, il secondo obiettivo di questa tesi è quello di individuare lo strumento migliore per diagnosticare con successo l’UMI e per assegnare la corretta unità di misura ai dati privi di tale informazione. A tal fine, vengono esaminati tre classificatori di Machine Learning supervisionato, cioè Support Vector Machine, Naive Bayes e K-Nearest Neighbor. Inoltre, viene proposta ed analizzata una nuova metodologia, chiamata Improved Nearest Neighbor.

Le potenzialità di ogni classificatore sono valutate mediante numerose analisi, per verificare come l’affidabilità dei dati utilizzati in fase di addestramento e il numero di classi influenzino le prestazioni delle varie metodologie.

Si dimostra che il classificatore Naive Bayes e l’Improved Nearest Neighbor sono i più promettenti in termini di efficacia, robustezza e generalità nel maggior numero di casi considerati. In questo modo, si può assegnare la corretta unità di misura e la diagnosi del sensore potrà quindi essere effettuata efficacemente.

Si segnala infine che tutte le analisi riportate in questa tesi utilizzano dati sperimentali acquisiti da sensori installati su turbine a gas di Siemens.

ABSTRACT

Monitoring and diagnostics of gas turbines is a key challenge that can be performed only if the unit is equipped with reliable sensors, thus providing the actual operating condition of the energy system under investigation.

Thus, the evaluation of sensor reliability is fundamental since only a reliable measurement can lead to proper decisions about system operation and health state. In fact, a faulty sensor may provide misleading information for decision making, at the expense of business interruption and maintenance-related costs.

For this reason, this thesis develops, tunes and validates comprehensive methodologies for the detection and classification of both faults and anomalies affecting gas turbine sensors. This purpose is achieved by means of two different analyses and related tools.

First, the Improved Detection, Classification and Integrated Diagnostics of Gas Turbine Sensors (I-DCIDS) tool is developed. The I-DCIDS tool comprises two kernels, namely Fault Detection Tool and Sensor Overall Health State Analysis (SOHSA). The former detects and classifies the most frequent fault classes. The latter evaluates the sensor overall health state. The novel diagnostic tool is suitable for assessing the health state of both single sensors and redundant/correlated sensors. The methodology uses basic mathematical laws that require some user-defined configuration parameters. Thus, a sensitivity analysis is carried out on I-DCIDS parameters to derive their optimal setting. The sensitivity analysis is performed on four heterogeneous and challenging field datasets referring to correlated sensors. Then, the I-DCIDS tool is validated by means of an additional field dataset, by proving its detection capability.

Furthermore, the I-DCIDS tool is also exploited to evaluate the health state of several single sensors, by analyzing a huge amount of field data that refer to six different physical quantities. These analyses provide some rules of thumb for field operation, with the final aim of identifying time occurrence and magnitude of faulty sensors.

The results demonstrate the diagnostic capability of the I-DCIDS approach in a real-world scenario. Moreover, the methodology proves to be suitable for all types of datasets and physical quantities and, thanks to its optimal tuning, also capable of identifying the actual time point of fault onset.

A further challenge addressed in this thesis relies on the evaluation of raw data reliability, which may be compromised because of process anomalies. Such anomalies, which have been rarely

investigated in the literature, may introduce errors whereby the unit of measure of a sensor is wrongly assumed. In this thesis such a situation is named Unit of Measure Inconsistency (UMI). Thus, this thesis is also aimed at identifying the approach that is mostly able to successfully detect UMI occurrence and classify unlabeled data. Among several alternatives, the capability of three supervised Machine Learning classifiers, i.e., Support Vector Machine, Naïve Bayes and K-Nearest Neighbors is investigated. In addition, a novel methodology, namely Improved Nearest Neighbor is proposed and investigated.

The capability of each classifier is assessed by means of several analyses, so that the influence of the reliability of the data used for training the classifier and the number of classes is investigated.

Among all tested approaches, the Naïve Bayes classifier and the novel Improved Nearest Neighbor prove to be the most effective, since they demonstrate their effectiveness, robustness and general validity in the majority of the cases.

Thanks to the selected classifiers, the actual unit of measure of raw data can be provided and further sensor diagnoses can be safely performed.

Finally, it has to be highlighted that all analyses reported in this thesis make use of field data acquired from sensors installed on Siemens gas turbines.

NOMENCLATURE

A	all sensors
c	number of classes
C	criterion
d	distance
F	feature
h	lube oil tank level
I	isolated anomaly
k	acceptability threshold
K	Number of neighbors
l	coefficient of theMinkowski distance
m	minor anomaly
Max	maximum acceptable value
med	median value
Min	minimum acceptable value
M	major anomaly
MLT	multiple
n	rotational speed
N	number of data/sensors
p	pressure
P	posterior probability
S	serial anomaly
SE	sensor
t	time point
T	temperature
Thr	threshold
u	weight
V	vibration
VGV	VGV position
w	number of measurements in the window sample (SOHSA)
W	window of consecutive data (Fault Detection Tool)
x	unlabeled data
Y	measurement in the time series
$\bar{Y}$	mean value of the measurements in the time series
$\hat{Y}$	bi-weight mean
Δ	difference
σ	standard deviation
$\hat{\sigma}$	bi-weight standard deviation

Subscripts and Superscripts

b	backward window
c	number of classes
current	real-time
dith	dithering
f	forward window

GTA	gross threshold analysis
max	maximum
min	minimum
norm	normalized
prev	previous data
rel	relative
std	standard deviation
t	time point
TC	trend coherence
tr	training
vs	virtual sensor

Acronyms

ACA	Anomaly Classification Algorithm
ADA	Anomaly Detection Algorithm
BA	Bias Analysis
BHM	Bayesian Hierarchical Model
CM	Confusion Matrix
DA	Dataset
DCIDS	Detection Classification and Integrated Diagnostics of gas turbine Sensors
DTA	Dithering Analysis
FNR	False Negative Ratio
FPR	False Positive Ratio
GT	Gas Turbine
GTA	Gross Threshold Analysis
INN	Improved Nearest Neighbor
I-DCIDS	Improved DCIDS
K-NN	K-Nearest Neighbor
LA	Linear function with OVA decomposition
LO	Linear function with OVO decomposition
ML	Machine Learning
NB	Naïve Bayes
OVA	One-Vs-All
OVO	One-Vs-One
PCA	Principal Component Analysis
PP	Posterior Probability
PQ	Physical Quantity
RBF	Radial Basis Function with OVA decomposition
RO	Radial Basis Function with OVO decomposition
SA	Spike Analysis
SDA	Standard Deviation Analysis
SOHSA	Sensor Overall Health State Analysis
SSA	Stuck Signal Analysis
SVM	Support Vector Machine
TCA	Trend Coherence Analysis
TPR	True Positive Ratio
UMI	Unit of Measure Inconsistency
UOM	Unit Of Measure

1. INTRODUCTION

Energy market demand sets high requirements to the productivity of gas turbine (GT) units, by imposing high availability and efficiency levels to achieve cost effectiveness. Furthermore, the complexity of the units implies a high level of insight on the health state of the turbines, so that potential problems can be identified long before they impact operation. In such a manner, downtime and maintenance costs can be reduced.

In the last years, an increasing number of sensors has been installed on GT units in order to provide a reliable piece of information regarding GT health state. As a result, a large amount of data, e.g., in the order of tens of GBs, can be acquired on a daily basis and the trend is to increase even further [1]. Thus, challenges related to GT monitoring and diagnostics interact with the requirement to the big data processing. In this context, analyses that involve frequent human decisions are both error-prone and highly impractical. Therefore, it is of great importance to provide automated analyses that are both effective and efficient with regard to the detection of problems.

Effectiveness refers to the ability to find all the real problems while reducing the number of false alarms, especially filtrating those generated by sensor malfunctioning. Despite their different purpose, in fact, state-of-the-art analysis tools rely on direct thermodynamic (e.g., temperatures, pressures) and mechanical (e.g., vibrations) measurements to monitor [2, 3] or forecast [4, 5, 6] the health state of GT units. Data processing effectiveness can generally be achieved through careful tuning of parameters. However, even if the detection capabilities are somewhat related to the number of model parameters, in most cases this is seen as a source of issues rather than a benefit. This is the main reason why complex methodologies based on heuristic rules tend to fail in field applications. In this regard, a desired feature is high usability of methodologies, meaning that their tuning process should be user-friendly without losing their detection capability.

Instead, efficiency means that analysis should be concluded in a reasonable amount of time, leaving enough room to react proactively to an identified problem. This stresses the suitability of computationally undemanding methodologies rather than complex model based approaches for the development of efficient methodologies for remote diagnostics.

It has to be highlighted that the extreme conditions in which sensors operate may cause hardware degradation and failure of measurement devices [7], thus generally lowering data quality over time. This fact, coupled to the high number of sensors, may make sensor faults more frequent compared to GT lifetime.

As monitoring and control applications employee sensor data for decision support, it is fundamental that acquired and stored data are accurate and reliable. In fact, a faulty sensor cannot properly perform its function, but instead may provide misleading information for decision making. The problem of measurement reliability is much more critical in case of physical quantities that are not directly measured, e.g., the engine thrust [8] or sensors where the measuring principle relies on estimated parameters, as e.g., heat transfer coefficient [9]. A further issue is given by systematic errors, of which the influence is analysed in detail in [10]. In particular, a critical source of GT performance loss related to sensor reliability assessment can be attributed to GT blades subject to dynamic loads caused by rotor imbalances and distortions in the intake flows [11]. These loads cause low or high cycle fatigue and the blades can fail over time [12, 13].

For these reasons, sensor fault detection constitutes a fundamental and challenging task in several scientific activities, which also include GT research field [14 –17]. According to [18], a sensor fault occurs when its measurement displays an unacceptable deviation from the true value of PQ.

Prasad *et al.* [19, 20] recently reviewed several diagnostic methodologies, by describing both traditional and modern approaches. In [19], the authors classified the diagnostic approaches that use prominent techniques (e.g., wavelet, artificial neural network and fuzzy logic approach) and hybrid techniques (e.g., neuro-fuzzy techniques [19, 21, 22]), which combine at least two techniques to overcome the respective drawbacks. The same authors analysed in [20] the most modern diagnostic methodologies, such as support vector machine (SVM), genetic algorithm and Principal Component Analysis (PCA). In the literature, artificial neural networks [23], the fuzzy logic approach and PCA [24, 25] are the most common data-driven diagnostic approaches. In particular, PCA is the most exploited data-driven technique since it allows to squeeze numerous multivariate data to obtain general information about the available measurements, as demonstrated by Bakdi *et al*. in [26]. Nevertheless, PCA usually generates several false alarms once slow time variations occur (Portnoy *et al*. [27]).

Sensor diagnostics can be also performed by exploiting repetition or duplication of elements within a system [28]. According to Verma and Singh [29], redundancy reproduces data resources, e.g., sensors, by generating similar data acquired from different sources. In general, three types of redundancy exist, i.e., temporal, information and spatial [29].

Temporal redundancy, or time redundancy, is achieved when sensor measurements are performed during consecutive time steps [30]. This redundancy definition is generally used for detecting intermittent faults, since they do not occur consistently [31]. Instead, information redundancy is achieved when complementary information is added to the original one. Finally, thanks to spatial redundancy, insights about a specific environment are provided by means of different sources.

Spatial redundancy can be achieved by means of both physical and analytical redundancy. Physical redundancy, also known as direct redundancy, arises when multiple independent sensors cover a specific area. Instead, analytical redundancy is a generalization form of physical redundancy, in which virtual sensors (also named virtual sensing or software sensors) are also accounted for. In the last years, novel virtual sensor approaches have been developed and successfully exploited in several fields to evaluate unmeasured physical quantities and validating the system status, by also simulating the healthy behaviour of the physical sensor [32].

Some of the most recent methodologies developed for the sensor diagnostics are documented below.

Zhao et al. [30] have recently developed a novel diagnostic approach that exploits both spatial and time redundancy models for sensor fault diagnostics on aero-engines. Instead, Stetter at al. [33] designed virtual sensors for both fault detection and prevention. Capriglione and Pietrosanto [34] verified the diagnostic capability of a nonlinear autoregressive with exogenous input neural networks that generated a virtual sensor for detecting stroke sensors. Mattera et al. [32] created virtual sensors by means of linear regression models, SVM regression and artificial neural networks. Losi et al. [35] exploited the Bayesian Hierarchical Model for generating a virtual sensor for GT purposes.

Stoumpos and Theotokatos [36] proposed a novel methodology for sensor diagnostics and health management of marine dual fuel engines, by employing a combination of thermodynamic, functional control and data-driven models. The paper was aimed at developing an intelligent engine monitoring, advanced sensor fault detection as well as prompt corrective actions. The methodology was based on a data-driven approach (i.e., neural network model) and proved to be sufficiently capable of capturing sensor faults.

Darvishi et al. [37] tuned a novel methodology for detecting sensor faults and replacing faulty observations with appropriate estimated data. The authors proposed a general machine-learning architecture based on a series of neural-network estimators and a classifier. The novel methodology successfully detected both hard and soft implanted faults.

Papers [24, 38 - 41] presented some approaches towards sensor fault detection and isolation in GTs. In [24], an adaptive kernel PCA solution was tuned to overcome the PCA drawback in time-varying processes. The methodology was then tested on an aeroderivative GT, thus detecting and isolating sensor degradation. Instead, the study [38] proposed a new sensor fault methodology based on robust Kalman filters that proved to be effective even in the case of noisy data and time-varying parameters. The proposed approach was tested on several sensor faults occurring in GT engines, proving its capability despite the presence of measurement noise and uncertainties. Rahme and Meskin [39] reviewed several sensor fault approaches, e.g., neural network, genetic algorithm and Kalman filter, summarizing their advantages and drawbacks. A robust nonlinear model was then developed in the same paper by means of an adaptive sliding mode observer. Finally, paper [40] proposed a data-driven methodology aimed at fault detection and isolation for both aircraft GT engines and sensors. The methodology requires evaluation of the frequency response of the GT engine by applying the Fast Fourier Transform to the input and output signal. This piece of information is employed to directly design the fault detection filters; the methodology therefore does not require any a priori knowledge of the system.

Another example is provided by paper [41], which presented a linear regression model for fault detection and signal reconstruction in industrial GTs. The detection of a faulty sensor installed in GTs was also performed by Jombo *et al*. in [42] by using the discrete wavelet transform, suitable to identify several fault classes, e.g. spike, bias and erratic faults.

Paper [43] reports the tuning of an Adaptive Gas Path Analysis to evaluate GT performance and assess component health state, correlating different physical quantities (PQs) such as flow rate, pressure, rotational speed and temperature. Nonlinear Gas Path Analysis is also used in [44] to evaluate unit degradation, in combination with a gray algorithm to enhance the robustness of the tool.

In addition to the sensor fault detection, it is also important to identify the fault type and magnitude, in order to find the origin of the malfunctioning, make a decision upon sensor replacement or design more robust sensors or data acquisition systems.

The task to identify the fault type can be labelled as a classification issue. As stated by Da Silva *et al*. [45], classification is a crucial step in several research fields. The main aim of any classification methodology consists in providing a class to each unlabelled data, characterized by attributes and features [45]. Similarly, Sanchez *et al*. [46] stated that the main goal of the classification task consists in implicitly or explicitly generating decision boundaries that distinguish the available classes within a *c*-dimensional feature space.

Thus, sensor fault classification represents the added value of every diagnostic methodology, since it allows to label occurring faults.

Some of the most frequent sensor fault types are reported in [47], which investigated seven fault classes, i.e., bias, drift, gain, precision degradation, constant, constant with noise and noise.

In order to both detect and classify sensor faults, Ceschini *et al*. have recently described the development and tuning of a novel statistics-based tool for anomaly and noise detection [1, 48, 49]. The proposed tool is a comprehensive approach for the Detection, Classification and Integrated Diagnostics of Gas Turbine Sensors (DCIDS). The DCIDS tool includes two kernels: an Anomaly Detection Algorithm (ADA) and an Anomaly Classification Algorithm (ACA). The former identifies the anomaly, noisy and reliable data by means of three levels of judgement, which include a statistical filter that can be selected by the user from four methodologies (i.e., k-σ, k-MAD, hybrid σ-MAD and bi-weight methodologies). The detected anomalies are then classified according to their time correlation, magnitude and the number of sensors in which the anomaly is detected at the same time point.

It has to be highlighted that both GT and sensor diagnostics can be efficiently performed only if the acquired data, which are categorized and usually normalized [50], correspond to the actual operating conditions in which data were collected [51]. In fact, the reliability of model–based, data-driven and hybrid methodologies rely on data quality, either directly or indirectly.

However, data quality may be compromised because of process anomalies, which make inconsistent the labelled data with the measured physical quantity.

For example, because of the availability of several systems of measurement, e.g., the SI and the Imperial system, Unit Of Measure (UOM) may be wrongly assumed, by causing misleading or inconsistencies. In this thesis, such an anomaly is named UOM Inconsistency (UMI).

Regardless of the research field, preliminary analyses dealing with the UOM consistency are crucial to avoid undesired consequences, which may be also catastrophic. In fact, several real-life events caused by such an inconsistency are documented. The most renowned example is represented by the failure of the NASA’s Mars Climate Orbiter spacecraft, which was designed and built by accounting for the Imperial system, whereas all simulations were carried out by means of the SI [52]. As a result, NASA lost $125 million because of the UOM misunderstanding. The “Gimli Glider” incident is a further historical example that highlights this problem, since the Air Canada’s Boeing 767 jet ran out of fuel because it was scaled in the incorrect UOM [53]. Finally, an American International Airways flight landed 15 tons heavier because of an incorrect kilogram-to-pound conversion. These examples show that a reliable and general approach for the UOM check is necessary.

In spite of its relevance, only a few studies deal with UOM consistency, which is usually investigated under an information technology point of view. For example, the studies [52, 54, 55] propose new routines aimed at the dimensional analysis, i.e., the procedure that leads to the evaluation of UOM consistency [52, 54], for programs written in the most popular programming languages. The tuned routines assess UOM consistency, by also providing the list of equations that violate the dimensional analysis. However, as stated in [55], computational packages generally lack UOM check, even for space missions.

Since UMIs may compromise GT diagnosis and the response of diagnostics approaches, their detection may be a preliminary step that precedes further diagnostics analyses.

Nowadays, UMIs are usually detected by comparing a given signal with respect to minimum and maximum acceptability thresholds in which the signal is expected to lie. In such a manner, if the measured values exceed the acceptability thresholds, a UMI is detected. However, this approach may not be general and efficient, especially if the range of thresholds is incorrectly set. As an alternative, UMIs can be detected by means of visual inspection. However, for industrial feasibility, an automated procedure for UMI check is required.

To this purpose, Machine Learning (ML) techniques may be employed to (i) detect UMIs and (ii) associate the correct UOM to a given measured physical quantity. In fact, ML is a data-driven approach that handles data in order to generate a model, which is subsequently employed to analyse and classify further measurements. ML approaches can be divided into two categories, i.e., supervised and unsupervised learning, based on information provided for data classification. Supervised ML requires labelled training datasets for model generation; conversely, unsupervised ML handles data whose label is not previously known.

In the literature, the effectiveness of both unsupervised and supervised ML approaches is documented in different research fields, also for GT applications. However, this thesis is only focused on supervised techniques, so that the UOM label can be provided.

Some of the most recent studies dealing with ML classifiers and their classification capability are briefly discussed.

Wu *et al*. [56] compared three ML approaches for GT fault diagnosis, by achieving promising results in terms of accuracy and computational time. Similarly, Jaw and Lee [57] examined some of the most popular ML techniques with the aim of selecting the most reliable technique for GT diagnosis. The author investigated the Naïve Bayes, Decision Tree, K-Nearest Neighbours (K-NN), linear and non-linear Support Vector Machine (SVM). Instead, Luo and Zeng [58] compared SVM and neural network, by stating that SVM is preferable if small size samples are available. Elangovan *et al*. [59] tested some classifiers, e.g., SVM, Naïve Bayes and Decision Tree for vibrational signal classification. Shi *et al*. [60] has recently exploited the K-NN method to classify faults affecting rotors, while the Naïve Bayes classifier is challenged by Niazi *et al*. [61] at detecting hotspots occurring in solar photovoltaic modules. Finally, Zhang *et al*. [62] tuned a multi-model ensemble, which comprises K-NN, SVM, Decision Tree, Random Forest and logistic regression, to the fault diagnosis in refrigeration systems.

1.1 NOVELTY

Based on the requirements highlighted by means of the literature survey, this thesis is aimed at developing and testing novel methodologies that evaluate the overall health state of GT sensors. First, this thesis deals with the improvement of the DCIDS approach previously developed in [1]. Unlike the previous approach, the improved methodology, called I-DCIDS, allows the detection and classification of seven classes of fault (i.e., out of range, stuck signal, dithering, standard deviation fault, trend coherence fault, spikes and bias). In fact, as discussed, these faults are recognized in the literature as the most frequent sensor faults. Moreover, thanks to the general approach, other types of fault may also be identified by means of I-DCIDS methodology.

In the DCIDS tool [1], anomalies were categorized according to three criteria: i) isolated vs. serial anomalies; ii) minor vs. major anomalies and iii) single vs. multiple data anomalies. Instead, the I-DCIDS tool allows each fault to be labelled with a specific recognizable pattern.

The procedure adopted in the DCIDS tool [1] for anomalies classification was based on statistical filters. Instead, the I-DCIDS tool performs a specific analysis for each fault type.

Consequently, the novelty of the methodology developed for the identification of sensor faults and anomalies lies in the development of a comprehensive, flexible and robust approach for fault detection and the classification and evaluation of sensor health state. In fact, the I-DCIDS methodology allows both detailed and general information about sensor reliability to be obtained. Moreover, it performs a systematic and general analysis by means of a “physical” approach, applicable to all the physical quantities (PQs). In addition, it allows classification of the faults, even if they occur simultaneously, while only a few approaches available in the literature allow the labelling of multiple faults.

One of the strengths of the I-DCIDS methodology is that it is straightforward. Thus, it is potentially suitable for different industrial applications and also non-expert users. Moreover, it is also suitable for real-time applications, since the computational time is very low (on the order of a few seconds to process one data point).

The I-DCIDS tool clearly requires some configuration parameters for fault detection and classification. This thesis therefore also addresses the optimal tuning of the fault detection parameters, regardless of the PQ under consideration and the operating conditions, in order to draw general guidelines for effective and reliable sensor fault detection and classification. This aim is achieved by means of a sensitivity analysis, which is performed on four heterogeneous and challenging field datasets acquired from Siemens GTs running in different operating conditions, both in steady-state and transient conditions.

Similar to the approach taken in Refs. [23, 63], it is assumed that the I-DCIDS methodology is only applied during regions of quasi steady-state GT operation. Applying this or other sensor diagnostic methodologies during rapid system transients (i.e., start-up and shutdown) may increase the risk of false alarms, as stated in [25] and [26].

The parameter configuration, identified by means of the sensitivity analysis, is tested in one further experimental dataset including only nominal, fault-free sensor measurements to highlight that no false positives are detected.

In addition, the thesis addresses an innovative sensor diagnostic approach developed by Losi *et al*. [64], based on a prognostic methodology, namely Bayesian Hierarchical Model, which is exploited to simulate a virtual sensor representative of reliable sensor behaviour. Such a sensor fault diagnostic approach has never been investigated in the literature.

A further main novelty of this thesis is represented by UMI detection, since, in spite of its relevance, it has not been adequately addressed in the literature. To this aim, several ML classifiers are reviewed and tested. Among several alternatives, this study evaluates the exploitation of three supervised classifiers, i.e., SVM, Naïve Bayes and K-NN; in fact, according to the literature survey, they proved to be the most popular and effective supervised ML classifiers, whose flexibility and versatility was assessed in several research fields. In addition, a novel methodology, namely Improved Nearest Neighbor is tuned with the aim to overcome K-NN drawbacks.

It has to be mentioned that it is the first time in the literature that the ML approaches are employed to detect UMIs occurring in GT field datasets.

Finally, an additional strength of this thesis is the fact that the methodologies and tools developed in this study extensively and exclusively employ field data, by illustrating the challenges associated with the application of sensor fault and anomaly diagnostics to actual system measurements.

This thesis is structured as follows.

The first part is focused on the detection and classification of sensor faults and data anomalies. After that the definition of “fault” and “anomaly” is defined and the classification of the most frequent fault and anomaly classes is reported, the development of the I-DCIDS methodology is documented. Then, all parameters required by the I-DCIDS methodology are tuned by means of field datasets acquired from Siemens GT in operation. Furthermore, the I-DCIDS methodology is validated by means of an additional dataset to prove its capability. These analyses, carried out on correlated sensors, allow the identification of sensor overall health state. In addition, the I-DCIDS methodology is also exploited to evaluate the health state of a large number of single sensors, by analysing approximately 70 million data. Thanks to the analysis of these data, this thesis also provides some rules of thumb about the occurrence of sensor faults and anomalies in GT field operation.

Then, the exploitation of the innovative BHM for diagnostic purposes is briefly documented.

The second part of the thesis id dedicated to the UMI detection. As previously mentioned, three supervised ML classifiers, i.e., SVM, Naïve Bayes and K-NN are accounted for. In addition, the Improved Nearest Neighbour classifier is developed. The characteristics of each classifier, as well as its benefits and drawbacks, are discussed. Then, the training and testing pipeline, field data, tests and metrics for evaluating the classification capability are shown. Finally, a direct comparison of the most promising classifiers is provided with the final aim to identify the optimal supervised ML classifier for the UMI detection.

2. FAULT AND ANOMALY

This thesis is aimed at detecting and classifying faults, anomalies and noise affecting sensors installed in gas turbines.

As discussed in the following, “fault” and “anomaly” have a different meaning.

2.1 SENSOR FAULT

Sensor faults are patterns in data that do not conform to a well-defined notion of normal/expected behavior [64]. The normal/expected behavior is provided by the vast majority of sensors, according to the principle that the decision regarding which system to trust is made democratically and that the majority rules [65].

The most frequent faults affecting sensors are documented in the following.

2.1.1 Out of range fault

Out of range faults, similarly to the “out of bounds” addressed in [66], are faulty data that exceed physical thresholds.

Figure 1 shows a sample nondimensional physical quantity affected by out of range faults. In fact, the simulated data number 4 and 16 are higher and lower, respectively, than the gross thresholds (assumed to be 0 and 1).

The gross thresholds usually depend on (i) the physical quantity (e.g., a temperature cannot be negative) and (ii) the minimum and maximum value that sensors can measure.

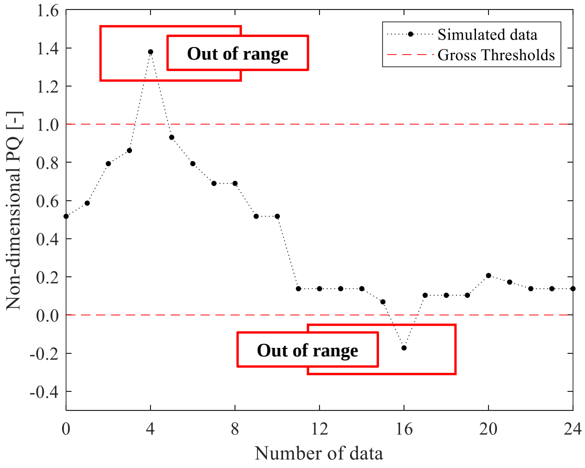


Figure 1 – Example of out of range fault

2.1.2 Stuck signal fault

In the literature, papers [66] and [67] deal with stuck signal faults by means of different approaches. In fact, according to [66], a “stuck-at” fault is detected if the signal is constant over time. A “stuck-at-one” fault is detected if the faulty signal is equal to the maximum value that the sensor can measure. Instead, if the faulty signal is equal to the minimum measurable value, a “stuck-at-zero” fault is detected. Conversely, paper [67] designs a switching law and filter aimed at detecting stuck signal faults.

Methodology reliability is tested in [67] on two case studies. The first case study deals with a sensor outage case, in which stuck values are equal to zero. The second case study is a sensor stuck fault affected by a small amplitude.

An example of stuck signal fault is shown in Fig. 2, where the signal is constant from data number 8 to 20.

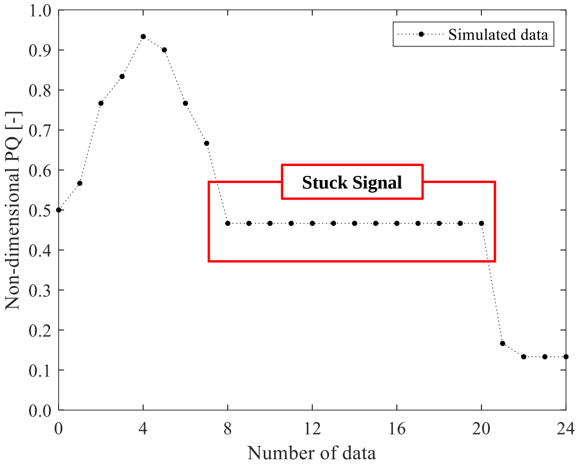


Figure 2 – Example of stuck signal fault

2.1.3. Dithering fault

As reported in [68], several definitions of dithering faults are documented. In fact, as stated by Iannelli *et al.*, dither signals must stand on two different values, which can exhibit a saw tooth, sinusoidal, square wave, triangular or trapezoidal profile.

In this thesis, only saw tooth profiles are investigated, as depicted by means of simulated data shown in Fig. 3.

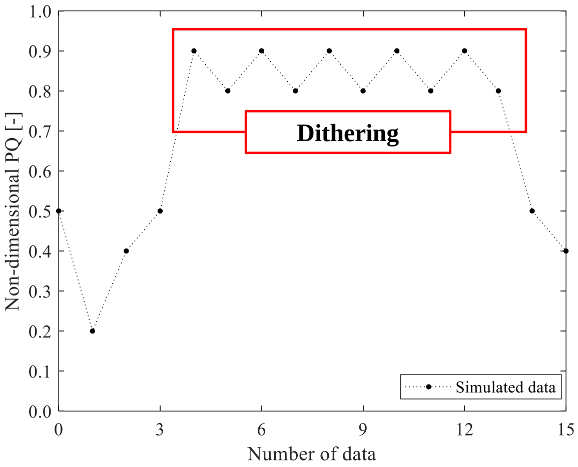


Figure 3 – Example of dithering fault

2.1.4. Standard deviation fault

Figure 4 shows an example of a standard deviation fault. According to the simulated data shown in Fig. 4, a signal, or part of it, is characterized by a standard deviation fault if its variation is greater or lower than a reference signal. For the sake of this general definition, erratic signal faults, one of the most common fault type [42], may also be detected. In fact, an erratic signal fault is characterized by high excursions of sensor readings, which can be viewed as a change in the signal variance [42].

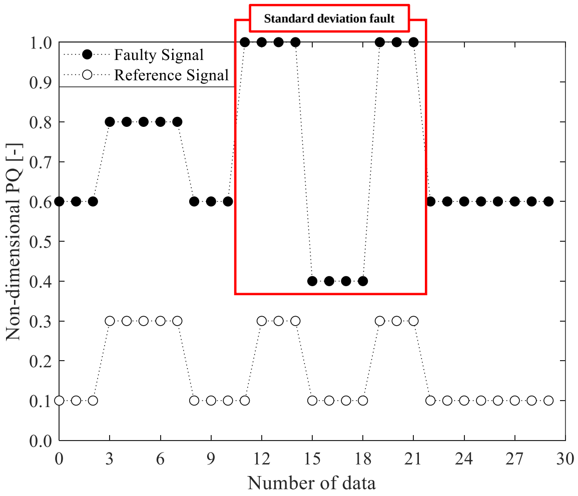


Figure 4 – Example of standard deviation fault

2.1.5. Trend coherence fault

Trend coherence faults are faulty data that exhibit an erroneous trend over time compared to a reference signal. An example of Trend Coherence Fault is depicted in Fig. 5.

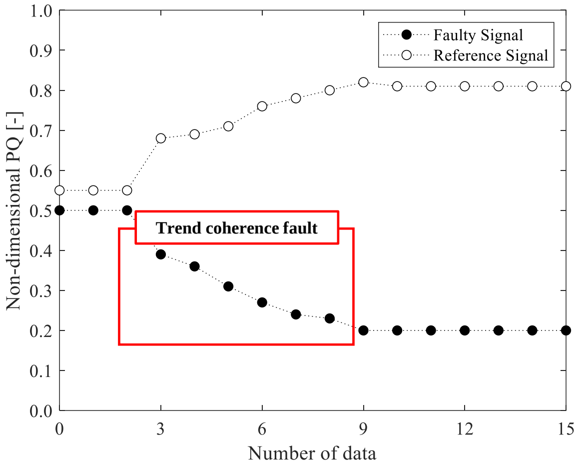


Figure 5 – Example of trend coherence fault

2.1.6. Spike fault

In the literature, studies [42], [66] and [69] define spike faults by means of similar definitions, i.e., a spike represents an abrupt increase or decrease in the PQ. Thus, a spike fault can be detected by analyzing just one sensor trend.

In the current thesis, a spike is defined as an isolated data point that does not correctly reproduce the behavior of a reference signal.

The simulated data in Fig. 6 include two examples of a Spike Fault.

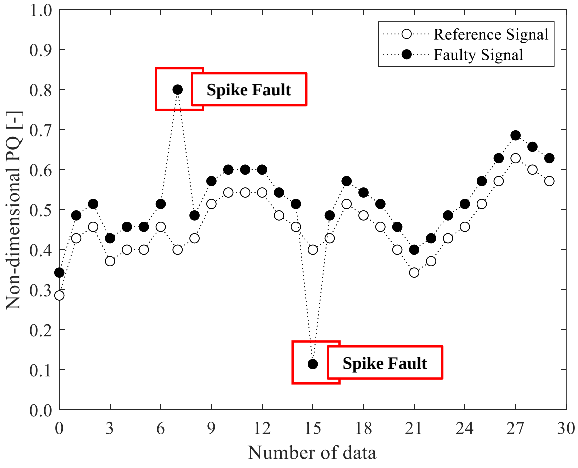


Figure 6 – Example of Spike Fault

2.1.7. Bias fault

Paper [70] specifically investigates bias faults, stating that their classification depends on the evaluation of the difference between pairs of redundant sensors. This paper classifies both a drift and a step change as representative of bias faults. On the other hand, study [43] only labels a step change as a bias fault.

This thesis includes the approaches documented in both [43] and [70], by assuming that a bias fault occurs if part of the signal differs from the reference value by a constant, showing a step change.

Figure 7 depicts an example of bias fault. According to the simulated data, a bias fault occurs from data 7 to 15. In fact, the signal reproduces the correct trend but, at data 7, it deviates from the reference sensor by means of a step change.

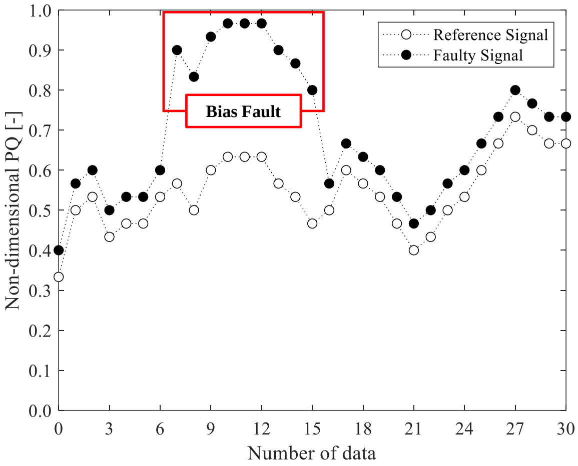


Figure 7 – Example of bias fault

2.2 ANOMALY

In this thesis, two different types of anomalies are investigated, i.e., (i) data anomalies and (ii) process anomalies.

It should be noted that both anomaly types are significantly correlated to raw data reliability. In this thesis, two anomaly types are accounted for, i.e., data anomalies and process anomalies.

2.2.1. Data anomaly

Data anomalies represent an anomalous behavior of data, which can be detected by means of the exploitation of statistical filters, as documented in Ceschini *et al*. [1]. As a result, an unreliable measure of the physical quantity is provided.

Monitoring of data anomalies is crucial, as they may be a warning about the incipient sensor failure.

Ceschini *et al.* [1] provided some rules of thumb for anomaly classification. Key information about anomalies classification concerns:

a. mutual time correlation;

b. magnitude, i.e. distance from acceptability boundaries;

c. detection score.

The analysis of these key features allows the discrimination between the following anomaly classes:

a. isolated anomalies vs. serial anomalies;

b. minor anomalies vs. major anomalies;

c. specific sensor(s) anomalies vs. multiple data anomalies.

Anomalous observations are related to the proposed anomaly classes according to three different criteria, which reflect the previously identified key information:

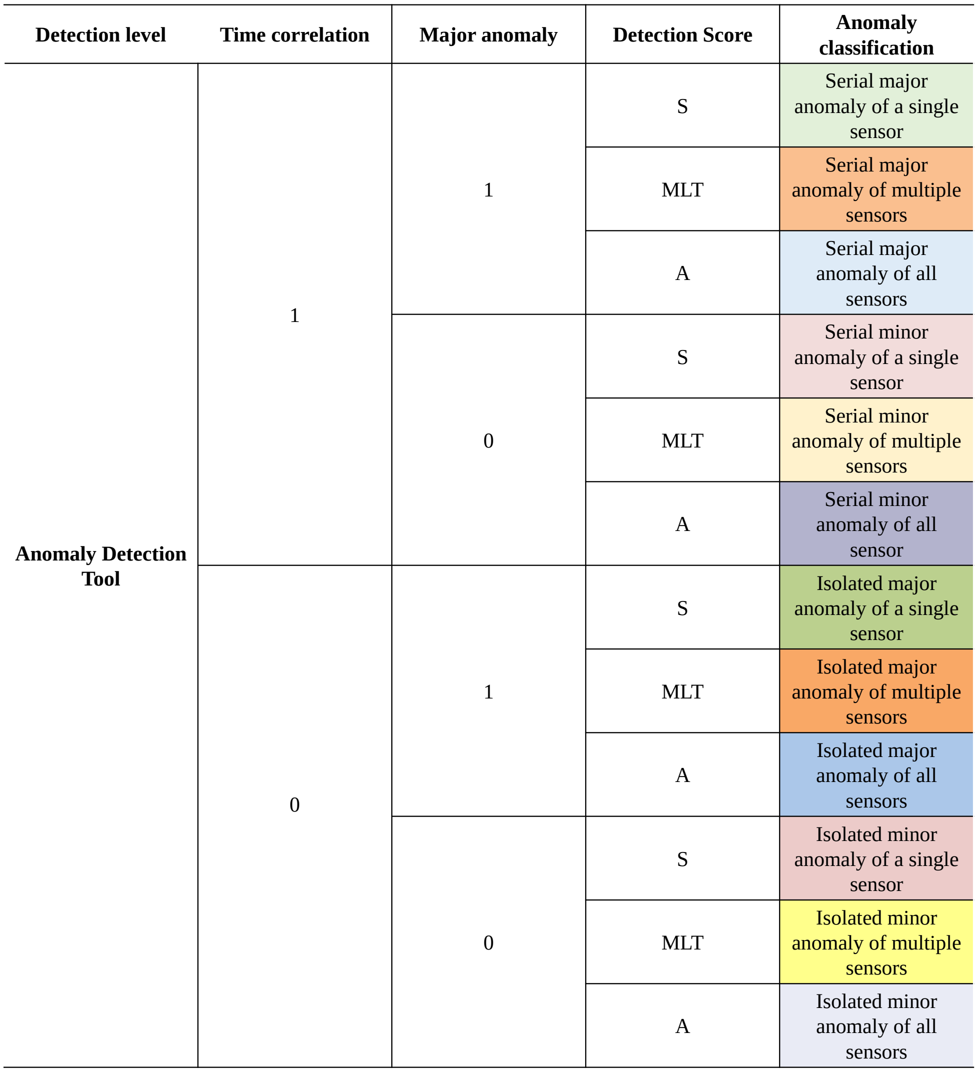
a. time correlation: two anomalies are time correlated if they occupy subsequent slots in the time series (e.g. at time point t and t+1);

b. major anomaly: anomalies whose magnitude is outside an interval, whose width is a function of measurement uncertainty. An anomaly that lies inside the same interval is classified as minor anomaly;

c. the value of the detection score allows the discrimination among single data anomaly, multiple data anomaly (i.e., at least two sensors are anomalous).

The classification pattern is summarized in Fig. 8, in which 1 indicates that a particular feature is verified, while 0 means the opposite. Symbols S, MLT and A for detection score mean that a certain anomaly is detected by single (S), multiple (MLT) or all but two sensors (A), respectively.

The proposed anomaly classes are designed to reflect typical field operation sensor faults, so that they can be predicted long before their occurrence. For example, the situation in which the classification changes from serial minor anomaly to serial major anomaly of a single sensor (identified in Fig. 8 as 1.0.S and 1.1.S, respectively) is representative of a series of time correlated anomalies with an increasing magnitude. This is the typical pattern of a sensor drift fault, in which sensor measurements gradually deteriorate over time. On the contrary, if no evolution of anomaly magnitude is detected, the occurrence of a bias fault is more likely. In this scenario, in fact, sensors produce faulty observations, of which the magnitude is constant over time. Another example is the case of isolated major anomaly(s) of single or multiple sensors, identified in Fig. 8 as 0.1.S and 0.1.MLT respectively, which corresponds to the occurrence of a spike.



Detection level	Time correlation	Major anomaly	Detection Score	Anomaly classification
Anomaly Detection Tool	1	1	S	Serial major anomaly of a single sensor
			MLT	Serial major anomaly of multiple sensors
			A	Serial major anomaly of all sensors
		0	S	Serial minor anomaly of a single sensor
			MLT	Serial minor anomaly of multiple sensors
			A	Serial minor anomaly of all sensor
	0	1	S	Isolated major anomaly of a single sensor
			MLT	Isolated major anomaly of multiple sensors
			A	Isolated major anomaly of all sensors
		0	S	Isolated minor anomaly of a single sensor
			MLT	Isolated minor anomaly of multiple sensors
			A	Isolated minor anomaly of all sensors

Figure 8 – Classification of data anomalies

For the sake of clarity, anomaly classes are hereafter gathered in four main classes, namely Isolated-Major (I-M), Isolated-minor (I-m), Serial-Major (S-M) ad Serial minor (S-m).

2.2.2. Process anomaly

Raw data reliability may be also compromised by process anomalies, as they may cause inconsistencies between field data and their labels. For example, UOM labels may be wrongly assumed. In this thesis such a situation is named Unit of Measure Inconsistency (UMI).

This thesis is addressed to develop a general methodology for detecting UMI occurrence and subsequently assigning the actual label to raw data affected by UMI.

2.3 SENSOR FAULT *VS.* DATA ANOMALY

In this thesis, “sensor fault” and “data anomaly” have a different meaning. Both anomalies and faults are patterns in data that do not conform to a well-defined notion of normal/expected behavior [64]. The normal/expected behavior is provided by the vast majority of sensors, according to the principle that the decision regarding which system to trust is made democratically and that the majority rules [65]. The difference between “faults” and “anomalies” is accounted for by the tool used for their detection. In fact, “faults” (as classified in [47]) are detected by the “Fault Detection Tool”. Instead, an “anomaly” is detected by means of the SOHSA (Sensor Overall Health State Analysis) tool, which, for example, allows noise detection. Moreover, a fault is detected by means of basic mathematical laws, while anomaly and noise detection is performed by means of a statistical approach.

2.4. SENSOR FAULT *VS.* UNIT OF MEASURE INCONSISTENCY

The discrimination between sensor faults and UMIs is a challenging task that can be performed only by means of a in deep insight of sensor behavior. In fact, at first sight, UMIs may be wrongly confused with one sensor fault, e.g., out of range fault, at the expense of the sensor, which is labelled as faulty.

As shown in Table 1, sensor faults (e.g., bias, spike and erratic signal) and UMIs can be both classified by means of the standard deviation and the mean value of the signal. For example, if a bias anomaly occurs, the signal standard deviation is comparable to the standard deviation of the healthy signal, whereas the mean value can be higher or lower than that of the healthy signal. Two different UMI anomalies can occur:

- if both standard deviation and mean value are lower than that of the reference signal, i.e., healthy sensor with consistent UOM, the actual UOM is a multiple of the supposed UOM;
- if both standard deviation and mean value are higher than that of the consistent UOM, the actual UOM is a sub-multiple of that UOM.

As sketched in Table 1, UMI characteristics could be similar to “constant with noise” signal in case of low standard deviation and mean, whereas they are spike or the combination of bias and erratic signal in case of high standard deviation and mean. However, for a given GT, UMIs could be detected by checking the amount of data collected by the same sensor that exhibits an anomalous behavior. In fact, UMIs affect the entire timeframe, whereas sensor faults are generally limited in time.

Table 1 - Sensor faults and UMI classification with respect to signal standard deviation and mean value

Mean value	**Standard deviation**		
	Low	*Average*	*High*
Low	Constant with noise or **UMI**	Bias	Bias + Erratic
Average	Constant with noise	**Healthy sensor with consistent UOM**	Erratic or Noise
High	Constant with noise	Bias	Spike or Bias+Erratic or **UMI**

2.5. NOISE

The diagnostic tool developed in Ceschini *et al*. [1] allowed to classify raw data in three different classes, i.e., anomaly, noise and reliable.

Thus, this thesis also discusses the detection of noisy observations, which are placed within the intermediate region between the sensor reliability and anomaly.

3. DETECTION AND CLASSIFICATION OF SENSOR FAULTS AND ANOMALIES

This Section of the thesis documents two different methodologies aimed at sensor fault and anomaly detection, namely Improved DCIDS tool (I-DCIDS) and Bayesian Hierarchical Models.

3.1. I-DCIDS TOOL

This thesis develops a novel diagnostic tool, i.e., I-DCIDS, which is the improvement of the DCIDS tool, discussed in Ceschini *et al*. [1]. As discussed in Section 1, the novelty of the I-DCIDS tool with respect to the DCIDS tool consists in the detection and classification of seven fault classes, i.e., out of range, stuck signal fault, dithering fault, standard deviation fault, trend coherence fault, spike fault and bias fault.

Thus, the I-DCIDS methodology is a comprehensive tool able to (i) efficiently detect and classify specific sensor faults and (ii) assess the overall health state of sensors, as described below.

Figure 9 depicts a diagram of the I-DCIDS methodology, which comprises two kernels, namely Fault Detection Tool and Sensor Overall Health State Analysis (SOHSA).

According to the methodology, all the raw data are analyzed by means of both the Fault Detection Tool and the SOHSA. The Fault Detection Tool allows the detection and classification of faults occurring in the time series.

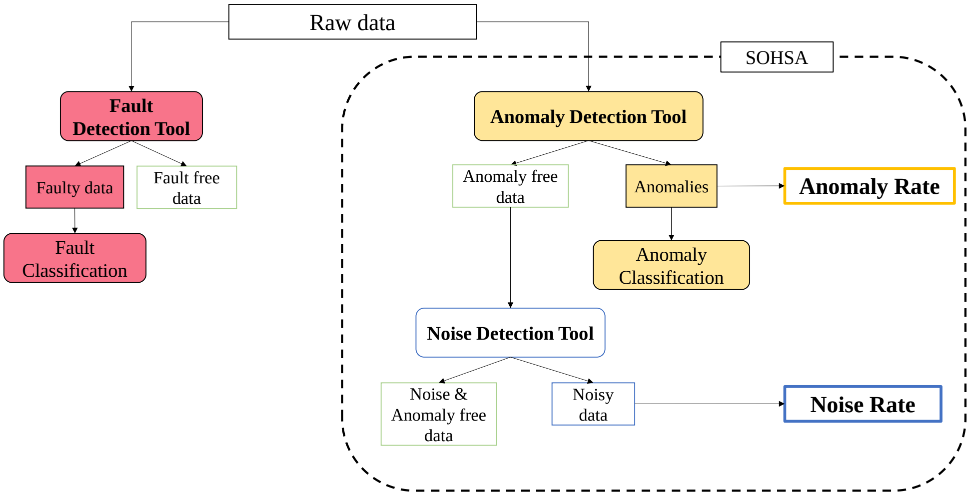


Figure 9 – Overall diagram of I-DCIDS methodology

The SOHSA comprises the Anomaly Detection Tool and Noise Detection Tool, which detect data anomalies and noise, respectively, so that the sensor overall health state can be evaluated. Such piece of information is provided by means of the Anomaly Rate and the Noise Rate. The Fault Detection Tool and SOHSA are described in Paragraphs 3.1.1, 3.1.2 and 3.1.6.

3.1.1. Fault Detection Tool

The Fault Detection Tool allows the detection of seven different faults: out of range, stuck signal, dithering, standard deviation, trend coherence, spike and bias. Fault detection is performed by means of basic mathematical laws that require some acceptability thresholds and parameters that have to be selected by the user.

Figure 2 depicts the logic of the Fault Detection Tool.

The Fault Detection Tool allows the diagnosis of both single and multiple sensors. Multiple sensors can be classified as *redundant* or *correlated*. *Redundant* sensors are installed in the same location and are expected to replicate the same signal. *Correlated* sensors are instead placed in different locations in the same gas turbine section (e.g., gas turbine exhaust section) to measure the same physical quantity. However, because of the different position, the value measured by correlated sensors may not be exactly the same, though a relationship exists between the different signals.

The Gross Threshold Analysis (GTA), the Stuck Signal Analysis (SSA) and the Dithering Analysis (DTA) are performed either in cases where a single sensor is analyzed, or redundant/correlated sensors are available. These three analyses (GTA, SSA and DTA) do not compare signals acquired by different sensors. For this reason, if redundant or correlated sensors are analyzed, the GTA, the SSA and the DTA do not require the same sampling frequency for all the sensors. Conversely, the Standard Deviation Analysis (SDA), Trend Coherence Analysis (TCA), Spike Analysis (SA) and Bias Analysis (BA) can only be performed if redundant or correlated sensors are available. In these cases, the sampling frequency must be the same for all the sensors under evaluation. This aspect is crucial, since all these analyses compare signals acquired by different sensors.

In general, the raw data is analyzed by a specific fault detection analysis according to its configuration parameters. If a fault is detected, its value, the time point of occurrence and the sensor in which it occurred are stored. All the raw data are then processed by the subsequent analyses. This means that *all* raw data are subject to *all* the analyses and, as a consequence, the same data may be labelled as faulty by more than one of the underlying Fault Detection Tools. Moreover, as demonstrated in Manservigi *et al*. [71], the I-DCIDS tool may also detect other sensor fault classes even if their specific detection logic is not implemented.

The outputs of the Fault Detection Tool include fault-free data, plus the identification and classification of any faulty data. Since the I-DCIDS methodology assesses (i) each time point of the fault occurrence, (ii) the sensor in which the faulty event occurred and (iii) the fault class, detailed information regarding sensor reliability is provided.

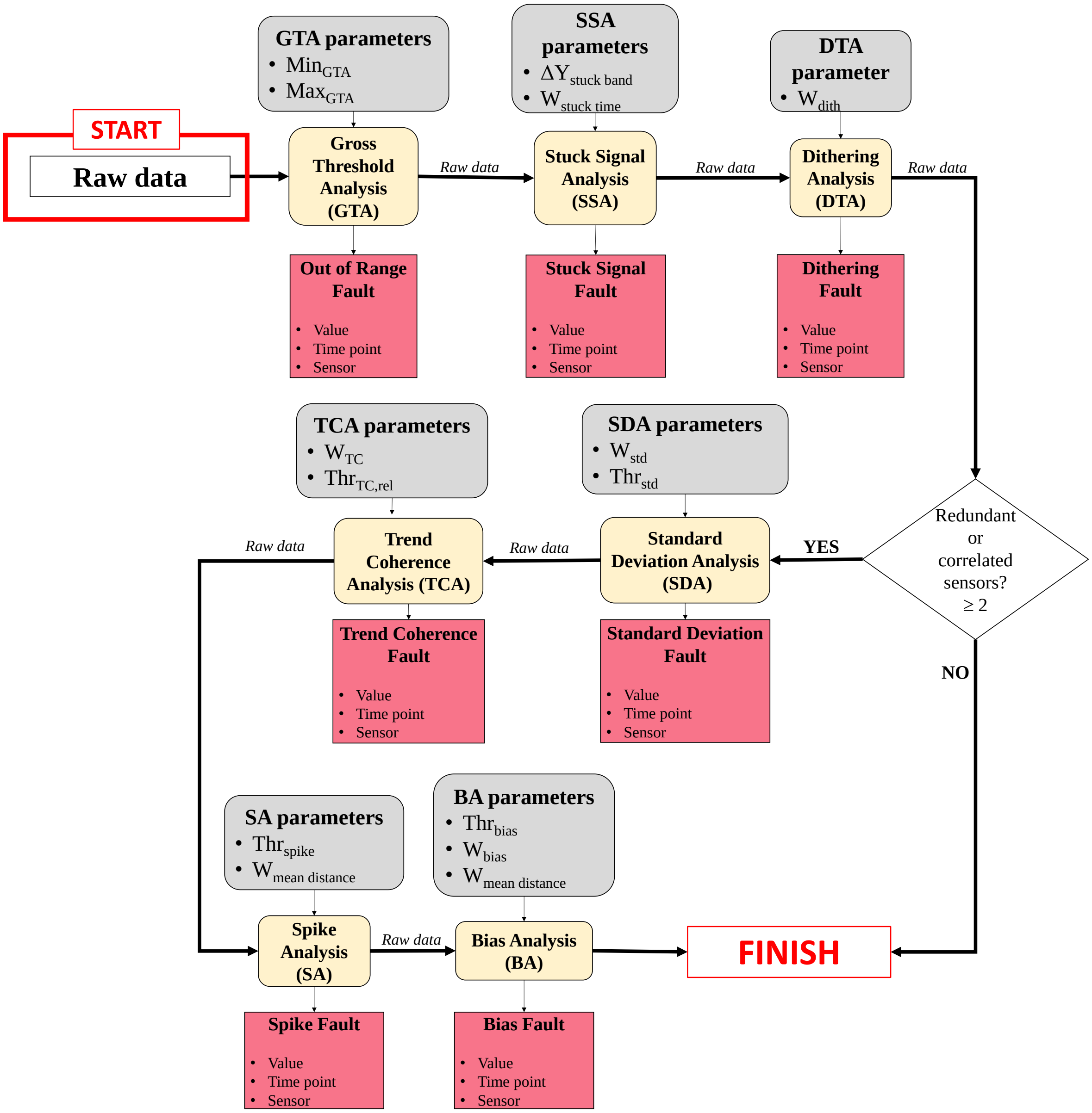


Figure 10 – Diagram of the Fault Detection Tool

Once the analyses of the Fault Detection Tool are finished, raw data are processed by means of the SOHSA, which provides general information concerning data reliability by means of the anomaly and noise rate.

Configuration parameters. Table 2 summarizes all the configuration parameters required by the Fault Detection Tool; the last column indicates the number of sensors required to perform the analysis. All the configuration parameters, five for a single sensor analysis or thirteen if redundant/correlated sensors are available, can be set by the user.

Table 2 – Configuration parameters required by the Fault Detection Tool

Configuration parameters	Analysis	No. of sensors
Max_{GTA}	GTA	≥ 1
Min_{GTA}	GTA	≥ 1
$\Delta Y_{\mathrm{stuck\ band}}$	SSA	≥ 1
$W_{\mathrm{stuck\ time}}$	SSA	≥ 1
W_{dith}	DA	≥ 1
W_{std}	SDA	≥ 2
Thr_{std}	SDA	≥ 2
W_{TC}	TCA	≥ 2
$Thr_{\mathrm{TC,rel}}$	TCA	≥ 2
$W_{\mathrm{mean\ distance}}$	SA, BA	≥ 2
Thr_{spike}	SA	≥ 2
W_{bias}	BA	≥ 2
Thr_{bias}	BA	≥ 2

The GTA requires two parameters: Max_{GTA} and Min_{GTA}, which represent the maximum and minimum acceptable values. These thresholds clearly depend on the physical quantity under investigation.

The SSA involves the parameters $\Delta Y_{\text{stuck band}}$ and $W_{\text{stuck time}}$. The former indicates the minimum allowable variation of the physical quantity; the latter is the minimum number of consecutive time points for stuck signal detection and classification. In fact, a stuck signal fault occurs if the signal does not exceed the $\Delta Y_{\text{stuck band}}$ parameter for more than $W_{\text{stuck time}}$ time points.

The DTA only requires one parameter, i.e., W_{dith}, which is the minimum number of consecutive time points in which a dithering fault must take place.

The subsequent eight configuration parameters are only defined if redundant/correlated sensors are available.

W_{std} and $W_{\text{mean distance}}$ represent windows of consecutive time points in which the standard deviation and the mean distance, respectively, are estimated.

W_{bias} and W_{TC} represent the minimum number of consecutive time points necessary to detect a bias fault and a trend coherence fault, respectively.

The parameter $Thr_{\text{TC,rel}}$ used in the TCA represents the relative threshold for detecting trend coherence faults. If within the W_{TC} window the maximum difference of the physical quantity exceeds the $Thr_{\text{TC,rel}}$ threshold, the W_{TC} time points are labelled as faults.

Finally, Thr_{std}, Thr_{spike} and Thr_{bias} are thresholds applied to isolate the faults in the SDA, SA and BA, respectively. These thresholds can be properly set to tune the filtering capability of the tools.

It should be noted that some configuration parameters, e.g., Max_{GTA} and Min_{GTA}, are strongly dependent on the physical quantity and the type of analyzed sensor. For this reason, they must be properly defined by the user.

Detection of the faulty sensor in multi-sensor analyses. Most of the analyses that involve redundant or correlated sensors (i.e., SDA, SA and BA) detect faulty data through the comparison of sensor measurement signal pairs. The TCA is the only exception because it analyzes all the signal trends at the same time. Figure 11 illustrates the procedure for the detection of the faulty sensor.

Each iteration compares two different signals (signal “i” and signal “j”) by analyzing all the pairs of signals “i-j”. If the fault equations, reported later in this thesis as Eq. (3), Eqs. (8) – (10) and Eqs. (13) – (15), are satisfied in (N-1) pairs of signals (with the exception of the pair “i-i”), the signal “i” is labelled as the faulty signal.

It should be noted that detection of the faulty sensor can only be performed if at least three redundant or correlated sensors are available. In fact, if only two redundant or correlated signals are known, the Fault Detection Tool indicates that a fault occurred at a given time point, but it is not able to specify in which sensor it occurred.

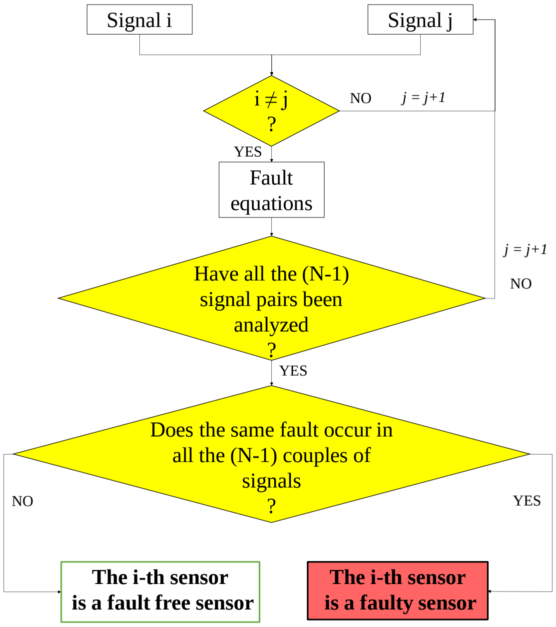


Figure 11 – Faulty sensor identification in multi-sensor analyses

Logic of the fault detection tool. This paragraph presents an in deep insight about the seven analyses carried out by means of the Fault Detection Tool to detect and classify sensor faults.

Gross Threshold Analysis. The GTA can be performed for both a single sensor and redundant/correlated sensors.

The GTA requires two configuration parameters:

- Min_{GTA};
- Max_{GTA}.

Min_{GTA} and Max_{GTA} represent the minimum and the maximum value that a given physical quantity can assume. According to Fig. 12, if data are included within the acceptability range, the GTA considers them reliable and the subsequent data are processed.

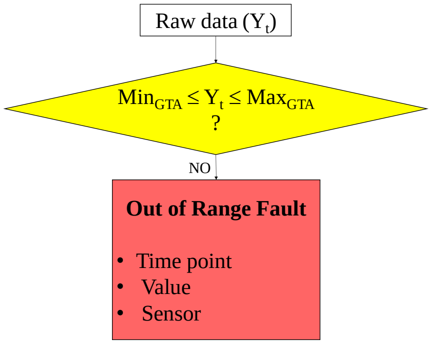


Figure 12 – Logic of the GTA

Conversely, if data reside outside of the defined acceptable range, the sensor is classified as having an out of range fault.

Stuck Signal Analysis. The SSA is performed for stuck signal fault detection.
The SSA is performed both for single and redundant or correlated sensors.
The tuning of the SSA requires two configuration parameters:

- $W_{\text{stuck time}}$;
- $\Delta Y_{\text{stuck band}}$.

$W_{\text{stuck time}}$ is a window of consecutive time points and $\Delta Y_{\text{stuck band}}$ is the minimum variation that the consecutive time points must assume to be considered reliable by the SSA.
Figure 13 shows the SSA logic. $W_{\text{stuck signal}}$ consecutive time points are isolated and the variation of their numerical values (ΔY) is calculated. If ΔY is lower or equal to the chosen threshold ($\Delta Y_{\text{stuck band}}$), the SSA detects all the $W_{\text{stuck time}}$ consecutive data as stuck signal faults.
Regardless of the SSA response, a new data is acquired and the window of $W_{\text{stuck signal}}$ data moves forward. The SSA is then repeated for all sensors.

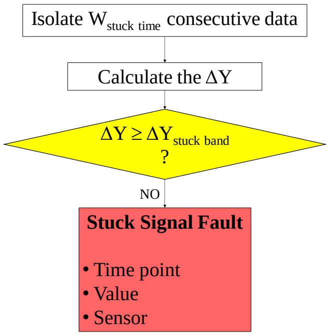


Figure 13 – Logic of the SSA

Dithering Analysis. As in the case of the GTA and the SSA, the analysis carried out by the DTA is performed regardless of the number of available sensors.
According to DTA logic, a dithering fault is detected if at least W_{dith} consecutive data assume two different values. Specifically, a generic data Y_t must be equal to Y_{t-2}, but it must differ from Y_{t-1}. This behavior corresponds to a saw tooth profile.
According to the fault definition considered in this thesis, i.e., saw tooth profile, a dithering fault, is captured by means of the logic shown in Fig. 14.
It should be noted that the DTA can effectively identify dithering faults, regardless of the magnitude of deviation (Y_t - Y_{t-1}).

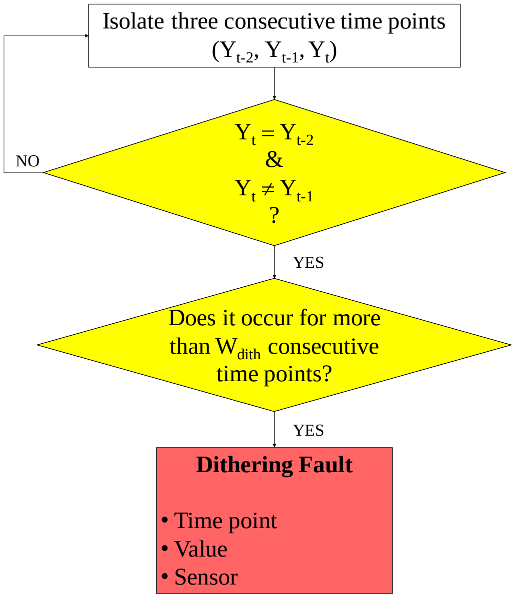


Figure 14 – Logic of the DA

Standard Deviation Analysis. The SDA can only be performed for redundant/correlated sensors. In fact, a standard deviation fault can only be detected if at least two signals are available, so that a comparison can be performed.

The SDA requires two configuration parameters:

- W_{std};
- Thr_{std}.

W_{std} is a window of consecutive time points in which the standard deviation of the data is calculated. The parameter Thr_{std} represents the threshold used to discriminate the standard deviation faults.

The flowchart shown in Fig. 15 shows the logic for detecting standard deviation faults.

The SDA compares the behavior of the last acquired W_{std} data to previous W_{std} data. The standard deviation of the previous data and of the current data is calculated.

Then, for all the pairs of signals, the SDA calculates the $\Delta\sigma_{\text{current}}$ and $\Delta\sigma_{\text{prev}}$ (Eqs. (1) and (2)), which are the differences in the standard deviations of the current and previous data, respectively.

The reliability of the current data is then estimated by means of Eq. (3). In fact, if $\Delta\sigma_{\text{current}}$ exceeds a threshold, the current data show faulty behavior and they are detected as standard deviation faults, since redundant/correlated signals must be characterized by the same standard deviation values.

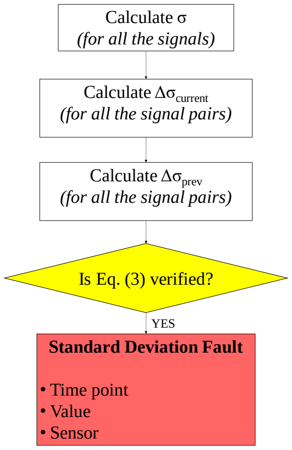


Figure 15 – Logic of the SDA

$$(\Delta\sigma_{\text{current}})_{\text{ij}} = (\sigma_{\text{current}})_{\text{i}} - (\sigma_{\text{current}})_{\text{j}} \qquad \text{i,j} = 1,\ldots, N \qquad (1)$$

$$(\Delta\sigma_{\text{prev}})_{\text{ij}} = (\sigma_{\text{prev}})_{\text{i}} - (\sigma_{\text{prev}})_{\text{j}} \qquad \text{i,j} = 1,\ldots, N \qquad (2)$$

$$|(\Delta\sigma_{\text{current}})_{\text{ij}}| > Thr_{\text{std}} \cdot |(\Delta\sigma_{\text{prev}})_{\text{ij}}| \qquad \text{i,j} = 1,\ldots, N \qquad (3)$$

Thanks to its flexibility and versatility, the SDA tool may also detect drifts and erratic signals, in which PQ statistical properties erroneously vary over time [42,72].

Trend Coherence Analysis. The TCA can only be performed if redundant/correlated sensors are available; in fact, the analyzed signal must be compared to at least one reference sensor.

The TCA requires two configuration parameters:

- W_{TC};
- $Thr_{\text{TC,rel}}$.

The former represents the minimum number of consecutive data necessary to detect the trend coherence fault. The latter is a relative threshold which is applied to isolate the trend coherence faults.

Figure 16 summarizes the logic of the TCA. When new data (Y_t) is acquired, the slope between the time points (t-2) and t is calculated by neglecting the intermediate value.

Thus, the calculation of the slope does not involve consecutive time points to smooth small variations between two consecutive data (that may be due to noise) and capture the overall trend. Though the results are not reported in this thesis for the sake of brevity, a sensitivity analysis was performed on the number of intermediate time points to be ignored in order to calculate the slope.

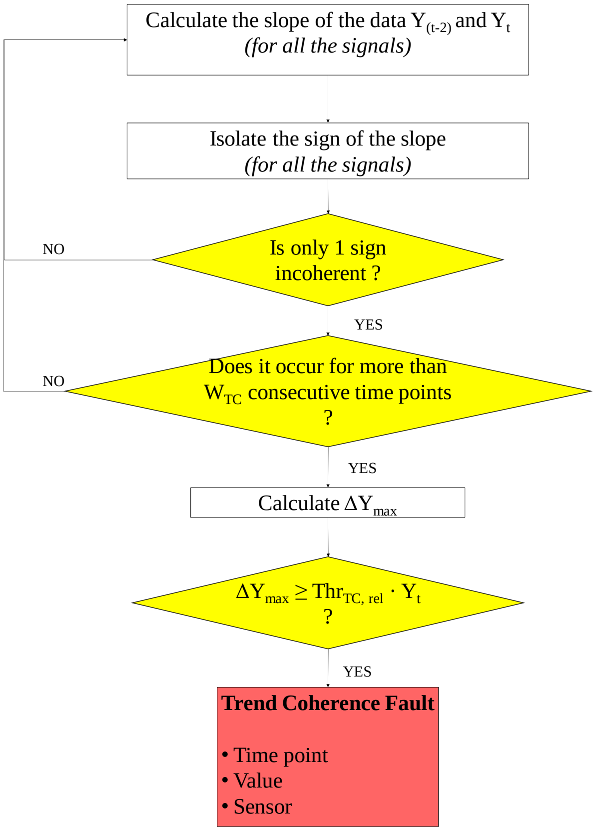


Figure 16 – Logic of the TCA

It was found that neglecting one intermediate time point was the best solution in most of the considered field datasets. The sign of the slope (positive, zero, or negative) is then isolated. This procedure is performed for all the sensors under consideration. If only one signal shows a different slope sign compared to the other signals for at least W_{TC} time points following the time point t (so the physical quantity is equal to Y_t), TCA classifies all these data as candidate trend coherence faults. A new check is then performed by calculating the maximum difference (ΔY_{max}) between the candidates trend coherence faults. If ΔY_{max} exceeds $Thr_{TC,rel}$ multiplied by Y_t, (see Eq. (4)) a trend coherence fault has occurred.

$$\Delta Y_{max} \geq Thr_{TC,rel} \cdot Y_t \tag{4}$$

It should be noted that, thanks to the estimation of the ΔY_{max} value, the TCA also allows the detection of stuck signal faults. In fact, a stuck signal is characterized by an incoherent trend compared to a time-varying signal.

Spike Analysis. In the Fault Detection Tool, the SA can only be performed if at least two sensors are available. In fact, according to SA, a spike is an isolated data point that does not correctly reproduce the behavior of a reference signal. More specifically, the faulty data represents a variation of the physical quantity that is higher than that of a reference signal. The SA requires two configuration parameters:

- $W_{\text{mean distance}}$;
- Thr_{spike}.

The former includes the consecutive data used for the calibration of the algorithm. The latter is a threshold applied to isolate only the major faults. The flowchart in Fig. 17 describes the logic of the spike fault detection.

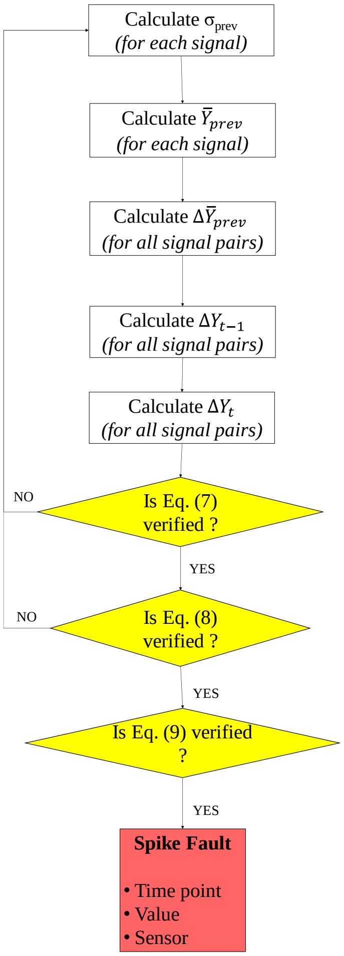


Figure 17 – Logic of the SA

The SA considers ($W_{\text{mean distance}}$ +2) data, of which the last two data points are the current time point t and the time point before the current time point (i.e., t-1). The data window $W_{\text{mean distance}}$ represents the past behavior of the signal, to be used for SA calibration, while the time point (t-1) is analyzed for spike detection.

The standard deviation and the mean value of the previous $W_{\text{mean distance}}$ data are calculated. Then, according to Eqs. (5) – (7), the difference between (i) the mean values of the previous $W_{\text{mean distance}}$ data, (ii) the time point (t-1) and (iii) the current time point t, are calculated for all of the signal pairs.

If Eqs. (8) – (10) are verified, the time point (t-1) is detected as a spike fault. In fact, according to Eqs. (8) and (9), a spike fault occurs if the data distance at time point (t-1) is higher than (i) the previous ones and (ii) the standard deviation of the previous data multiplied by the Thr_{spike} coefficient. Finally, the current data are analyzed.

According to Eq. (10), the difference at time point t between the current data must be comparable to that of the previous data. The last condition is fundamental for spike detection; in fact, this allows the detection of isolated faults.

$$(\Delta\bar{Y}_{\text{prev}})_{\text{ij}} = (\bar{Y}_{\text{prev}})_{\text{i}} - (\bar{Y}_{\text{prev}})_{\text{j}} \qquad \text{i,j} = 1,\ldots, N \qquad (5)$$

$$(\Delta Y_{\text{t-1}})_{\text{ij}} = (Y_{\text{t-1}})_{\text{i}} - (Y_{\text{t-1}})_{\text{j}} \qquad \text{i,j} = 1,\ldots, N \qquad (6)$$

$$(\Delta Y_{\text{t}})_{\text{ij}} = (Y_{\text{t}})_{\text{i}} - (Y_{\text{t}})_{\text{j}} \qquad \text{i,j} = 1,\ldots, N \qquad (7)$$

$$\Delta Y_{\text{max}} \geq Thr_{\text{TC,rel}} \cdot Y_{\text{t}} \qquad (8)$$

$$|(\Delta Y_{\text{t-1}})_{\text{ij}}| > Thr_{\text{spike}} \cdot (\sigma_{\text{prev}})_{\text{i}} \qquad (9)$$

$$|(\Delta Y_{\text{t}})_{\text{ij}}| \leq |(\Delta\bar{Y}_{\text{prev}})_{\text{ij}}| \qquad (10)$$

Bias Analysis. Based on its definition, the I-DCIDS tool detects a bias fault by comparing pairs of sensors. For this reason, the BA can only be performed if redundant/correlated sensors are available.

The BA includes three configuration parameters:

- $W_{\text{mean distance}}$;
- Thr_{bias};
- W_{bias}.

$W_{\text{mean distance}}$ and W_{bias} represent windows of consecutive data and Thr_{bias} is a threshold applied for the detection of the bias faults.

The flowchart in Fig. 18 shows the logic of the BA.

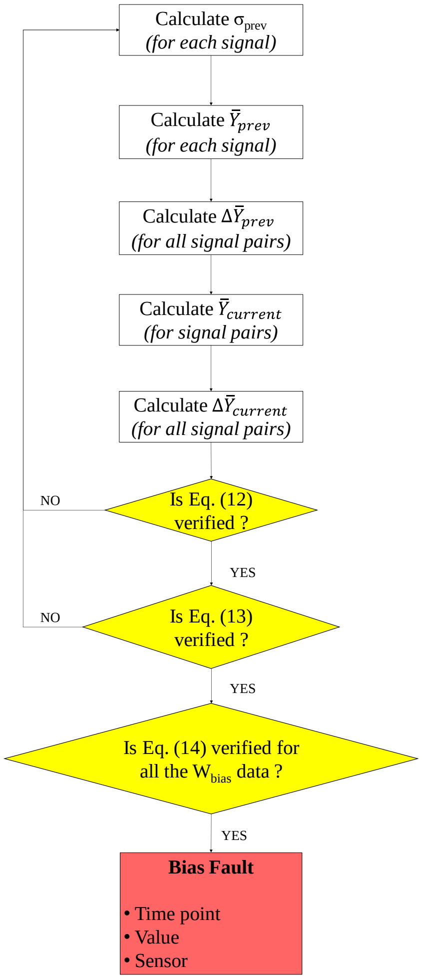


Figure 18 – Logic of the BA

The BA compares two windows of consecutive data. The previous $W_{\text{mean distance}}$ data are used for the calibration of the algorithm. Conversely, the current W_{bias} data are analyzed. For all the signals, the standard deviation and the mean value of the previous data are calculated. For all the signal pairs, the difference between the mean value of the previous and current data are calculated, according to Eqs. (11) and (12).

The bias fault detection depends on the difference between the mean of the current data. First, Eqs. (13) and (14) are applied to isolate the time points in which candidate faults could occur. Then, if Eqs. (13) and (14) are satisfied, the difference between the two signals is calculated at each time step of the data. According to Eq. (15), if all the current data deviate from the previous behavior, it means that a step change has occurred and a bias fault is identified.

$$(\Delta\bar{Y}_{\mathrm{prev}})_{\mathrm{ij}} = (\bar{Y}_{\mathrm{prev}})_{\mathrm{i}} - (\bar{Y}_{\mathrm{prev}})_{\mathrm{j}} \qquad \mathrm{i,j} = 1,\ldots, N \tag{11}$$

$$(\Delta\bar{Y}_{\mathrm{current}})_{\mathrm{ij}} = (\bar{Y}_{\mathrm{current}})_{\mathrm{i}} - (\bar{Y}_{\mathrm{current}})_{\mathrm{j}} \qquad \mathrm{i,j} = 1,\ldots, N \tag{12}$$

$$|(\Delta\bar{Y}_{\mathrm{current}})_{\mathrm{ij}}| > |(\Delta\bar{Y}_{\mathrm{prev}})_{\mathrm{ij}}| \tag{13}$$

$$|(\Delta\bar{Y}_{\mathrm{current}})_{\mathrm{ij}}| > Thr_{\mathrm{bias}} \cdot (\sigma_{\mathrm{prev}})_{\mathrm{i}} \tag{14}$$

$$|(\Delta Y_{\mathrm{current}})_{\mathrm{ij}}| > |(\Delta\bar{Y}_{\mathrm{prev}})_{\mathrm{ij}}| \tag{15}$$

3.1.2. Fault Detection Tool parameters

The Fault Detection Tool requires thirteen configuration parameters (shown in Table 2), which represent the acceptability thresholds and windows of consecutive data required for the calibration of the tool.

The five configuration parameters required by GTA, SSA and DA, i.e., Max_{GTA}, Min_{GTA}, $\Delta Y_{\mathrm{stuck\ band}}$, $W_{\mathrm{stuck\ time}}$ and W_{dith}, depend on the considered PQ and application.

Therefore, a sensitivity analysis on the remaining eight parameters is performed in this thesis to evaluate the optimal tuning of the tool. In fact, this thesis is aimed at tuning the most challenging fault detection tools (i.e., SDA, TCA, SA and BA), since fine tuning is required. In addition, the thesis aims to provide general configuration parameters, which are independent of the considered physical quantity. Therefore, GTA, SSA and DTA are not investigated, since the detection of out of range, stuck signal and dithering faults is affected by the considered physical quantity.

The effect of different values of W_{std}, W_{TC}, $W_{\mathrm{mean\ distance}}$ and W_{bias}, i.e., 5, 10, 15, 30, is analyzed in conjunction with four different acceptability thresholds, i.e., Thr_{std}, $Thr_{\mathrm{TC,rel}}$, Thr_{spike} and Thr_{bias}. For the sake of brevity, the results of the sensitivity analysis on the windows W is not reported. In fact, the output of the Fault Detection Tool generally proves independent of the windows W. For this reason, the sensitivity analysis focuses on the acceptability thresholds required by the multi-sensor analyses, i.e., Thr_{std}, $Thr_{\mathrm{TC,rel}}$, Thr_{spike}, Thr_{bias}, while the other configuration parameters are assumed to be constant, as follows:

- $W_{\mathrm{std}} = 5$;
- $W_{\mathrm{TC}} = 5$;

- $W_{\text{mean distance}} = 5$;
- $W_{\text{bias}} = 5$.

All the W_{std}, W_{TC}, $W_{\text{mean distance}}$ and W_{bias} values are set equal to 5, meaning that the given fault class is detected if the fault occurs for at least 5 consecutive time points.

The sensitivity analysis carried out in this thesis to tune Thr_{std}, $Thr_{\text{TC,rel}}$, Thr_{spike} and Thr_{bias} is performed by means of four different acceptability threshold values, which are progressively increased.

According to Eq. (3), Eq. (4), Eq. (9) and Eq. (14), it is clear that increasing Thr_{std}, $Thr_{\text{TC,rel}}$, Thr_{spike} and Thr_{bias} leads to the detection of progressively severe faults.

Table 3 reports the values considered for the sensitivity analysis. The Thr_{std}, Thr_{spike} and Thr_{bias} values are equal to 3, 5, 10 and 15, meaning that a fault is detected if the standard deviation value is exceeded three, five, ten or fifteen times, respectively. Thr_{std}, Thr_{spike} and Thr_{bias} equal to 10 or 15 are clearly dramatically high. Nevertheless, the calibration of SDA, TCA, SA and BA is based on a few data, since the length of all windows is assumed equal to 5.

Table 3 – Configuration parameters considered in the sensitivity analysis

Thr_{std}	3	5	10	15
$Thr_{\text{TC,rel}}$	1.0 %	1.5 %	2.0 %	2.5 %
Thr_{spike}	3	5	10	15
Thr_{bias}	3	5	10	15

Thus, if only one candidate fault occurs, its weight is greater compared to the same case study in which the tool calibration exploits much more data. Four different $Thr_{\text{TC,rel}}$ values (in the range from 1.0% to 2.5%) are also analyzed.

Finally, it should be noted that all the fault analyses (i.e., SDA, TCA, SA and BA) are completely independent from each other. Thus, the sensitivity analysis allows identification of the optimal combination of the four parameters.

The most suitable setting of the configuration parameters is evaluated by means of the subsequent three criteria (C):

- C #1: only the faulty sensor (one or more) must be detected by the Fault Detection Tool, i.e., only true positives are identified;

- C #2: the *onset* of the fault (one or more) must be detected, so that a warning can be raised in a timely manner;
- C #3: faults are not detected within the reliable sensors, i.e., no false positive is identified.

These three criteria were selected since the I-DCIDS methodology is validated in this thesis on experimental data. According to Cejnek and Bukovsky [72], several difficulties can occur when real data are considered. For example, the type and level of noise affecting real data is hard to evaluate. In addition, the exact time point of fault occurrence cannot be precisely annotated. For these reasons, in the literature most sensor detection tools are tested on reliable or simulated data ([23, 38 – 40, 43, 64, 72]), in which some faults are implanted, and their effectiveness is assessed by means of quantitative indices, e.g., true positive rate and false positive rate.

Conversely, general criteria to evaluate detection effectiveness on experimental data are not available.

3.1.3. Field data

The I-DCIDS tool is validated in this thesis by means of four datasets acquired from Siemens gas turbines. Three of the faulty datasets (dataset #1, #2 and #4) are characterized by at least one sensor that is clearly faulty, while the identification of the candidate faulty sensors in the dataset #3 is not as straightforward, as discussed below.

Dataset #1 through #4 are employed to tune the optimal values of the configuration parameters, which are further tested in dataset #5, which includes only fault-free sensors.

All the field data are characterized by distinctive characteristics, since the number of available sensors differs and the GT load varies in different ways. In addition, the effect of sampling frequency on the I-DCIDS methodology is investigated. In fact, the sampling frequency is different in all datasets and in dataset #2 through #4 it varies over time. Thus, the selected datasets represent test cases that are also suitable for evaluating the I-DCIDS methodology, even by processing data with different sampling frequency.

For a given dataset, all the redundant/correlated sensors are acquired with the same sampling frequency, so that multi-sensor analyses are allowed.

More detailed information about the considered sensors cannot be disclosed for confidentiality reasons. In fact, even though the analyses were performed on the dimensional PQs, the trends reported in this thesis refer to their nondimensional values, calculated with respect to the maximum value recorded in each dataset.

Dataset #1. Though it is hard to grasp from Fig. 22, dataset #1 includes three sensors. In fact, sensor #1 and sensor #2 record almost (but not exactly) the same value.

As can be observed, the non-dimensional PQ measured by both sensors #1 and #2 significantly decreases between data 6 and 16 and again between data 74 and 81.

In this dataset, the sampling frequency is equal to one per minute.

Given the nature of the PQ, the I-DCIDS tool should detect sensor #3 as a completely faulty sensor. In fact, it always measures a constant and negative value.

Thus, all sensor #3 data should be labelled as both an out of range fault and a stuck signal fault.

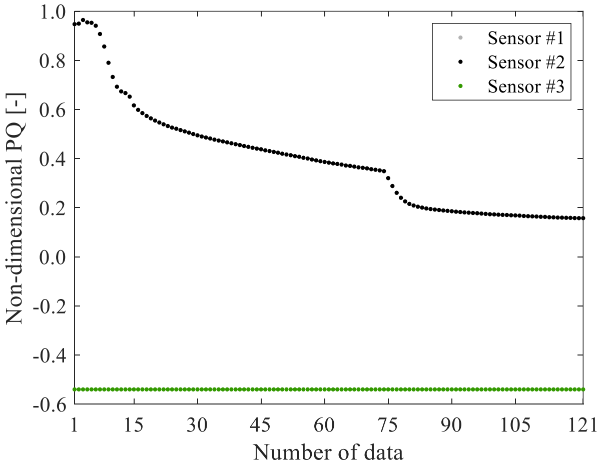


Figure 22 – Dataset #1

Dataset #2. Figure 23 shows dataset #2, which is composed of two sensors. Sensor #2 shows an almost constant trend from time point 1 to time point 400.

A rapid transient then occurs and the nondimensional PQ continues to decrease until the end of the dataset.

In this dataset, the sampling frequency varies over time. In fact, the sampling frequency is in the range of one per minute and one per 43 minutes.

Sensor #1 is the candidate faulty sensor. In fact, sensor #1 and sensor #2 show approximately the same behavior in the first 42 data points.

Subsequently, sensor #1 shows more irregular and scattered behavior until the time point 400. Thus, at least one standard deviation fault should be detected.

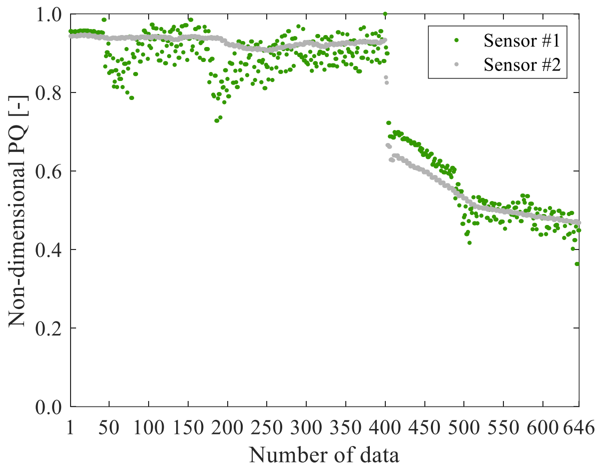


Figure 23 – Dataset #2

Dataset #3. The dataset #3, shown in Fig. 24, includes three sensors, whose sampling frequencies are in the range of one per minute and one per 17 minutes.

This dataset is characterized by three steady-states, separated by two rapid transients, corresponding to different GT operating conditions. In fact, from time point 1 to 241 and from time point 425 to the end, the nondimensional PQ is lower than 0.1.

During the second steady-state, the values of the nondimensional PQ are in the range of 0.94 - 0.95 (sensor #1), 0.93 - 0.94 (sensor #2) and 0.89 - 0.90 (sensor #3).

However, even though all the signals show the same trend, their mean distance during the second steady-state is greater than that of the first and third steady-state.

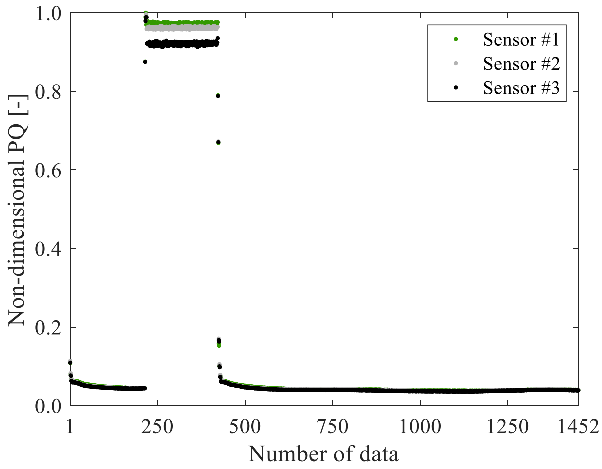


Figure 24 – Dataset #3

Since the Fault Detection Tool makes use of antecedent data for calibrating the rejection rules, standard deviation fault and bias fault may be detected in sensor #3, which measures the lowest values of the PQ.

Dataset #4. Dataset #4 is shown in Fig. 25 and includes seven correlated sensors. The sampling frequency of all sensors varies in the range of one per minute and one per 175 minutes. The frequent and rapid variation of GT operating conditions represents the most challenging feature of this dataset. Sensor #2 is clearly the faulty sensor since its behavior differs significantly from the other sensors, starting from time point 105. In fact, while the reliable sensors vary over time, sensor #2 usually measures the same constant value. Moreover, between time point 107 and 111, time point 548 and 551 and time point 733 and 739, sensor #2 includes a bias fault.

Thus, the Fault Detection Tool is expected to detect a stuck signal fault, a trend Coherence fault and a bias fault, all occurring in sensor #2.

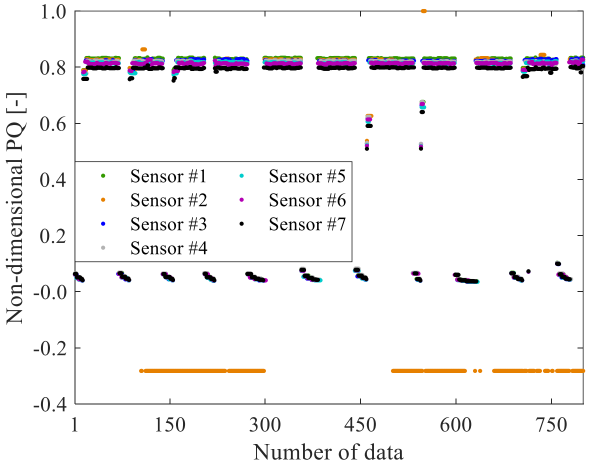


Figure 25 – Dataset #4

3.1.4. Optimal tuning of Fault Detection Tool parameters

This section discusses the results of the sensitivity analysis performed on the configuration parameters listed in Table 2. As previously highlighted, GTA, SSA and DTA are not considered in this thesis, since the definition of the corresponding parameters has to account for the PQ under consideration. The faults identified by the Fault Detection Tool setup with the optimal tuning are then reported and analyzed.

Dataset #1. *SDA results.* If Thr_{std} = 3, SDA successfully detects two standard deviation faults in sensor #3; the two faults of sensor #3 occur between the time points 6 to 10 and between the time points 71 to 75. It should be noted that SDA correctly detects the faulty event since sensor #1 and sensor #2 measure a significant variation of the PQ, while the signal of sensor #3 is constant. When Thr_{std} = 5 or Thr_{std} = 10, only the time frame in which the sensor fault occurs (i.e., time points 6 to 10) is detected by SDA. If the acceptability threshold is further increased, i.e., Thr_{std} = 15, SDA does not detect any fault.

TCA results. Regardless of the $Thr_{\mathrm{TC,rel}}$ values, all time points of sensor #3 are detected by TCA.

SA results. Regardless of the acceptability threshold, i.e., Thr_{spike}, SA does not detect any faults.

BA results. BA does not detect any faults.

In summary, all three acceptability criteria are satisfied. In fact, only the faulty sensor is detected (C #1), the time point at which the fault starts (C #2) is correctly identified by both SDA (Thr_{std} = 3, 5, 10) and TCA, while only TCA detects the fault occurrence if $Thr_{\mathrm{TC,rel}}$ is increased. Finally, faults are not detected in reliable sensors (C #3).

Dataset #2. *SDA results.* SDA results highlight that the beginning of the fault occurrence between data points 41 and 45 is correctly detected if Thr_{std} = 3. The results obtained by means of Thr_{std} = 5 differ slightly from the case in which Thr_{std} = 3. The fault detection by means of Thr_{std} = 10 and Thr_{std} = 15 proves to be very similar to Thr_{std} = 3 and Thr_{std} = 5. As a general comment, false positives are not detected. However, some minor faults are no longer detected by increasing Thr_{std}.

The beginning of fault occurrence is correctly detected by means of SDA.

TCA results. According to $Thr_{\mathrm{TC,rel}}$ = 1.0%, several trend coherence faults are detected between time points 100 and 400 and between time points 500 until the end of the dataset. All the trend coherent faults are confirmed by setting $Thr_{\mathrm{TC,rel}}$ = 1.5%, $Thr_{\mathrm{TC,rel}}$ = 2.0% and $Thr_{\mathrm{TC,rel}}$ = 2.5%.

SA results. If Thr_{spike} = 3, several spike faults are detected across the dataset; a few of them are also detected before fault occurrence (i.e., false positives). Most spike faults are confirmed by using Thr_{spike} = 5 or Thr_{spike} = 10. If Thr_{spike} = 15, several spike faults are detected during the faulty behavior. As a consequence, Thr_{spike} should be increased to reduce an excessive number of warning alarms. However, a few false positives are detected before the faulty event.

BA results. Several bias faults are detected within the dataset by means of Thr_{bias} = 3. In particular, bias faults are detected at the beginning of the fault. Moreover, most bias faults are confirmed if the acceptability threshold is increased, i.e., Thr_{bias} = 5, Thr_{bias} = 10 and Thr_{bias} = 15, but the onset of the fault occurrence is always detected.

As a general comment, it should to be considered that the I-DCIDS tool cannot detect the faulty sensor if only two sensors are available.

Thus, if at least one fault is detected, both sensors are labeled as faulty. This is also verified for dataset #2, which only includes two sensors. However, even though C #1 and C #3 cannot be verified, the Fault Detection Tool correctly provides a warning to the user.

Dataset #3. *SDA results.* By imposing Thr_{std} = 3, standard deviation faults are detected in all steady-state regions of the dataset. Thus, the selected threshold proves to be inappropriate, since some false positives are detected. In all sensors, faults are detected as soon as rapid transients occur at data 215 and 423. If Thr_{std} = 5, standard deviation faults decrease by more than 50 % in all sensors. All the standard deviation faults detected during the second steady-state are confirmed, while most faults previously detected during the third steady-state are no longer identified. Imposing Thr_{std} = 10 allows a 30 % further reduction in the data labelled as faulty. All sensors are characterized by standard deviation faults at the same time point at the beginning of the step change. Moreover, in sensor #1, some data included in the second transient are also classified as faults. SDA results do not vary significantly if the acceptability threshold Thr_{std} is increased to 15.

TCA results. An unexpected trend coherence fault is detected during the first steady-state of sensor #1 if $Thr_{\text{TC,rel}}$ = 1.0%, but, if the TCA acceptability threshold is increased, trend coherence faults are no longer identified.

SA results. The use of Thr_{spike} = 3 detects spike faults in all sensors, but most of them occur during the third steady-state. If the acceptability threshold is increased, i.e., Thr_{spike} = 5, SA results do not vary significantly with respect to Thr_{spike} = 3. By means of Thr_{spike} = 10, only a few spike faults are detected in sensors #1 and #2; on the other hand, more spike faults (18 out of 1,452 data points) are identified in sensor #3. If Thr_{spike} is increased to 15, spike faults are reduced in all sensors and sensor #3 shows five faults, while only one fault is detected in sensor #1.

BA results. If Thr_{bias} = 3 and Thr_{bias} = 5, bias faults are highlighted in sensor #1 and sensor #3, mainly during the third steady-state. If the BA acceptability threshold is increased,

i.e., Thr_{bias} = 10, only two bias faults are detected in sensor #3 during the second steady-state. Bias faults are reduced if Thr_{bias} = 15; in fact, only one bias fault is detected in sensor #3.
As a result, sensitivity analyses prove that the time point in which the step change occurs is correctly detected in all sensors by means of SDA (C #2). In sensor #3, five data are also classified as spike and bias faults.

Dataset #4. *SDA results.* If Thr_{std} = 3, all sensors are characterized by standard deviation faults, by including several false positives in reliable sensors. Thus, similarly to dataset #3, the investigated Thr_{std} value suggests that the threshold should be increased to reduce or remove false positives. The faulty sensor (i.e., sensor #2) includes the highest number of standard deviation faults compared to the reliable sensors; the beginning of the fault is also correctly detected. If the SDA acceptability threshold is slightly increased, i.e., Thr_{std} = 5, the standard deviation fault rate is reduced in all sensors, but the beginning of the fault in sensor #2 is confirmed. Standard deviation faults are slightly reduced if Thr_{std} = 10 or Thr_{std} = 15, but the sensor #2 fault and its beginning are confirmed. Regardless of the Thr_{std} values, SDA detects some faults in all sensors when the PQ significantly changes.

TCA results. Regardless of the value of $Thr_{\text{TC,rel}}$, TCA only detects a few faults and only in sensor #2.

SA results. Imposing Thr_{spike} = 3 allows the detection of some spike faults in the faulty sensor trend when the measured PQ is equal to the minimum value, while sensor #4 only includes a few spike faults. On the contrary, SA does not detect faults in the other reliable sensors. The increased Thr_{spike} value, i.e., Thr_{spike} = 5, only allows the detection of spike faults in sensor #2 and the beginning of the faulty trend is correctly detected. SA results do not vary if Thr_{spike} = 10 or $\text{Thr}_{\text{spike}}$ = 15.

BA results. A few faults in sensor#1 and sensor #7 are detected if Thr_{bias} = 3. Moreover, in sensor #2 (i.e., the faulty sensor), some bias faults are detected when the PQ assumes the minimum value and the beginning of the fault is correctly detected. If the acceptability threshold is increased, i.e., Thr_{bias} = 5, Bias Faults are reduced, but the beginning of the fault is still detected. If Thr_{bias} = 10, BA results do not vary significantly, with the exception of sensor #1, in which only a few data are labelled as bias faults. Using Thr_{bias} = 15 allows the detection of only the faulty sensor once the fault occurs.

Optimal tuning. On the basis of the results obtained by means of the sensitivity analysis, the optimal configuration parameters to perform SDA, TCA, SA and BA are listed below:

- $Thr_{\text{std}} = 10$;
- $Thr_{\text{TC, rel}} = 1.5\%$;
- $Thr_{\text{spike}} = 15$;
- $Thr_{\text{bias}} = 15$.

It should be noted that the tuning parameter configuration stands regardless of the considered variable type, as demonstrated in this thesis by employing datasets collected in different gas turbine sections and engines.

The faults identified by means of the optimal configuration parameters are shown in Fig. 26 through Fig. 29.

As can be seen, Fig. 26 proves that fault occurrence is successfully detected. In fact, the Fault Detection Tool detects all sensor #3 data in dataset #1.

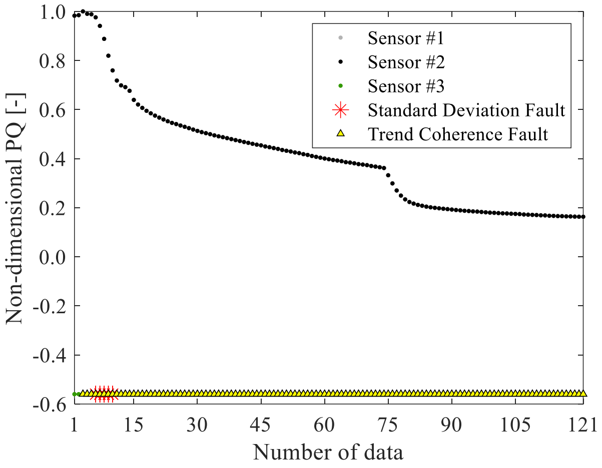


Figure 26 – Fault detection in dataset #1

In dataset #2 (Fig. 27), the SDA correctly detects the incipient standard deviation fault, which continues to the end of the dataset. This is further confirmed by the fact that several trend coherence, spike and bias faults are also detected. However, Fig. 27 also shows that some false positives are detected by the spike analysis before sensor fault occurrence.

Dataset #3 and dataset #4 (Fig. 28 and Fig. 29) prove to be the most challenging datasets since abrupt transients occur. In both datasets, once a transient occurs, the Fault Detection Tool detects faults in all sensors, but, as discussed in the following subparagraph, detection effectiveness can be enhanced by limiting the analysis to steady-states.

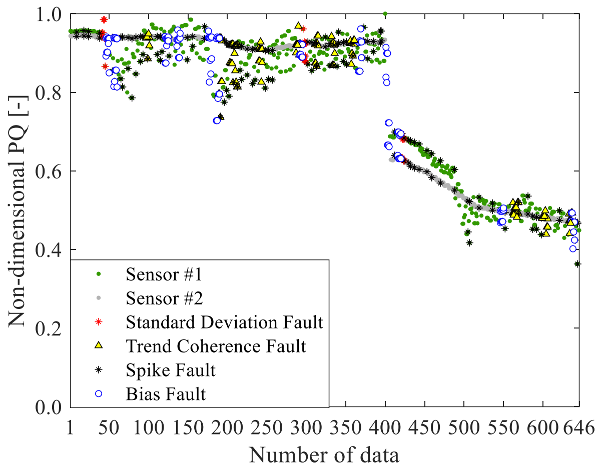


Figure 27 – Fault detection in dataset #2

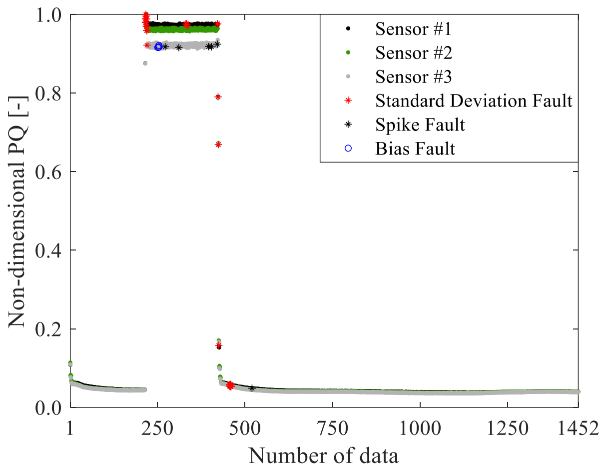


Figure 28 – Fault detection in dataset #3

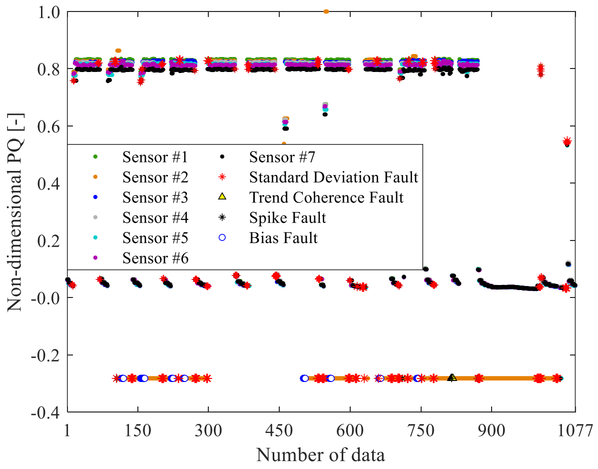


Figure 29 – Fault detection in dataset #4

SDA results on steady-states. According to the sensitivity analyses on dataset #4, SDA is affected by the frequent and rapid changes of GT operating conditions. In fact, SDA usually detects some faults in all the reliable sensors, when step changes occur.

Nevertheless, sensor #2 proves to be the sensor with the highest fault rate. This suggests that a more effective SDA analysis could be performed by neglecting transients and focusing on steady-states in which the measured value of the PQ in reliable sensors (sensors #1 and #3 through #7) is greater than 0.7. The results of the SDA performed on filtered data are shown in Fig. 30.

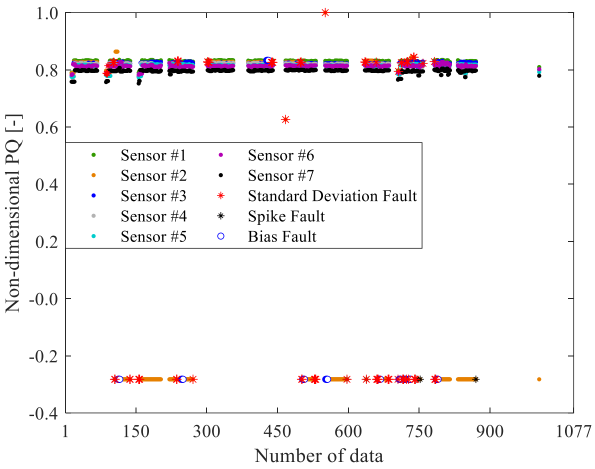


Figure 30 – SDA results on filtered data

(value measured by sensors #1 and #3 through #7 greater than 0.7)

The main outcome is that faults are only detected by the SDA in sensor #2. Moreover, these faults are detected when an actual standard deviation fault occurs in sensor #2, according to the definition provided in Eqs. (1-3).

Table 6 provides a description of the three degrees of fulfillment (high, medium, low) of the three criteria outlined above.

In particular, the third column in Table 7 indicates the fault which is expected to be detected by the four fault analyses.

As can be seen in this table, the Fault Detection Tool generally proves its effectiveness and is therefore validated. In fact, the three criteria are usually met with a high degree of fulfillment (✓✓). This notation is used in Table 7 to summarize the effectiveness of the tuned parameters Thr_{std}, $Thr_{\text{TC, rel}}$, Thr_{spike}, Thr_{bias}.

Table 6 – Degree of fulfilment of criteria #1, #2 and #3

	✓✓	✓	✗
C #1	Only the faulty sensor is detected	The faulty sensor and a few faults in the other sensors are detected	The faulty sensor is not detected
C #2	The beginning of the fault is detected	The tool detects the fault some time before/after its beginning	The beginning of the fault is not detected
C #3	No fault is detected in the reliable sensors	Only a few faults are detected in the reliable sensors	Several faults are detected in the reliable sensors

Table 7 – Effectiveness of the Fault Detection Tool

	Analysis	**Fault?**	**C #1**	**C #2**	**C #3**
Dataset #1	SDA	YES	✓✓	✓✓	✓✓
	TCA	YES	✓✓	✓✓	✓✓
	SA	NO			✓✓
	BA	NO			✓✓
Dataset #2	SDA	YES		✓✓	
	TCA	YES		✓	
	SA	YES		✓✓	
	BA	YES		✓✓	
Dataset #3	SDA	YES	✓	✓✓	✓
	TCA	NO			✓✓
	SA	NO			✓
	BA	YES	✓	✓	✓✓
Dataset #4	SDA	YES	✓	✓✓	✓
	TCA	YES	✓✓	✓	✓✓
	SA	YES	✓✓	✓	✓✓
	BA	YES	✓	✓✓	✓

Finding the optimal values of the parameters used for tuning the Fault Detection Tool represents the main outcome of this thesis and a guideline for fault identification.

It should be noted that these guidelines can be considered of general validity, since optimal tuning was inferred by means of heterogeneous trends representative of very different scenarios that can occur in field applications. Moreover, the sensitivity analyses proved that the I-DCIDS approach is a suitable methodology for fault detection and isolation of GT sensors.

It is worth noting that the I-DCIDS approach includes numerous fault detection analyses, allowing the detection of the same faulty data by means of different analyses. For example, stuck signal faults can also be detected as trend coherence faults and contemporarily as standard deviation faults, as demonstrated in this thesis. Thus, reliable and effective fault detection is guaranteed.

3.1.5. Validation of the Fault Detection Tool parameters

This section deals with further validation of the Fault Detection Tool parameters, whose optimal values were selected by means of the sensitivity analysis carried out in Paragraph 3.5.4. The validation of the Fault Detection Tool parameters is achieved by means of dataset #5, by evaluating false positives.

Similarly to dataset #1 through #4, dataset #5 includes field data time series acquired from one Siemens gas turbine in operation.

The field dataset includes temperature measurements taken from thirteen correlated sensors at combustion chamber shaped section (“turbine interduct temperature”) and used to detect successful ignition by means of a temperature rise.

The time series was collected with a one-minute sampling frequency and consists of 800 observations, which refer to steady-state conditions.

For confidentially reasons, field data reported in Fig. 31 were normalized by means of the first datum measured by a given sensor.

The time series data approximately vary from 0.85 to 1.05, though most data lie in the range from 0.95 to 1.00. Sensor time series #1 and #12 are characterized by the lowest values, while sensors #2, #9 and #11 include the highest values.

The available raw data do not include major faults, as can be inferred from visual inspection. Thus, dataset #5 allows the suitability of the optimal configuration parameters (i.e., $Thr_{\mathrm{std}} = 10$, $Thr_{\mathrm{TC,\ rel}} = 1.5\%$, $Thr_{\mathrm{spike}} = 15$ and $Thr_{\mathrm{bias}} = 15$) to be tested using reliable field data.

Similarly to dataset #1 through dataset #4, all the W_{std}, W_{TC}, $W_{\mathrm{mean\ distance}}$ and W_{bias} values are set equal to 5.

As a result. the Fault Detection Tool applied to this dataset does not detect any fault, i.e., false positives are null.

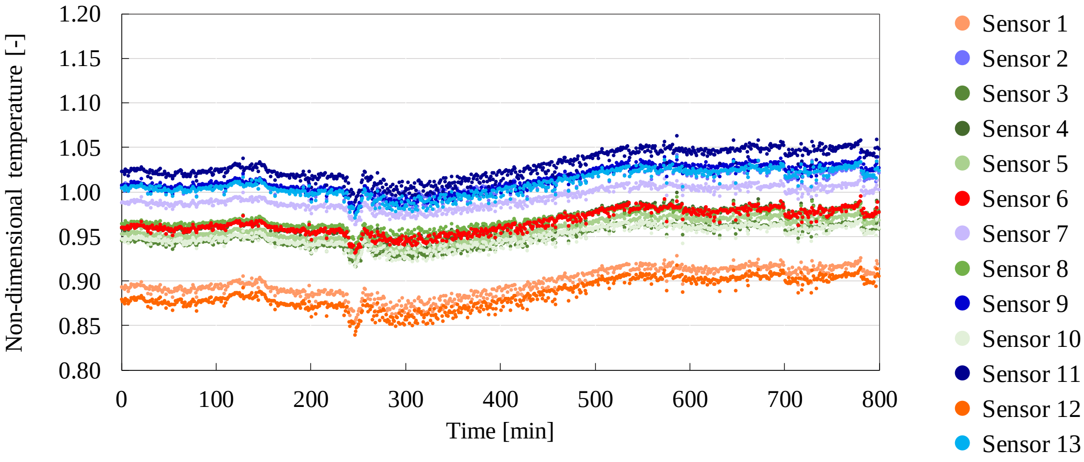


Figure 31 – Dataset #5 [35]

3.1.6. Sensor Overall Health States Analysis

Sensor Overall Health State Analysis (SOHSA) provides an overall evaluation of sensor health state by means of two tools, namely Anomaly Detection Tool and Noise detection Tool.

According to the pipeline defined by the I-DCIDS tool (see Fig. 9), all raw data are first analyzed by the Anomaly Detection Tool and then anomaly free data are analyzed by the Noise Detection Tool.

An in deep insight about both Anomaly Detection Tool and Noise Detection Tool is provided as follows.

Anomaly Detection Tool. As depicted in Fig. 19, the Anomaly Detection Tool comprises two algorithms, namely Anomaly Detection Algorithm (ADA) and Anomaly Classification Algorithm (ACA). Raw data are processed by ADA by means of user-specified thresholds [1]. As a result, raw data are distinguished between anomaly free data and anomalous observations. The former subsequently feeds the Noise Detection Tool, while the latter are further analyzed according to their characteristics.

Anomalies, together with the respective characteristics, serve as inputs to ACA, which does not require any additional user specified parameter to perform its analysis.

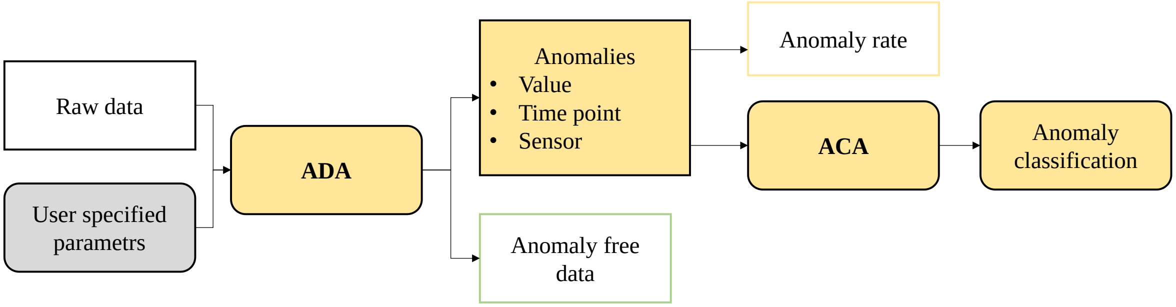


Figure 19 – Anomaly Detection Tool structure

It is worth mentioning that the ADA is also provided with one additional capability, which is referred as sensor muting. In field applications, some sensors may produce anomalous observations with sporadic reliable observations.

The presence of these isolated reliable readings in the post processing time series may produce noisy and unclean results that can be difficult to interpret during trend or diagnostic analyses. The sensor muting capability is designed to tackle this issue.

After the application of ADA, the user can reject all observations of specific sensors that proved to be faulty for most of the time. In this way, all measurements from "muted".

In the following sections, the two kernels of the Anomaly Detection Tool, i.e., ADA and ACA, are described in detail. Guidelines and best practices for the tuning of each methodology implemented in ADA are provided as well.

Anomaly Detection Algorithm. The architecture of the ADA [1] is illustrated in Fig. 20. ADA allows to distinguish between anomalies and anomaly free data by means of user-specified parameters, which feed a statistical filter. Such statistical filter analyzes one signal in turn, by means of one out of four alternatives, i.e., k-σ, k-MAD, hybrid σ-MAD and bi-weight methodologies [49].

Independently of the selected statistical filter, data acquired at time point t is analyzed by means of four user-specified parameters, i.e, one backward moving window (w_b), one forward moving window (w_f), and two thresholds, i.e., k_b and k_f. Backward and forward windows include data that precede and come after, respectively, time point t.

Anomaly characteristics are stored as well, namely distance from the acceptability threshold and detection score (DS), i.e., number of sensors that identified an anomaly at a given time point. This feature can be calculated only in case redundant sensors are available.

k-σ methodology. According to the k – σ methodology, raw data at time point t is labeled as anomalous if two conditions reported in Eq. (16) are verified. The exploitation of Eq. (16)

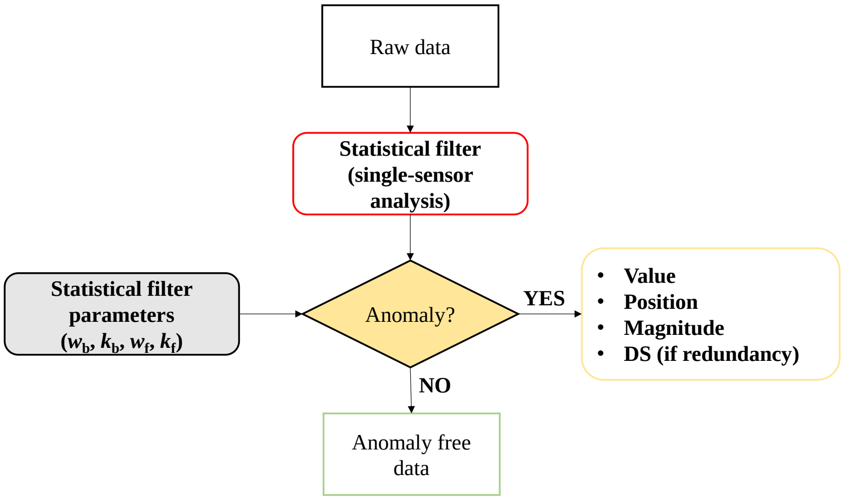


Figure 20 – ADA architecture

requires to set the length of backward and forward moving window, so that the mean value ($\bar{Y}$) and standard deviation (σ) of raw data included within both windows, i.e., $\bar{Y}_b$, $\bar{Y}_f$, σ_b and σ_f, are calculated.

$$\frac{|Y_t - \bar{Y}_b|}{\sigma_b} > k_b \text{ AND } \frac{|Y_t - \bar{Y}_f|}{\sigma_f} > k_f \tag{16}$$

As stated in Ceschini *et al.* [49], the k-σ methodology proved its effectiveness towards simulated dynamic time series in [48], achieving large percentages of detection combined with small percentages of false positive calls.

However, the employment of mean and standard deviation as estimators may result ineffective in case assumptions regarding the underlying data distributions are not completely verified [73]. Despite some specific cases, the set of random observations is assumed as Gaussian distributed. In this case, sample mean and standard deviation provide an efficient and unbiased estimate of location and scale with minimum variance [74]. However, as deviations from the Gaussian models occur in the dataset, the performance of such estimators severely decreases [73, 75]. Anomalous observations, i.e., outliers, represent a perfect example of deviations from the assumed underlying distribution and their negative effect on mean and SD is particularly severe, being the breakdown point for theses estimators equal to 0%. Consequently, the performance of the k-σ methodology can be lowered, because of the scarce robustness and null resistance of the estimators implemented in its parametric test.

k-MAD methodology. The Median Absolute Deviation (MAD) [75, 76] became popular for outlier identification [76, 77] mainly thanks to its breakdown point, i.e. 50 %. In spite of its exceptional robustness, the performance of the MAD in terms of efficiency is not so encouraging, achieving 37% for the Gaussian case [77]. Another drawback of the MAD is that MAD is equal to 0 if more than 50% of data are equal [76]. In such a case, observations are identified as anomalous despite their absolute difference from the median. This phenomenon, called implosion [73], can occur for rounded data coming from sensor readings. As the MAD replaces standard deviation in Eq. (16), the median replaces the mean. Therefore, if Eq. (17) is met, an anomaly is identified.

$$\frac{|Y_{\mathrm{t}} - med_{\mathrm{b}}|}{MAD_{\mathrm{b}}} > k_{\mathrm{b}} \text{ AND } \frac{|Y_{\mathrm{t}} - med_{\mathrm{f}}|}{MAD_{\mathrm{f}}} > k_{\mathrm{f}} \tag{17}$$

Hybrid k-MAD methodology. The tuning of the hybrid scheme relies on the fact that, according to the k-σ methodology, observations prior to the one under assessment have already been processed and consequently they can be considered reliable. Therefore, the probability that the backward window contains anomalous observations is rather low and the distribution is more likely to verify the Gaussian assumption. Under this condition, efficient estimators like mean and standard deviation become more desirable than robust and resistant ones in the backward window. Conversely, outliers can still occur among observations in the forward window, thus potentially affecting the quality of the estimate and consequently of the test criterion. In this case, robust estimators with consistent breakdown point, such as the median and the MAD, are required to prevent the reliability assessment from being excessively conditioned by outliers. The hybrid scheme consists of the application of the k-σ test criterion to the backward window and of the k-MAD test criterion to the forward window. A given anomalies is detected if Eq. (18) is met.

$$\frac{|Y_{\mathrm{t}} - \bar{Y}_{\mathrm{b}}|}{\sigma_{\mathrm{b}}} > k_{\mathrm{b}} \text{ AND } \frac{|x_{\mathrm{t}} - med_{\mathrm{f}}|}{MAD_{\mathrm{f}}} > k_{\mathrm{f}} \tag{18}$$

This resistance towards the presence of outliers is fundamental, but at the same time the stiffness of median and MAD is likely to cause a high number of false positive calls.

Bi-weight methodology. The bi-weight methodology is an intermediate solution between the k-σ methodology and the k-MAD methodology and it can be offered by the bi-weight mean and SD estimators, which present a more adaptive behavior than median and MAD, while keeping a 50% breakdown point. Lanzante *et al.* [73] demonstrated that the bi-weight estimate of dispersion is more efficient than pure MAD in a sample composed of thirty elements, which is very similar to the size of the window analysed by the authors in [48] for the back and forward moving window scheme. The estimate is performed by assigning different weights to observations according to their proximity to the center of the distribution, which is inversely proportional to their probability of being outliers. The weights are assigned according to a bi-weight function, gradually decreasing as measurements detach from the location estimate until they drop to 0 when a certain distance is reached. This threshold value, determined by the parameter a, influences the measure in which each weight u_t for each observation x_t is assigned and consequently the performance of the methodology, according to Eq. (19).

$$u_{\mathrm{t}} = \frac{Y_{\mathrm{t}} - med}{a \cdot MAD} \tag{19}$$

However, despite the interesting property of assigning less impact on measurements with high probability of being outliers, the tuning of the methodology may result complex. Directions for tuning are available in literature, but they refer to specific cases of application. Hoaglin *et al.* [78] suggest values of a between 6 and 9. Lanzante *et al.* adopt $c = 7.5$, while Kafadar [74] identifies 4-6 as a suitable range, concluding that the best performance is obtained with the value of 6.

In [49], on the basis of the information available in literature, the parameter a is assumed equal to 6. The acceptability criterion for the bi-weight methodology is expressed by replacing mean and standard deviation in Eq. (16) with their counterparts calculated by means of the previously described weighting procedure:

$$\frac{|Y_{\mathrm{t}} - \hat{Y}_{\mathrm{b}}|}{\hat{\sigma}_{\mathrm{b}}} > k_{\mathrm{b}} \text{ AND } \frac{|Y_{\mathrm{t}} - \hat{Y}_{\mathrm{f}}|}{\hat{\sigma}_{\mathrm{f}}} > k_{\mathrm{f}} \tag{20}$$

It can be seen that, if the value of the scale estimator in the test criteria in Eqs (16-18) and (20) is zero, an undetermined solution occurs. In order to prevent the denominator from being equal to zero in the test criterion, a non-biasing infinitesimal quantity can be added to standard deviation, MAD and bi-weight standard deviation, respectively.

Even if each of the mentioned methodologies can be selected in the Anomaly Detection Tool, the application of the k-σ is considered as the primary option in this thesis, as well as in Ceschini *et al*. [49], as this proved to be effective towards different scenarios which dealt with field data [79].

The parameters requested to the user at this stage are the same as the ones that characterize the statistical moving window methodology, i.e., number of observations in the backward w_b and forward w_f windows and their respective acceptability thresholds k_b and k_f.

Optimal values for these parameters were identified in [79] and are reported in Table 4 for each methodology. Thanks to the results obtained towards both simulated and field data, these values can be recommended for field application.

Table 4 – Statistical filter optimal tuning [79, 80]

Methodology	w_b	k_b	w_f	k_f
k – σ	50	3	25	2
k – MAD	50	3	25	3
Hybrid σ - MAD	50	3	25	3
Bi-weight	50	3	25	3

Noise Detection Tool. Noise Detection Tool identifies an intermediate region between reliability and anomaly boundaries, i.e., noise. To this aim, the statistical filters exploited by means of ADA, i.e., k-σ, k-MAD, hybrid σ-MAD or bi-weight methodologies, are used, but acceptability thresholds of the selected filter are properly adjusted. Starting from the optimal tuning values, a rule of thumb configuration, based on the results achieved in [79, 80] and on engineering practice, was conceived and applied in this thesis. The values of the noise removal filter parameters are reported in Table 5.

The main difference of the noise removal filter parameters with respect to the four methodologies listed in Table 4 is the value of k_b, which is equal to 2. This is coherent with the fact that the noise removal filter performs a refined analysis on data which have been already processed by three levels of judgment.

Table 5 – Noise removal filter parameters

w_b	k_b	w_f	k_f
50	2	25	2

In summary, SOHSA distinguishes between anomalies, noise and noise and anomaly free data observations according to the rules reported in Fig. 21. The procedure for both the anomaly and noise detection is depicted in Fig. 9. First, anomalies are detected by means of k_b and k_f. Then, anomaly free data are filtered out by means of the $k_{b,noise}$ and $k_{f,noise}$ thresholds, by distinguishing between noise and reliable data (i.e., noise and anomaly free data).

$\frac{\lvert Y_t - \bar{Y}_b \rvert}{\sigma_b} > k_b$ AND $\frac{\lvert Y_t - \bar{Y}_f \rvert}{\sigma_f} > k_f$	**Anomaly**
$\frac{\lvert Y_t - \bar{Y}_b \rvert}{\sigma_b} > k_{b_{noise}}$ AND $\frac{\lvert Y_t - \bar{Y}_f \rvert}{\sigma_f} > k_{f_{noise}}$	**Noise**
$\frac{\lvert Y_t - \bar{Y}_b \rvert}{\sigma_b} \leq k_{b_{noise}}$ OR $\frac{\lvert Y_t - \bar{Y}_f \rvert}{\sigma_f} \leq k_{f_{noise}}$	**Noise and anomaly free data**

Figure 21 – Detection rules of statistical filter and noise removal filter

3.1.7. Sensor Overall Health State Analysis on multiple and single sensors

Sensor Overall Health State Analysis has been exploited to evaluate the sensor overall health state of a large number of experimental datasets, so that advantages and drawbacks of SOHSA can be also identified.

Results reported within this Paragraph are divided into two different subparagraphs. This choice relies on the fact that, based on field data availability, different analyses are performed.

In Subparagraph "Multiple sensors", the anomaly and noise detection is carried out on dataset #1 through #5, which were used for setting the Fault Detection Tool parameters. Since these datasets account for multiple sensors, a comprehensive analysis dealing with fault, anomaly and noise detection and classification is reported.

Instead, analyses summarized within Subparagraph “Single sensors” are carried out on single sensors only, since no redundant/correlated sensors are available within the datasets. For this reason, only the detection of data anomalies and noise can be performed.

Multiple sensors. In this Subparagraph, the analyses of dataset #1 through #5 reported in Section 3.1.4 is integrated by detecting anomalous and noisy observations by means of the Anomaly Detection Tool and Noise Detection Tool.

Similarly to the results reported for the Fault Detection Tool, Figs. 32, 35, 38, 41 and 44 show where data anomalies and noise have been detected within the dataset, while Figs. 33, 36, 39, 42 and 45 summarize the anomaly rate and noise rate calculated for each sensor.

It has to be reminded that both Anomaly Detection Tool and Noise Detection Tool require one backward and one forward window to process field data. According to Tables 4 and 5, the backward window includes the first 50 data points of the dataset, while the forward window the last 25 measurements. As a result, 75 field data in total cannot be analysed.

In addition, the fault rate, obtained by means of the optimal tuning of the Fault Detection Tool, is also reported, so that the comprehensive health state information of each sensor can be provided.

Unlike the anomaly rate and noise rate, the fault rate is calculated as the ratio of the detected faults to the total number of field data acquired from each sensor. In fact, though SDA, SA and BA require a backward window, i.e., equal to 5, trend coherence faults can be detected since the beginning of the dataset. It has to be reminded that GTA, SSA and DTA are not considered.

Finally, Figs. 34, 37, 40, 43 and 46 deal with the anomaly classification.

As mentioned in Section 2, anomalies are gathered in four classes, namely Serial-Major (S-M), Serial-minor (S-m), Isolated-major (I-M) and Isolated-minor (I-m).

Dataset #1. Sensor Overall Health State Analysis for dataset #1 is shown in Figs. 32 through 34. As can be grasped from Fig. 32, the faulty sensor, i.e., sensor #3, is not detected by the Anomaly Detection Tool and Noise Detection Tool. In fact, stuck signal faults are properly detected and classified by means of the Fault Detection Tool. As a result, both the anomaly rate and the noise rate of sensor #3 are null (see Fig. 33). Conversely, fault rate clearly highlights fault occurrence, since 98 % of data are labelled as faulty.

However, it has to be highlighted the fact that the dataset was specifically selected for tuning the Fault Detection Tool. Thus, the fault rate referring to sensor #3 is not representative of the reliability of sensors installed in Siemens GTs, which instead are constantly monitored to guarantee a continuous and effective piece of information of unit operating conditions.

Instead, anomaly rate of sensor #1 and #2 is equal to 30 % and 37 %, respectively, while noise rate is roughly equal to 4 % for both sensors (Fig. 33). These rates may appear dramatically overrated with respect to anomalies and noise highlighted in Fig. 32. However, as previously mentioned, such anomaly and noise rates are calculated by considering only field data that can be properly analysed by means of the Anomaly Detection Tool and Noise Detection Tool, i.e., 38 % of the entire dataset.

As shown in Fig. 32, anomalies and noise are localized in the second transient.

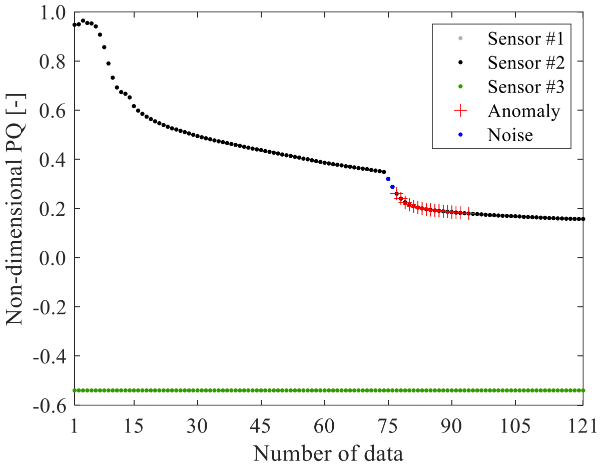


Figure 32 – Sensor Overall Health State Analysis in dataset #1

Instead, fault rate for both sensor #1 and #2 is equal to 0 %, meaning that no faults occurred. This result confirms that results sensor overall health state has to be provided by means of Fault Detection Tool and SOHSA as a whole.

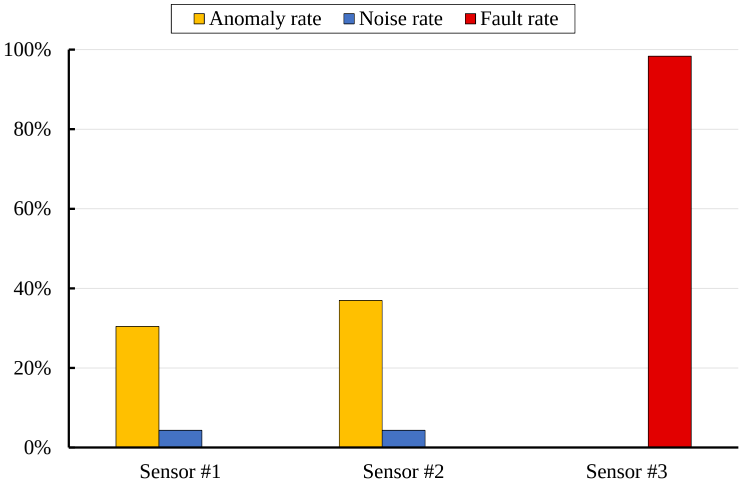


Figure 33 – Anomaly rate and noise rate in dataset #1

Anomaly classification is reported in Fig. 34. As can be seen, all anomalies detected in sensor # 1 are classified as S-M; in fact, subsequent time points are labelled as anomalous. As in sensor #1, majority of anomalies (94 %) are classified as S-M in sensor #2, while 6 % of anomalous data are labelled as I-M.

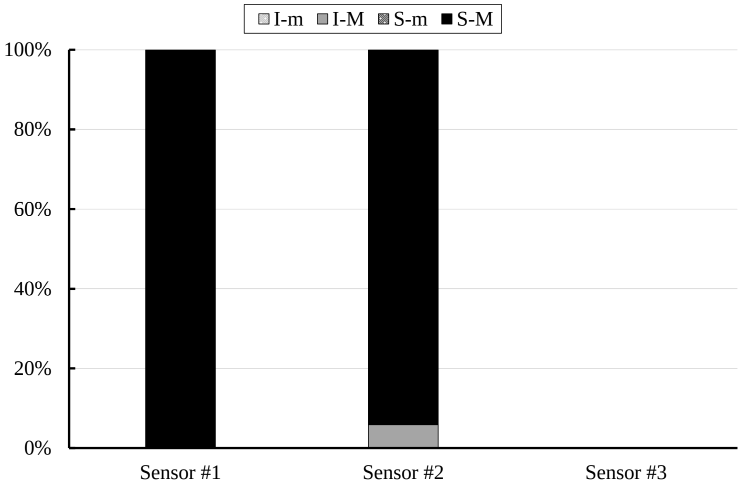


Figure 34 – Anomaly classification in dataset #1

Dataset #2. Anomaly and noise detection in dataset #2 is reported in Fig. 35 through 37. It has to be highlighted that dataset #2 represents a challenging scenario for SOHSA; in fact, sensor fault started occurring within the backward window. For this reason, the first anomalous observation is detected at time point 186 in sensor #1 (see Fig. 35). Instead, transient at time 400 is labelled as anomalous for both sensors.

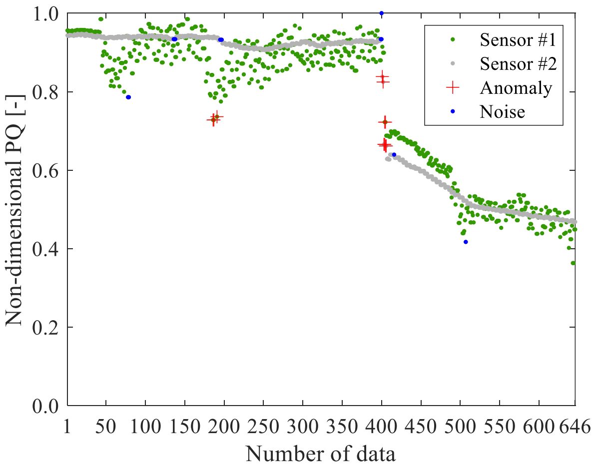


Figure 35 – Sensor Overall Health State Analysis in dataset #2

As a result, only a few anomalies are detected. In fact, anomaly rate is roughly equal to 1 % for both sensors, while maximum noise rate is slightly lower than 2 % (Fig. 36). Instead, the fault rate clearly highlights fault occurrence, since 31 % of data are labelled as faulty. As mentioned in Paragraph 3.1.2, both sensors are labelled as faulty.

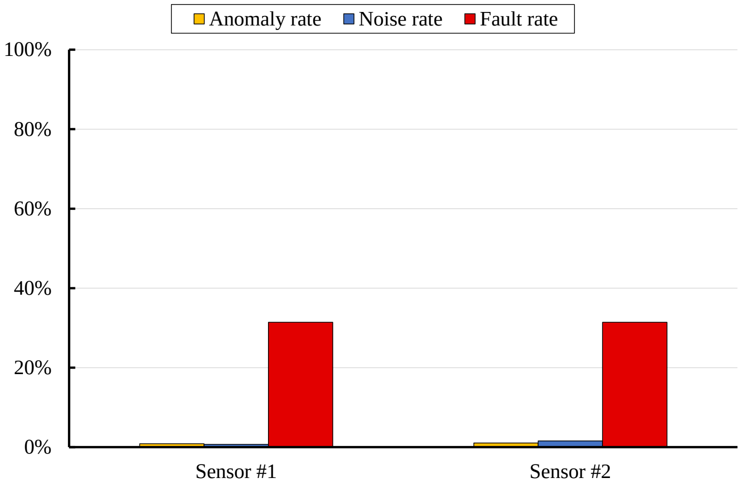


Figure 36 – Anomaly rate and noise rate in dataset #2

In spite of the reliability of sensor #2, all anomalies are classified as major (Fig. 37). However, all anomalies are detected when GT operating condition varies. Instead, only minor anomalies are detected by ACA in sensor #1. This result can be explained by the fact that sensor #1 exhibits an almost scattered behaviour throughout the considered time frame, so that transient operation is not labelled as rapid as in sensor #2. As a result, I-m and S-m rates are equal to 20 % and 80 %, respectively.

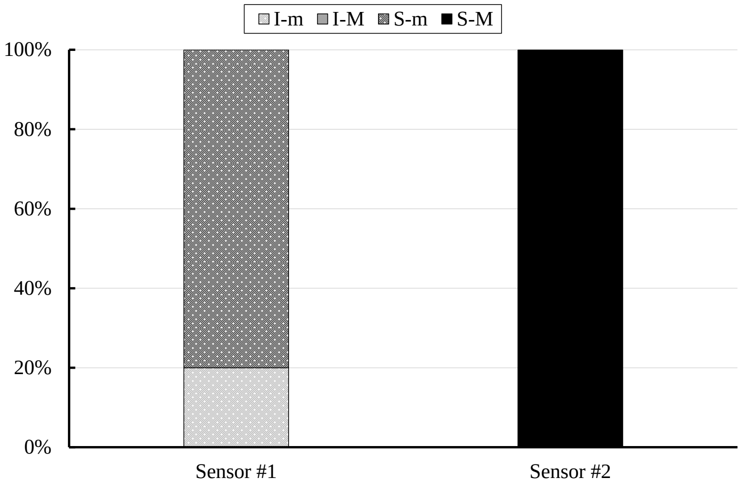


Figure 37 – Anomaly classification in dataset #2

Dataset #3. Figure 38 reports anomalies and noise detected by means of the Anomaly Detection Tool and Noise detection Tool. Similarly to dataset #1 and #2, anomalies are mainly detected throughout GT transients. Instead, only a few data are labelled as anomalies in steady-states conditions, on which noisy data are also identified.

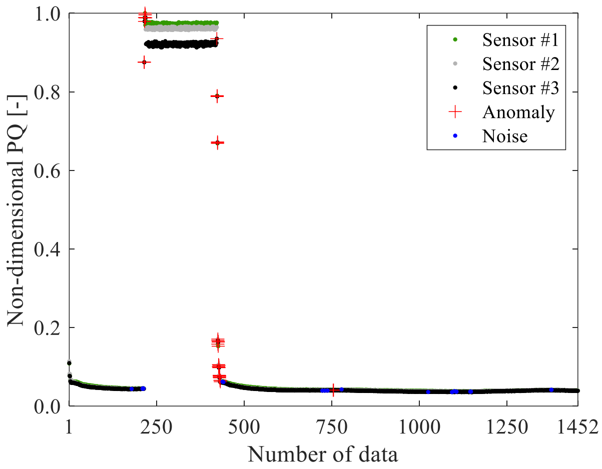


Figure 38 – Sensor Overall Health State Analysis in dataset #3

As a result, independently of the considered sensor, approximately 1 % of data are labelled as anomaly and noise (Fig. 39). Sensor health state is also confirmed by the Fault Detection Tool, which approximately detects 1 % of faults.

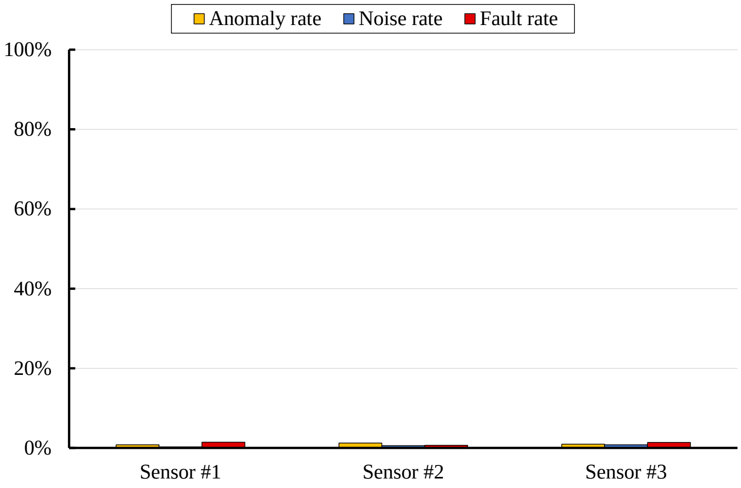


Figure 39 – Anomaly rate and noise rate in dataset #3

Since most anomalies are subsequently located within the dataset, most anomalies are classified as serial. As a result, 100 %, 82 % and 92 % of anomalies are labelled as S-M in sensor #1, #2 and #3, respectively. Some isolated anomalies are identified in sensor #2 and #3. More in detail, 6% and 12% of anomalies are I-M and S-m, respectively, in sensor #2, while 8% of anomalies are classified as I-M in sensor #3 (Fig. 40).

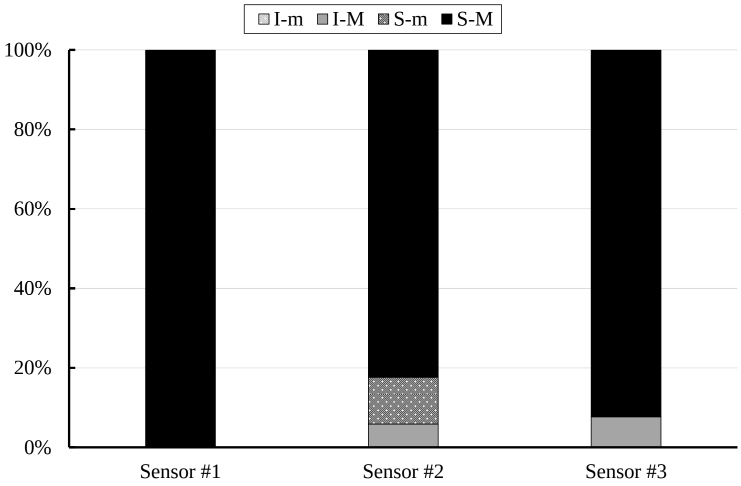


Figure 40 – Anomaly classification in dataset #3

Dataset #4. Similarly, to dataset #1, stuck signal fault occurring in sensor #2 is not detected (Fig. 41). However, bias fault from time point 548 to 551 is identified.

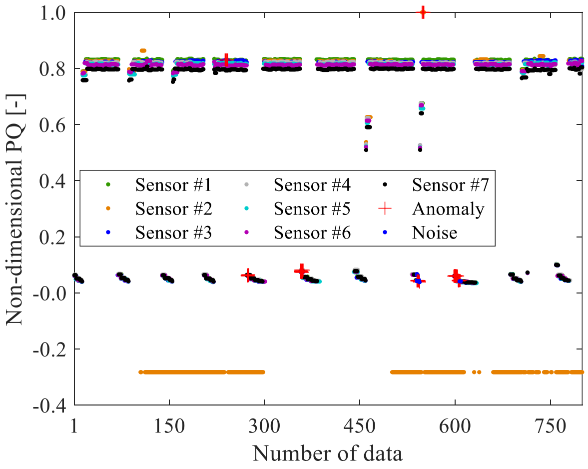


Figure 41 – Sensor Overall Health State Analysis in dataset #4

Regarding the other sensors, only a few field data are detected, especially throughout transients. As a result, both anomaly rate and noise rate are equal to 3 % at maximum (Fig. 42).

Similarly, fault rate is approximately equal to 1 %, with the exception of sensor #7, in which the fault rate is equal to 5 %.

Instead, Fault Detection Tool labels as faulty roughly 15 % of data acquired from the faulty sensor, i.e., sensor #2, by achieving the highest fault rate among all correlated sensors.

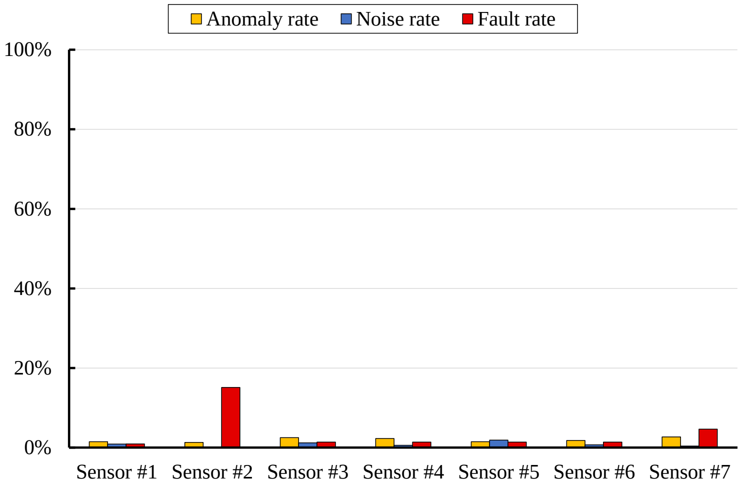


Figure 42 – Anomaly rate and noise rate in dataset #4

As depicted in Fig. 43, most anomalies are classified as S-M. In fact, the S-M rate generally varies in the range from 60 % (i.e., sensor 1) to 100 % (i.e., sensor #2 and #3).

Instead, the S-m rate ranges from 13 % (sensor #5) to 33 % (sensor #1), while no S-m anomalies are detected in sensor #2, #4 and #6. Instead, isolated anomalies are generally less frequent, since I-m rate and I-M rates are not higher than 6 % and 11 %, respectively.

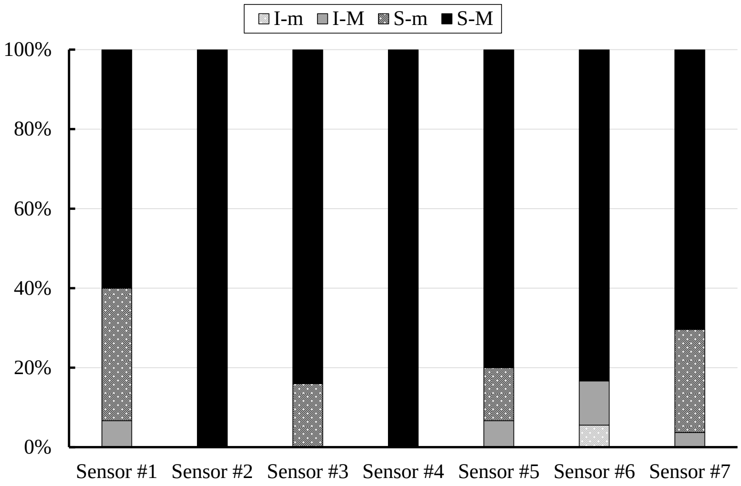


Figure 43 – Anomaly classification in dataset #4

Dataset #5. Figures 44 through 46 report results of Anomaly Detection Tool and Noise Detection Tool for dataset #5. As can be seen by means of Fig. 44, only a few anomalies and noisy data are detected, so that anomaly rate and noise rate is not higher that 3 % and 1 %, respectively (Fig. 45). This piece of information, examined with the fault rate equal to zero, allows to confirm the reliability of all sensors composing dataset #5.

Finally, Fig. 46 summarizes anomaly classification carried out by means of ACA. As can be seen, most of anomalies are classified as I-M. In fact, the I-M rate is in the range from 43 % (sensor #6) to 80 % (sensor #8). Minor anomaly rates, i.e., I-m and S-m, are lower than 42 % and 25 %, respectively. In addition, in some sensors, I-m and S-m rates assume approximately the same value.

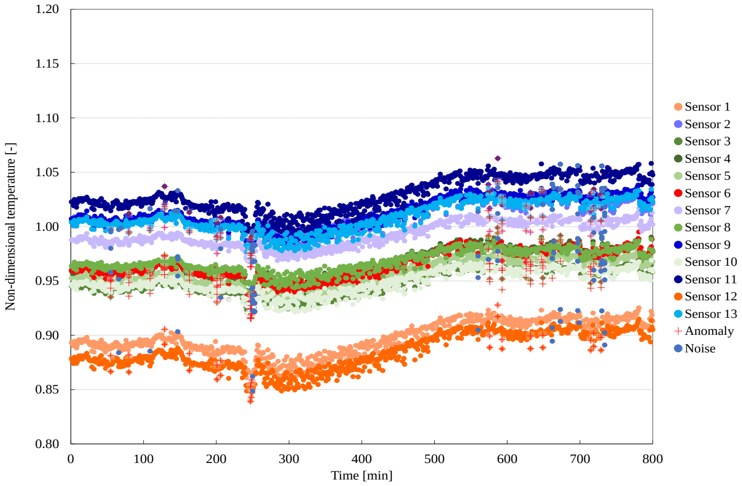


Figure 44 - Sensor Overall Health State Analysis in dataset #5

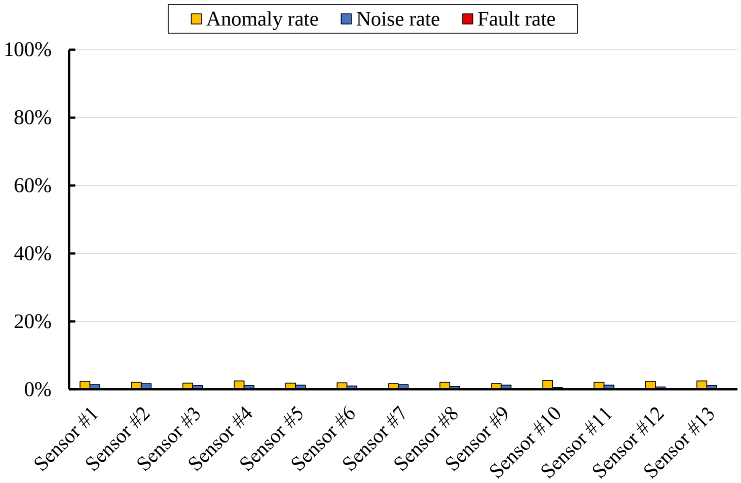


Figure 45 - Anomaly rate and noise rate in dataset #5

Finally, S-M anomalies are detected only in sensor #1, #4, #10 and #12, proving that a long-lasting fault does not occurred.

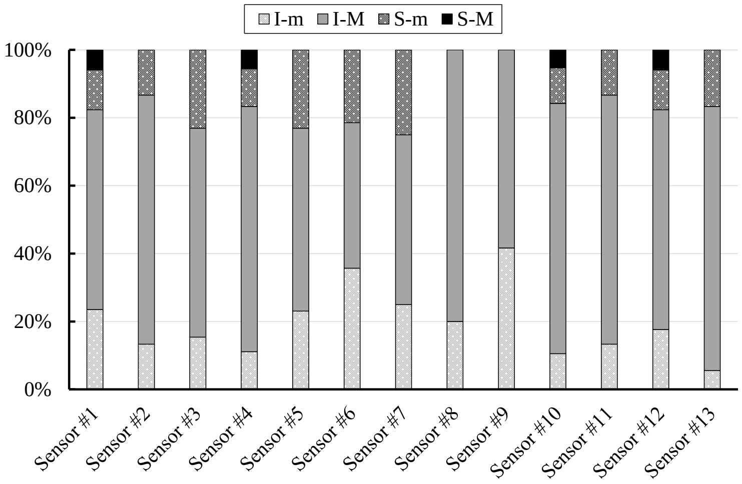


Figure 46 - Anomaly classification in dataset #5

Discussion. Sensor diagnostics provided by means of the Fault Detection Tool, Anomaly Detection Tool and Noise Detection Tool reveal that the I-DCIDS methodology can efficiently detect sensor fault occurrence.

As demonstrated, the analysis of the anomaly rate, noise rate and fault rate have to be jointly evaluated. For example, to avoid a redundant piece of information, stuck signal faults are detected by means of the Fault Detection Tool.

However, such a fault may not be detected by means of the Anomaly Detection Tool.

Thus, if a frozen signal is considered, its anomaly rate and noise rate may equal to zero, but the fault rate may reveal the fault occurrence, as demonstrated in dataset #1.

For this reason, the three indices, i.e., anomaly rate, noise rate and fault rate, have to be jointly taken into account to provide a comprehensive view on a given dataset or group of sensors.

As demonstrated, Anomaly Detection Tool and Fault Detection Tool generally flags as faulty and anomalous rapid transients, proving that the reliability of the developed methodologies can be maximized if quasi steady-state GT operation are considered.

In fact, its application to rapid system transients (i.e., start-up and shutdown) may increase the risk of false alarms, as shown in [25] and [26].

When redundant or correlated sensors are analyzed, faults post-processing may be a straightforward solution to avoid false alarms during transients.

In fact, if a measured value at a given time point is labeled as faulty in all sensors, the considered data can be flagged as fault free data, since the occurrence of one anomaly affecting all sensors at the same time point may be an unlikely event. Instead, GT operating conditions may be varied through the time.

The same post-processing operation may be a viable solution also for the Anomaly Detection Tool and Noise Detection Tool.

Single sensors. The performance of the SOHSA is further assessed in this thesis by means of a significant amount of field data taken on several Siemens gas turbines in operation. Thus, results reported in this thesis can provide a rule of thumb about data anomaly occurrence in field operation.

Field data. The datasets (DAs), registered by Siemens acquisition system during GT normal operation, are 15 in total (DA1 through DA15) and refer to 8 different gas turbines (GT1 through GT8).

The main characteristics of the gas turbine field data processed by means of SOHSA are described in Table 8.

The datasets include the recordings of 47 different sensors (SE1 through SE47) corresponding to 6 measurable physical quantities (PQ1 through PQ6, i.e., vibration, pressure, temperature, VGV position, lube oil tank level, rotational speed).

Most sensors measure vibrations (24) and pressures (14), while few of them measure temperatures (3), VGV position (4), lube oil tank level (1) and rotational speed (1).

As can be seen in Table 8, not all the sensors of a given PQ belong to a single DA. In fact, the datasets include a different combination and number of sensors (up to 43 at maximum).

The datasets cover a period of three months each, with the exception of two DAs, which instead cover a period of one month and four months.

Therefore, the total covered period is almost four years.

Since the sampling frequency is one minute, the total number of analyzed field data is approximately 70 million. Finally, it has to be highlighted that some datasets refer to the same GT (i.e. GT2, GT3, GT6 and GT7).

As an example of the trend of the available data of the 47 sensors, Figs. 47 through 52 show the trend over time of the six sensors (one for each considered PQ) which, according to the analyses performed and discussed in this Subparagraph, presented the highest anomaly rate, calculated as the ratio between the total number of anomalies and the sum of (i) anomalies, (ii) noise and (iii) noise and anomaly free data (hereafter named “reliable data”, for the sake of brevity).

Table 8 – Overview of analyzed field data

		V [# SE]	p [# SE]	T [# SE]	VGV [# SE]	h [# SE]	n [# SE]	Total [# SE]	# days	# data
GT1	DA1	1	0	0	0	0	0	1	91	131,040
GT2	DA2	18	14	3	3	1	1	40	90	5,051,588
	DA3	18	14	3	3	1	1	40	92	5,296,160
GT3	DA4	16	14	3	3	1	1	38	90	4,921,914
	DA5	18	14	3	3	1	1	40	92	5,296,160
	DA6	18	14	3	4	1	1	41	92	5,428,564
GT4	DA7	12	13	3	2	1	1	32	92	4,236,928
GT5	DA8	22	13	3	2	1	1	42	31	1,871,688
GT6	DA9	22	14	3	2	1	1	43	92	5,693,448
	DA10	22	14	3	2	1	1	43	90	5,569,532
	DA11	13	14	3	3	1	1	35	91	4,583,740
	DA12	22	14	3	2	1	1	43	92	5,693,372
GT7	DA13	14	13	3	2	1	1	34	91	4,452,776
	DA14	16	13	3	2	1	1	36	123	6,373,548
GT8	DA15	22	14	3	2	1	1	43	92	5,428,564

For the sake of confidentiality reasons, all the data in Figs. 47 through 52 were normalized according to Eq. (21), where the minimum and maximum value are specific to the considered dataset:

$$Y_{\text{norm}} = \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{21}$$

It should be noted that anomalies seem to be much more numerous in Figs. 47 through 52 than actual occurrence, which, in the worst case (pressure sensor) is lower than 5.3% (see Table 9). As for the multiple sensor analyses, each DA, 75 data points are not processed by the Anomaly Detection Tool and Noise Detection Tool.

Figure 47 shows that the normalized signal is in the range 0.3 – 0.6. SOHSA classifies as anomalies only a few data points at the boundaries of this range.

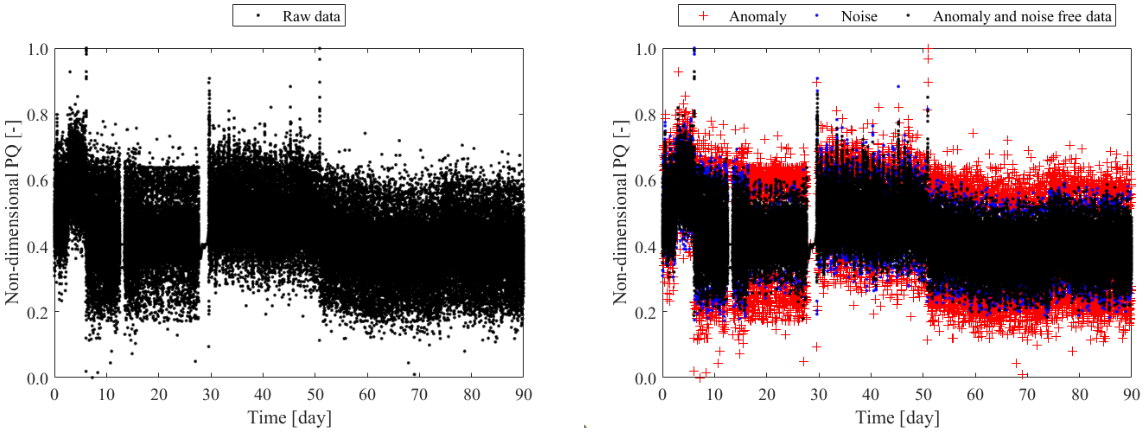


Figure 47 – Non-dimensional vibration data with highest AR (left: raw data; right: data processed by SOHSA)

In Fig. 48, the signal exhibits a particular behaviour; in fact most data belong to seven levels in the range 0.6 – 1.0.

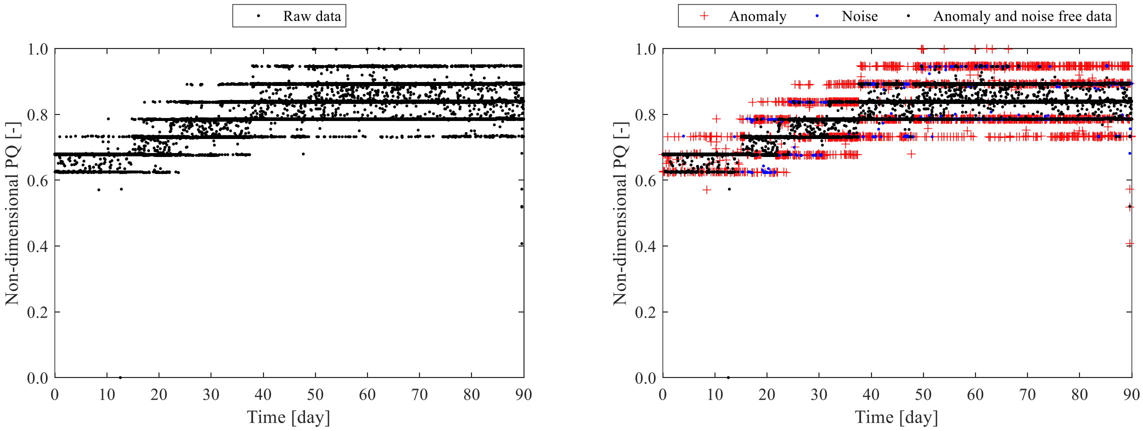


Figure 48 – Non-dimensional pressure data with highest AR (left: raw data; right: data processed by SOHSA)

The signal shown in Fig. 49 is in the range 0.8 – 1.0. As can be grasped from the figure, the signal exhibits a periodic behavior over time and only one severe transient occurs slightly before the 25th day, when the GT is not running (in fact, the non-dimensional temperature is null). Thus, such a PQ behavior may refer to a base load GT, as discussed in Losi *et al*. [81].

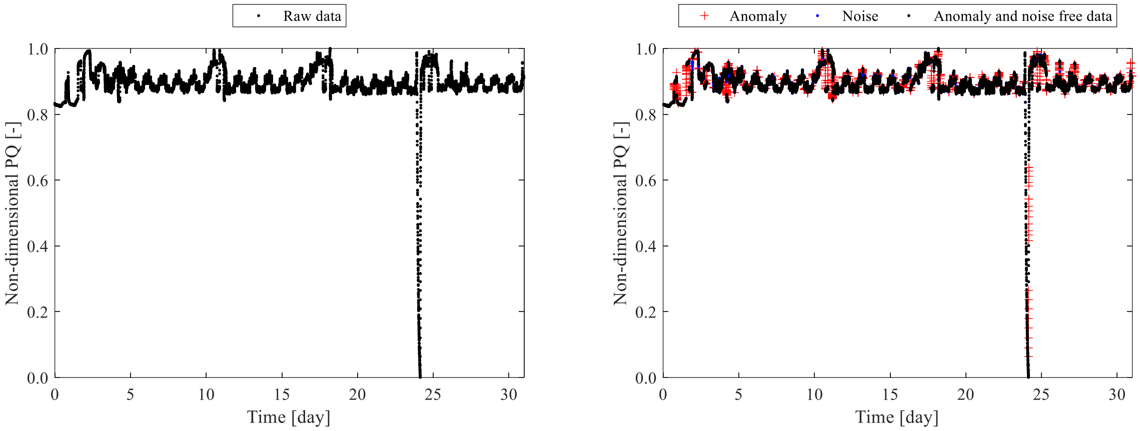


Figure 49 – Non-dimensional temperature data with highest AR (left: raw data; right: data processed by SOHSA)

The signal of VGV position (Fig. 50) seems to be constant over time, with the exception of some transients. Differently from what observed for other PQs, most anomalies do not take place during the transients.

In fact, even though the trend seems almost constant, the signal experiences some variations that are flagged as anomalies by SOHSA.

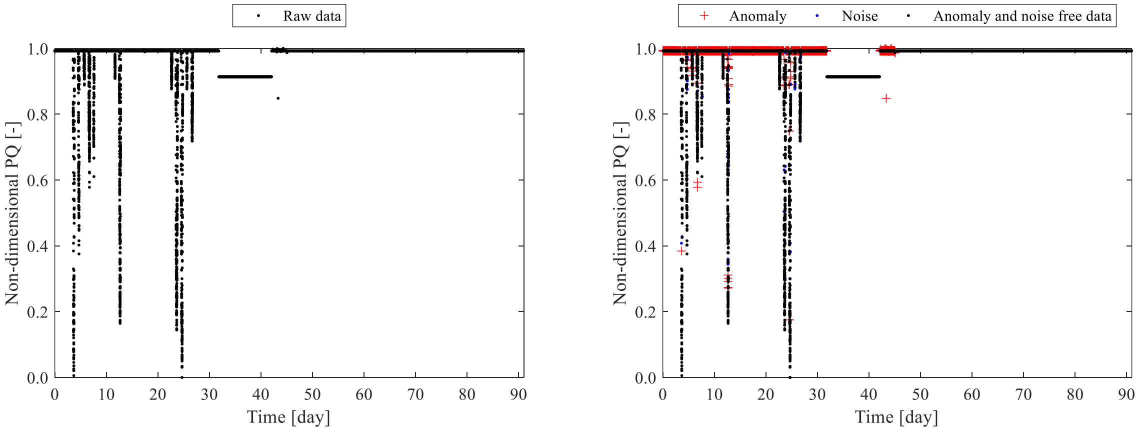


Figure 50 – Non-dimensional VGV position data with highest AR (left: raw data; right: data processed by SOHSA)

The signal of lube oil tank level (Fig. 51) varies in the range 0.9 - 1 and exhibits a somewhat periodic behaviour. Also in this case, most anomalies are identified during transients.

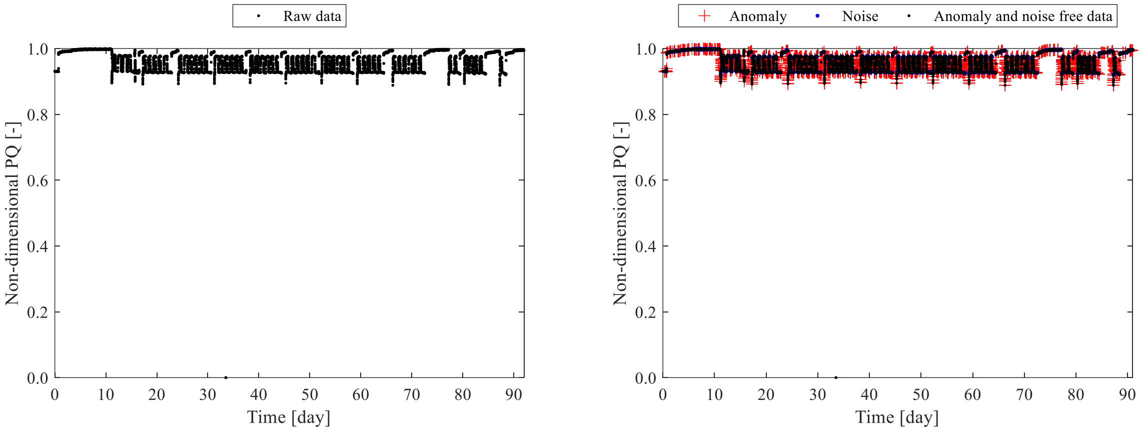


Figure 51 – Non-dimensional lube oil tank level data with highest AR (left: raw data; right: data processed by SOHSA)

The rotational speed (Fig. 52) varies in the range 0 - 1, with several step changes at 0.1, meaning that the GT is frequently shut down over time.

Anomalies take place during transients, in addition to the cases in which the GT is very close to the maximum non-dimensional specific speed and when it is not running (i.e. null rotational speed).

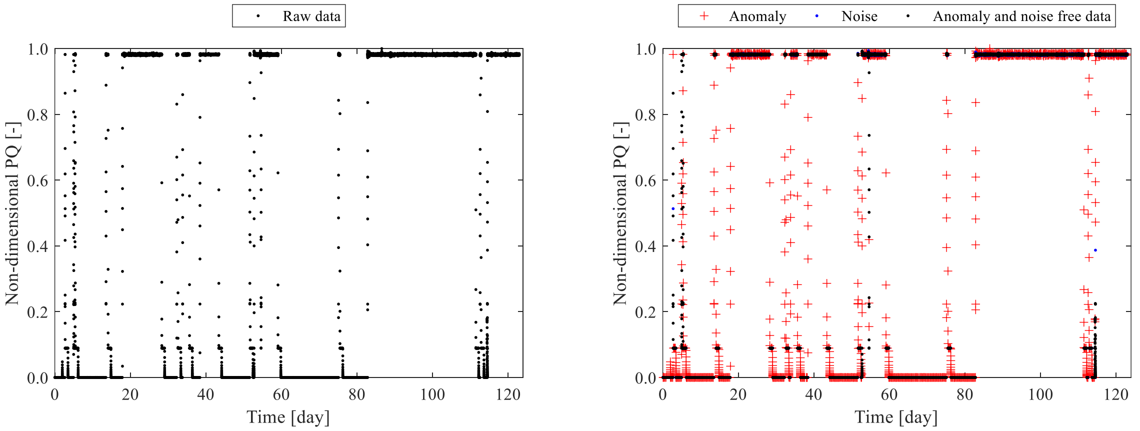


Figure 52 – Non-dimensional rotational speed data with highest AR (left: raw data; right: data processed by SOHSA)

Data anomaly detection. The analysis of field data by means of SOHSA allows the identification of three categories, i.e., reliable (i.e., noise and anomaly free data), noise and anomalies.

Figure 53 reports the rate of anomalies and noise of all the PQs for all the available DAs.

Figures 54 through 59 highlight anomaly detection output for every single PQ.

According to Fig. 53, the sum of anomalies and noise is lower than 3% for 10 out of 15 DAs. The DAs characterized by the highest anomaly rate are DA2 and DA3 (1.8%), while DA8 is the dataset affected by the highest noise (2.0%). Finally, as a combination of anomaly and noise rate, DA2 and DA3 are the datasets with the lowest rate of reliable data: in fact, anomalies and noise account for about 3.5% of all data. Conversely, DA1 has the highest rate of reliable data, since anomalies and noise affect less than about 0.05% of data. Finally, as a general comment, noise is very close to anomaly rate for all DAs.

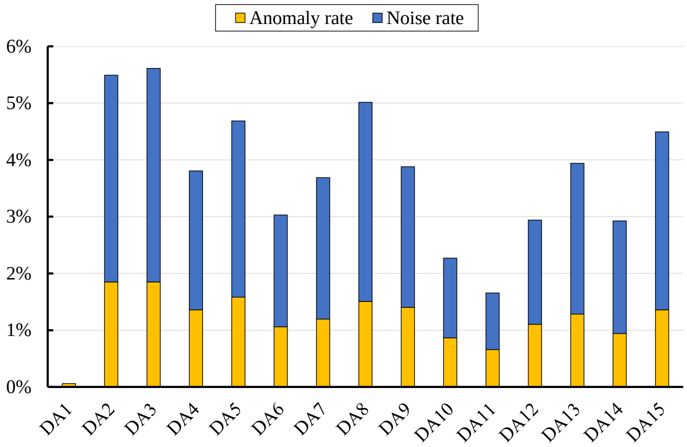


Figure 53 – ADA results on all sensors

The analysis of anomalies occurring in the different GTs shows that GT2 seems to be the most critical machine, followed by GT5, for which the sum of anomaly and noise rate is about 3.5%. The anomalies observed in GT3 range from 1.1% (DA6) to 1.6% (DA5). GT4 and GT8 have approximately the same anomaly rate (1.2%) and noise close to 1.5%. GT6 includes 4 DAs (DA9 through DA12): DA9 is characterized by the highest anomaly rate (close to 1.5%) and noise rate (close to 1.1%), while DA11 has the minimum noise (0.3%) and anomaly (0.6%) rates. Finally, GT7 (DA13 and DA14) shows a similar anomaly rate (close to 1.0%) and noise rate (close to 1.2%).

More in detail, the most relevant results about noise and anomaly rate affecting the six considered physical quantities are discussed in the following.

Vibration. The dataset DA2 presents the highest anomaly rate (approximately 2.0%). Conversely, DA1, as already observed in Fig. 53, shows the minimum sum of anomaly and noise rate (very close to zero), even though it should be remarked that it includes only one sensor of vibration. The sum anomaly and noise rates in DA4 (GT3) presents the highest difference compared to the average value in Fig. 53. All the datasets are characterized by a higher amount of noise than anomalies.

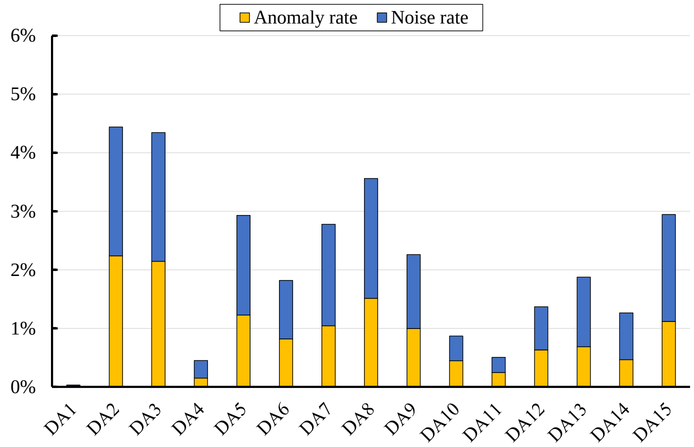


Figure 54 – ADA results on vibration sensors

Pressure. DA4 proves to be the dataset affected by the highest sum of anomaly and noise rate. For all datasets, the total amount of anomalies and noise is very close or lower than the average of all the type of sensors in Fig. 53. All datasets, with the exception of DA8 (GT5), present a higher rate of anomalous observations than the noisy ones.

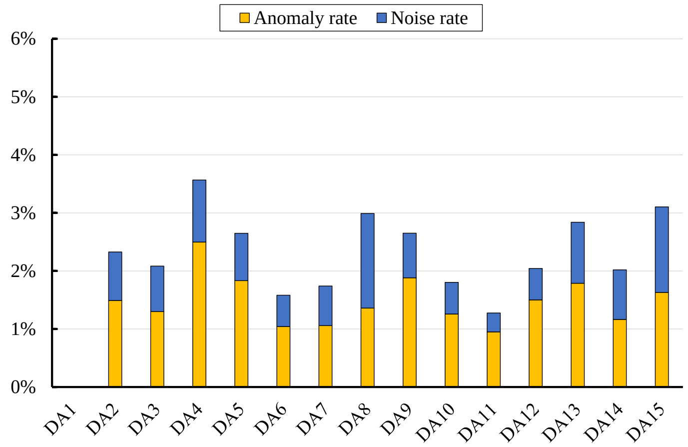


Figure 55 – ADA results on pressure sensors

Temperature. DA3 (GT2) is the most critical dataset. In fact, it includes the lowest rate of reliable data of all temperature sensors. It should be noted that most of these sensors are affected by a higher noise rate than the anomaly rate. GT6 is the gas turbine characterized by the highest value of reliable data. Generally, temperature sensors are affected by a lower rate of reliable data compared to the other sensors.

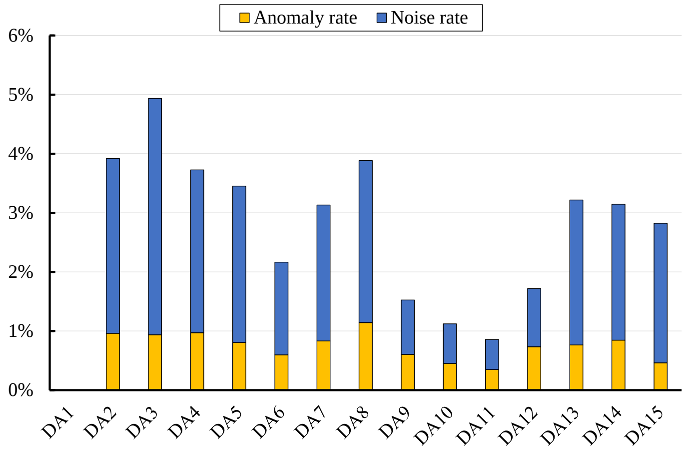


Figure 56 - ADA results on temperature sensors

VGV. It can be observed that the sum of anomaly and noise rate, with the exception of GT4 and GT6, is in the range 2% - 6%. DA3, followed by DA4, includes the most critical sensor of VGV position. In fact, they are the only two datasets where the sum of anomaly and noise rates is higher than 5%. Therefore, similarly to temperature, sensors of VGV position may also be critical, since, generally, they are affected by a greater number of anomalous and noisy observations than the average value (Fig. 53).

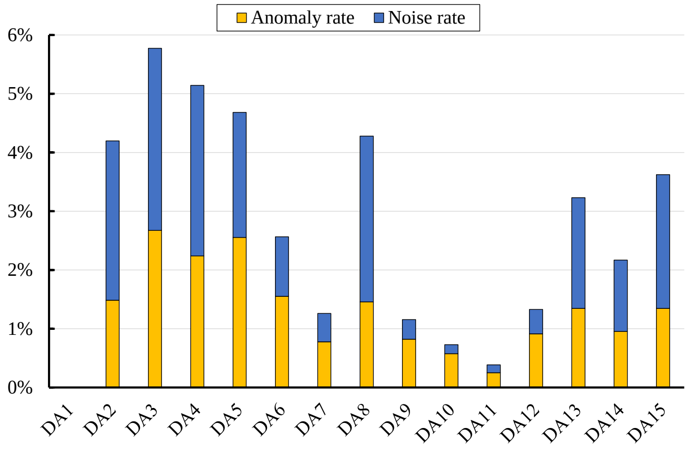


Figure 57 - ADA results on VGV position

Lube oil tank level. It should be recalled that only one sensor measures lube oil tank level (SE46). DA12 is the most critical dataset, since anomaly and noisy observations represent approximately 5 % of the entire dataset. GT6 (i.e. DA9 through DA12) seems to be the gas turbine with the most critical sensor, since the sum of anomaly rate and noise rate is higher than 4% for 3 out of 4 DAs.

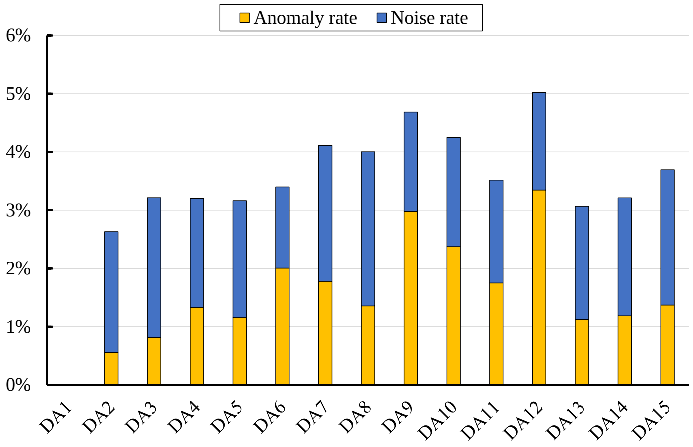


Figure 58 - ADA results on lube oil tank level

Rotational speed. Also in this case, only one sensor measures the rotational speed (SE47). It should be noted that DA11 (GT6) presents the lowest sum of anomaly rate and noise rate (4.6%).

In general, the datasets are affected by a larger amount of noise compared to anomalies; in addition, their sum is higher than the average values reported in Fig. 53.

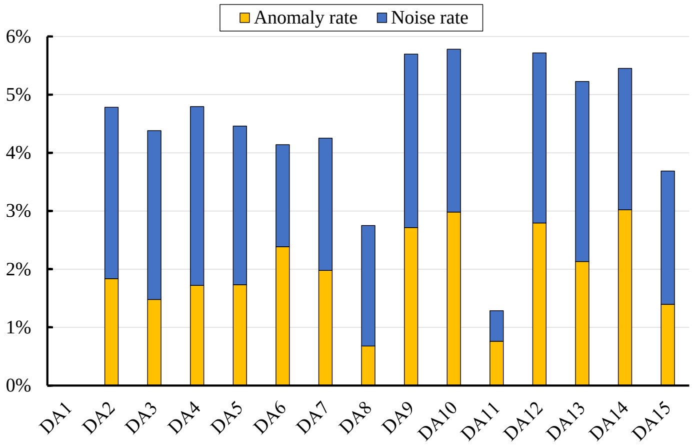


Figure 59 - ADA results on rotational speed

Discussion. Table 9 summarizes the characteristics of the six sensors with highest anomaly rate, previously reported in Figs. 47 through 52.

It should be noted that, on the basis of the analysis carried out on the available field data, GT3 proves to be the most critical machine, since the highest anomaly rate occurs in this GT in two out of six cases.

It should be noted that SE10 and SE34 are the most critical sensors; in fact, their anomaly rate is close to 5%, while this rate of the most critical sensors of temperature, VGV position, lube oil tank level and rotational speed is close to 3%.

More in general, rotational speed (13 out of 14 DAs), temperature and lube oil tank level (11 out of 14 DAs) sensors are the most critical, followed by VGV position (9 out of 14 DAs) and pressure (6 out of 14 DAs).

On the other side, the sum of anomaly and noise rates of vibration sensors is generally lower than the average.

It should be also highlighted that in some cases the sum of anomaly and noise rates in sensors of temperature, VGV, lube oil tank level and rotational speed is significantly higher than the average value, but rotational speed data are few compared to the total number of data. For this reason, the overall anomaly and noise rate is always lower than 4%.

Table 9 - Characteristics of the sensor datasets with highest anomaly rate

SE	PQ	DA	GT	Anomaly rate	Total data	# days
SE10	V	DA2	GT2	4.7%	129,600	90
SE34	p	DA4	GT3	5.3%	129,600	90
SE40	T	DA8	GT5	2.8%	44,640	31
SE45	VGV	DA6	GT3	3.1%	132,480	92
SE46	h	DA12	GT6	3.3%	132,480	92
SE47	n	DA14	GT7	3.0%	177,120	123

Data anomaly classification. As in Subparagraph “Multiple sensors”, anomalies detected by ADA are subsequently processed by ACA and can be classified as I-m, I-M, S-m and S-M.

Figure 60 shows the anomaly classification for all datasets, by considering all sensors and PQs. The most frequent anomaly classes are S-m and I-m. In GT1, most vibration anomalies are classified as S-m (52%), while S-M and I-M are at about 39% and 9%. No anomaly is classified as I-m.

Conversely, in DA2 through DA15, the anomalies labelled as I-m range from 16% to 55%, followed by S-m (in the range 17% - 40%), I-M (in the range 10% - 28%) and S-M (in the range 7% - 28%).

Major faults (S-M and I-M) are usually less frequent than minor faults.

Figures 61 through 66 show the results of anomaly classification by considering the different PQs and DAs. In particular, it should be remembered that DA1 only includes one sensor of vibration, temperature and VGV position are measured by three and four sensors respectively, while lube oil tank level and rotational speed are both measured by one single sensor.

The most relevant comments about the six considered PQs are reported in the following.

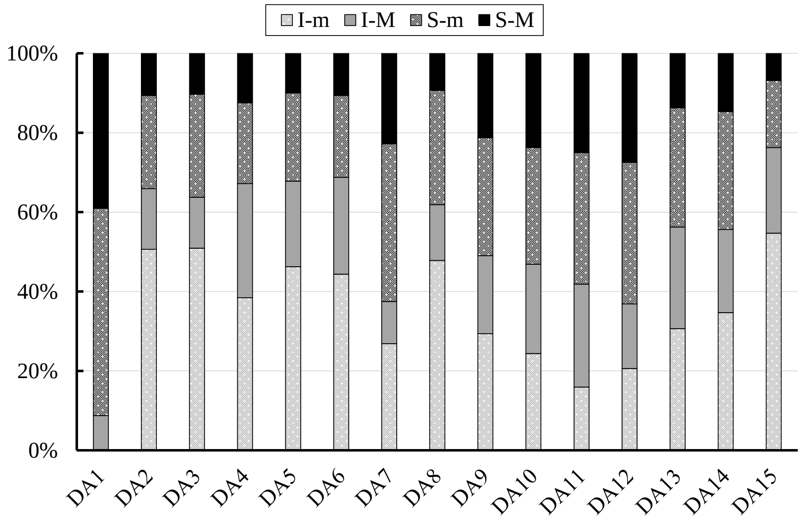


Figure 60 – Overall anomaly classification

Vibration. According to Fig. 61, most anomalies are classified as I-m (10 out of 15 cases); conversely, in DA1, DA10, DA11, DA13 and DA14 the largest class of anomalies is S-m. It should be noted that anomalies present a common behaviour.

In fact, S-M, S-m and I-m anomalies are generally lower, compared to the general results showed in Fig. 60, while I-m is slightly larger.

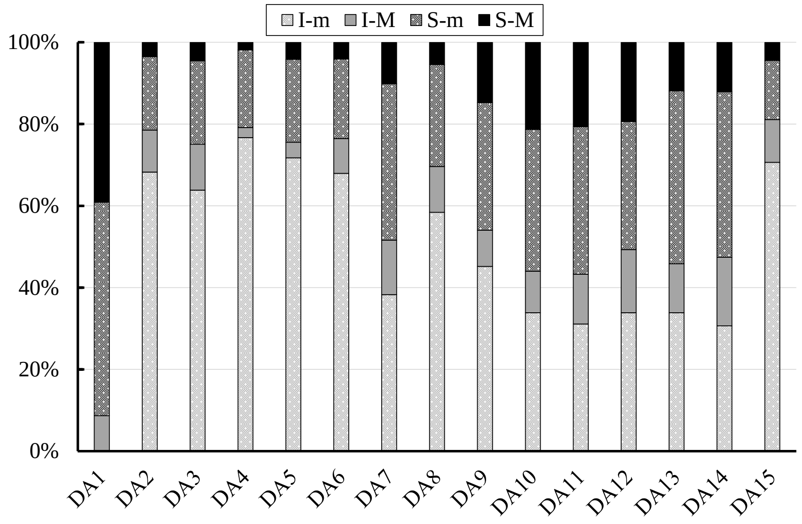


Figure 61 - Classification of anomalies in vibration sensors

Pressure. Differently from the overall results presented in Fig. 60, most anomalies (9 out of 14 cases) are classified as I-M (Fig. 62). On the contrary, most anomalies are S-M in DA2, the most frequent class is S-m in two cases (DA7 and DA12), while the most frequent class is I-m in DA8 and DA15. In general, pressure sensors are characterized by a lower occurrence of I-m then the average value shown in Fig. 60.

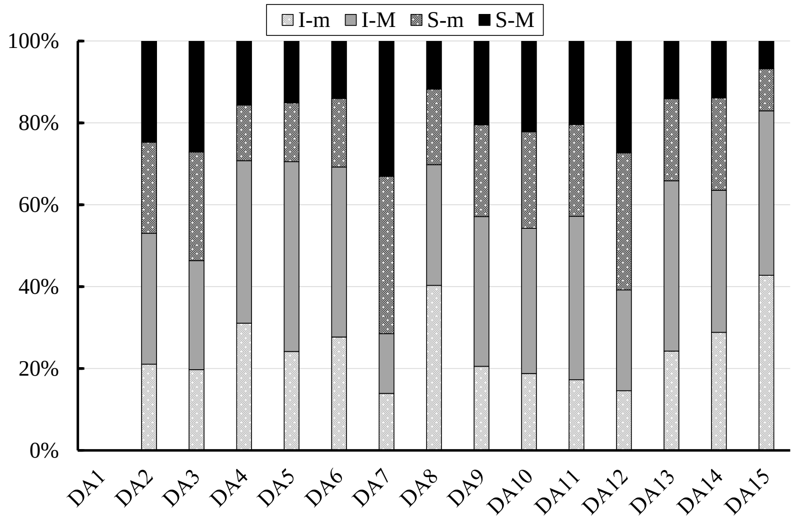


Figure 62 - Classification of anomalies in pressure sensors

Temperature. In case of temperature sensors, anomaly classification presents a particular scheme (Fig. 63). In fact, the most frequent anomaly class is always S-m. Moreover, in all DAs, the percentage of S-m is significantly higher than the average value shown in Fig. 60. As a consequence, I-M, S-M and I-m are significantly lower.

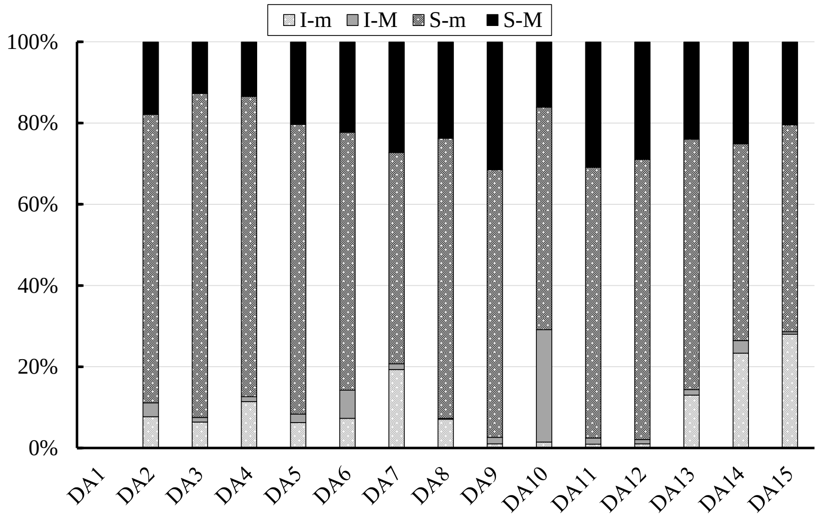


Figure 63 - Classification of anomalies in temperature sensors

VGV position. As shown in Fig. 64, the most frequent type of anomaly in DA2 through DA6, DA8 and in DA13 through DA15 is I-m. In DA7 and DA9 through DA12, the largest class is S-m.

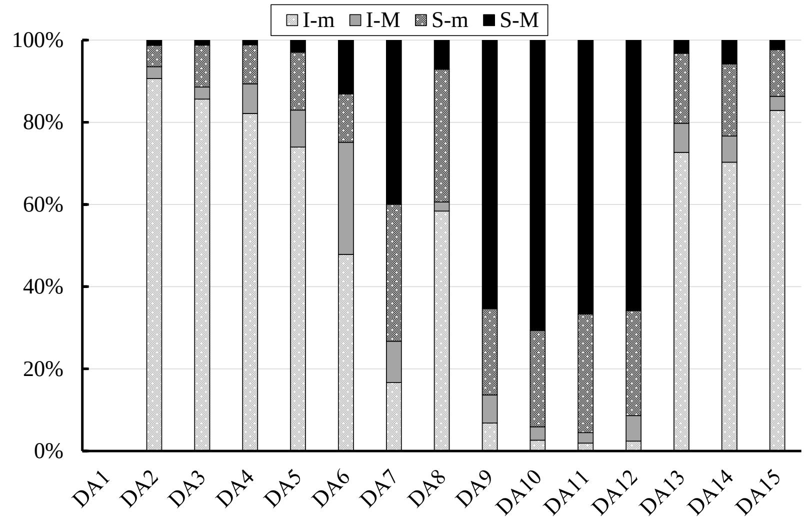


Figure 64 - Classification of anomalies in VGV position sensors

Lube oil tank level. Results shown in Fig. 65 are in agreement with the overall behaviour of Fig. 60, most anomalies are classified as I-m (7 out of 14 DAs) and their rate is higher than the average value shown in Fig. 60. GT6, GT7 and GT8 represent the only exceptions: in fact, most anomalies are classified as S-M (in the range 10% - 50%) and S-m (in the range 30% - 60%). Independent of the considered DAs, anomalies classified as I-M are generally in a lower percentage compared to the average value.

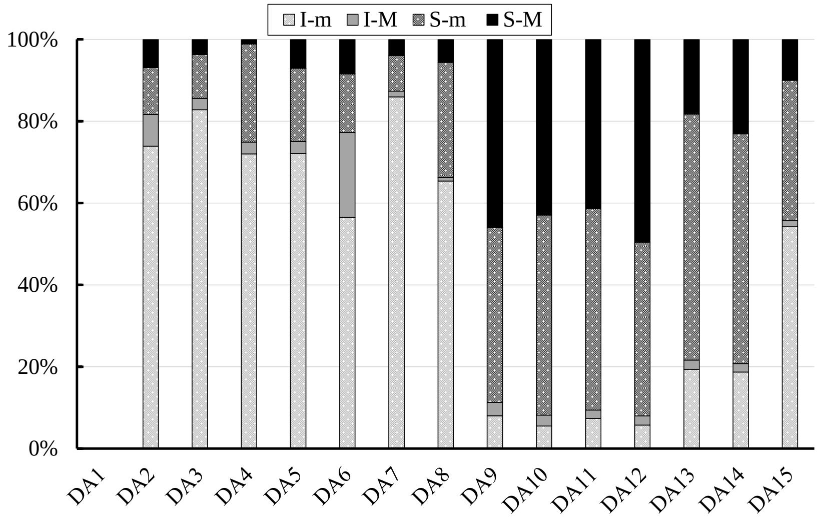


Figure 65 - Classification of anomalies in lube oil tank level sensors

Rotational speed. As reported in Fig. 66, most anomalies are classified as S-m (7 out of 14 cases) and as I-m (6 out of 14 cases). In DA6, most anomalies are I-M. With the exception of this dataset, anomalies classified as I-M are lower or even significantly lower than the average value.

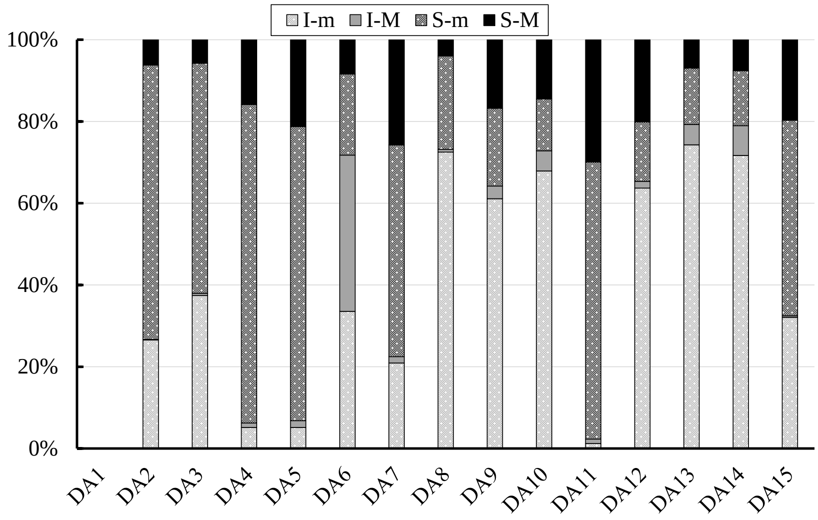


Figure 66 - Classification of anomalies in rotational speed sensors

3.2. BAYESIAN HIERARCHICAL MODELS

The Bayesian Hierarchical Model (BHM) employed in this thesis was derived by Zaidan *et al.* in [82] and Losi *et al.* in [83], where the methodology was exploited for prognostics purposes. In the current thesis, the BHM is adapted to multi-sensor diagnostics, so that data acquired from each individual unit, i.e., a given sensor, can be organized into a hierarchical structure.

Thanks to its multi-level structure, the BHM methodology is able to incorporate group effects, by performing a parallel analysis of redundant or correlated sensors. Such analysis is carried out by means of the simulation of a virtual sensor representative of healthy (i.e., reliable) sensor behavior. The methodology comprises four main steps that are shown in Fig. 67.

In step 1, model prior parameters are estimated by exploiting the information held by recorded observations taken in the past (i.e., past data).

In step 2, observations acquired in real-time (i.e., real-time data) from multiple sensors are used to provide the training dataset to estimate the individual likelihood functions.

Step 3 exploits the information held by the training dataset as a base of knowledge to perform inference on model parameters and, in particular, to provide an estimate of the pooled mean and covariance matrix that are used to generate the desired virtual sensor.

Finally, at step 4, a single-step-ahead prediction for the virtual sensor can be obtained by using the outputs of step 3.

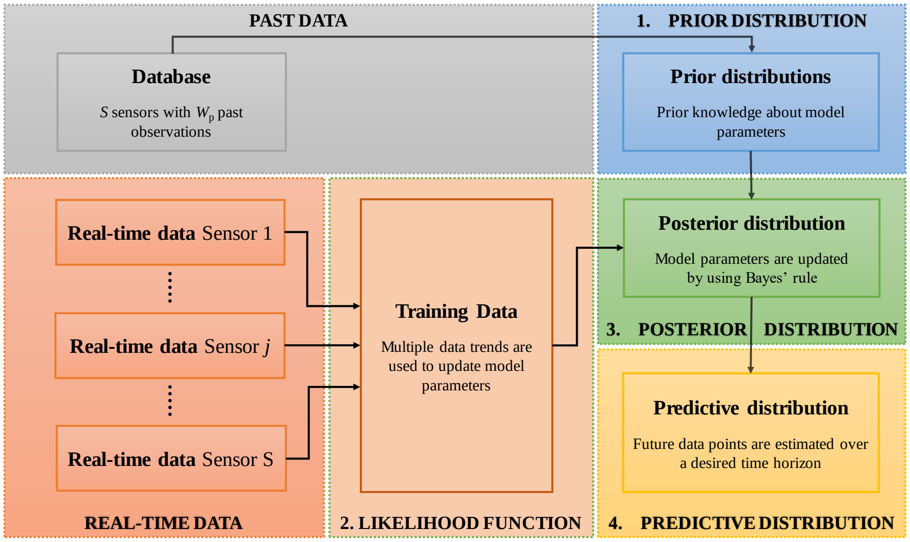


Figure 67 - Block diagram of the BHM

The BHM allows to estimate a future time point of the virtual sensor by means of a moving window of fixed length, i.e., in this case the moving window comprises 15 data, that includes data used for training the methodology. Then, when a new field data is acquired, the window moves forward and model parameters are updated.

In order to detect faults in sensor time series, each sensor is compared to the virtual sensor generated by the BHM and the absolute deviation is calculated at each time.

The fault detection criterion flags an observation as "faulty" if the deviation of a given sensor from the virtual sensor exceeds k times the predicted standard deviation across sensors.

A detailed insight about model equations and characteristics is reported in [35].

The quantitative assessment of the BHM-based methodology is provided by means of the following indices:

- True-positive rate (TPR), i.e., ratio between the true positives identified by the BHM and the total number of observations flagged as faults by the BHM, which is equal to the sum of true positives and false positives:

$$\mathrm{TPR} = \frac{N_{\mathrm{tp}}}{N_{\mathrm{tp}} + N_{\mathrm{fp}}} \tag{22}$$

- False-negative ratio (FNR), i.e., ratio between the number of faulty data incorrectly flagged as reliable and the number of true faults implanted in the time series:

$$\mathrm{FNR} = \frac{N_{\mathrm{fn}}}{N_{\text{true anomalies}}} = 1 - \frac{N_{\mathrm{tp}}}{N_{\text{true anomalies}}} \tag{23}$$

- False-positive ratio (FPR), i.e., ratio between the number of false positives and the number of true reliable observations:

$$\mathrm{FPR} = \frac{N_{\mathrm{fp}}}{N_{\text{true reliable}}} \tag{24}$$

3.2.1 Field data

The BHM methodology is tested on dataset #5 (see Fig. 44), also used for validating the I-DCIDS methodology.

The coefficient of variation, i.e., the ratio between standard deviation and mean, across the thirteen correlated sensors was calculated for each time point. The average value over the total time frame, which provides a measure of the variability of raw data across the pool, resulted equal to 4.5 %. This means that the thirteen time series are not completely homogeneous.

Fig. 68b shows both the raw data and the virtual sensor generated by BHM with lower and upper confidence boundaries for three different values of k, i.e., $k = 1$, $k = 2$ and $k = 3$. Given that σ_{vs} is the standard deviation of the virtual sensor, the boundaries corresponding to each value of k

are obtained by adding/subtracting the quantity $k*\sigma_{vs}$ to/from the virtual sensor. It has to be noticed that virtual sensor values start at time point 16, since the first 15 data represent past observations used to estimate prior distribution parameters.

For each sensor, the FPR is calculated on raw data for the three different values of *k*. When $k = 1$, i.e. very restrictive fault detection criterion, a high rate of false positives (100 % for sensor 1 and sensor 12 and higher than 50 % for sensors #2, #9 and #11) are obtained.

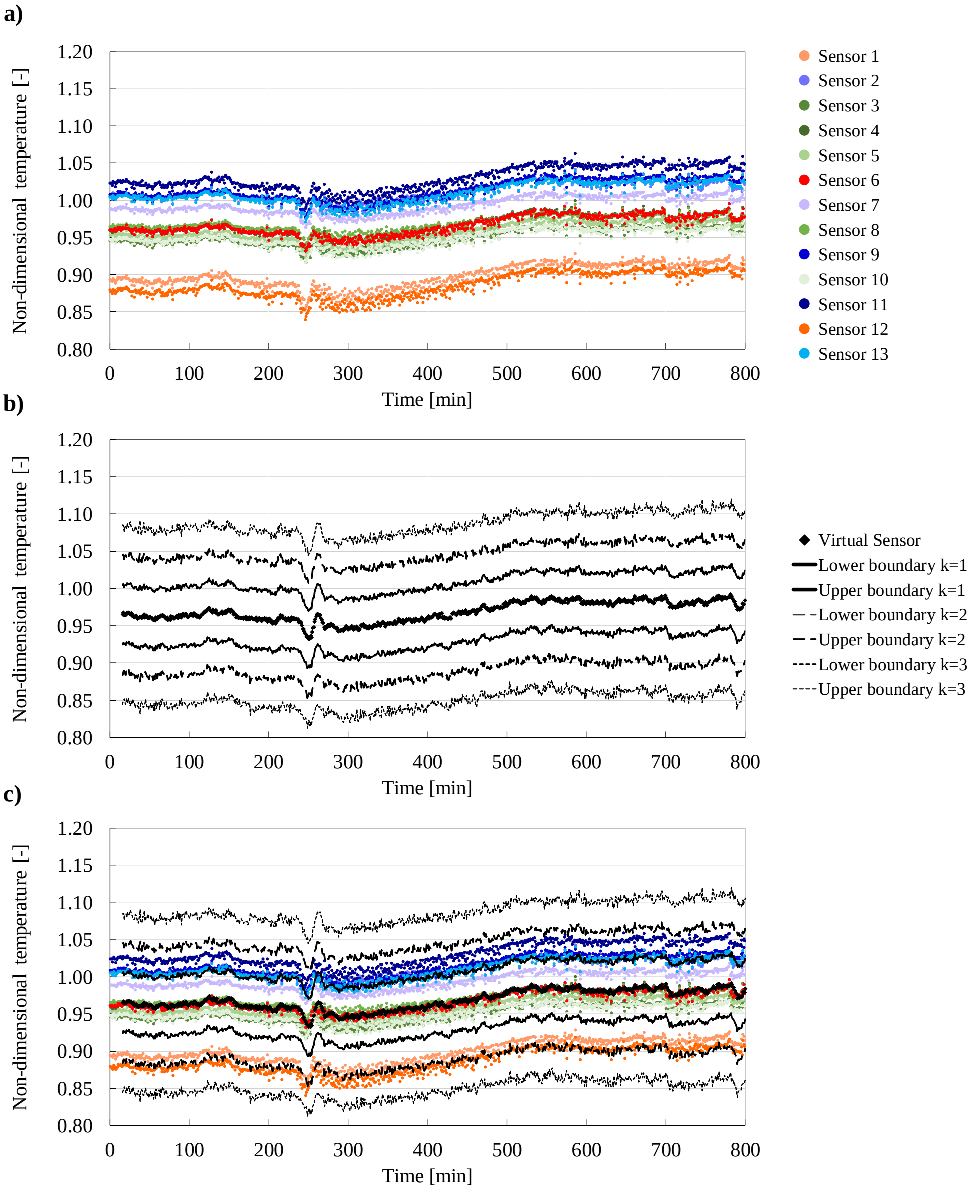


Figure 68 - Raw data (a), virtual sensor and boundaries (b) and synoptic view of raw data, virtual sensor and boundaries (c)

These results are related to the specific morphology of the considered time series and the fact that these time series include the highest and lowest measured values. In fact, when the *k* value is enhanced up to 3, FPR is null for all sensors.

3.2.2 Implanted faults

Two different types of faults are implanted into the dataset, i.e., spikes and bias. For both types of implanted faults, the BHM methodology is tested in different scenarios by varying the number of faulty sensors within the pool (1, 3 or 13).

Spike faults. Fifteen spikes are added at randomly selected time points to a single time series (*scenario 1*), three time series (*scenario 2*) or all time series simultaneously (*scenario 3*). With regard to *scenario 1*, three different cases are investigated since three different data levels can be identified:

- case (i): the time series with the lowest measured values as faulty and thus the sensor #12 is selected;
- case (ii): the faults are placed in sensor #6 (almost in the middle of the sensor trends);
- case (iii): the time series with highest measured values as faulty (sensor #11).

In *scenario 2*, the three faulty sensors are sensor #6, sensor #11 and sensor #12, i.e., the same three sensors individually investigated in *scenario 1*.

In *scenario 3*, all the available sensors are characterized by a fault at the same time points. Moreover, three different fault magnitude values are investigated, i.e. +5 %, +10 % and +30 %, with respect to the actual observations.

Bias faults. Three bias faults of different lengths, i.e., 10, 20 and 40 time points are considered and, also in this case, they are implanted into a single time series (*scenario 1*), three time series (*scenario 2*) or all time series simultaneously (*scenario 3*). Three different magnitudes are considered, which are the same as in the simulated spikes.

3.2.3 Results and discussion

For the sake of brevity, this thesis only reports results dealing with $k = 3$. Further comments dealing with the lowest values of the *k* threshold, i.e., $k = 1$ and $k = 2$, are documented in [35].

Spike faults. The results for scenario 1 are summarized in Fig. 69 for all the three sensors (sensors #6, #11 and #12). It should be noted that the 5 % spike magnitude is approximately equal to the variability of raw data (calculated equal to 4.5 %) across the complete pool of sensors; thus, it can be considered a challenging test for the methodology. The BHM flags all the implanted spikes as reliable observations for both sensor #6 and sensor #12.

The 10% spike magnitude leads to similar results as those obtained with the 5% magnitude, especially for sensor #12, while the spikes in the other two sensors are not detected.

By considering a 30 % fault magnitude, BHM is able to detect all the fifteen implanted faults for cases (i), (ii) and (iii). In fact, the FNR is zero.

This result does not rely on the predicted virtual sensor, which also accounts for the implanted faults, but depends on the variability of the raw data. In fact, the FPR is very high for sensor #11 and sensor #12, also when the virtual sensor predicts raw data.

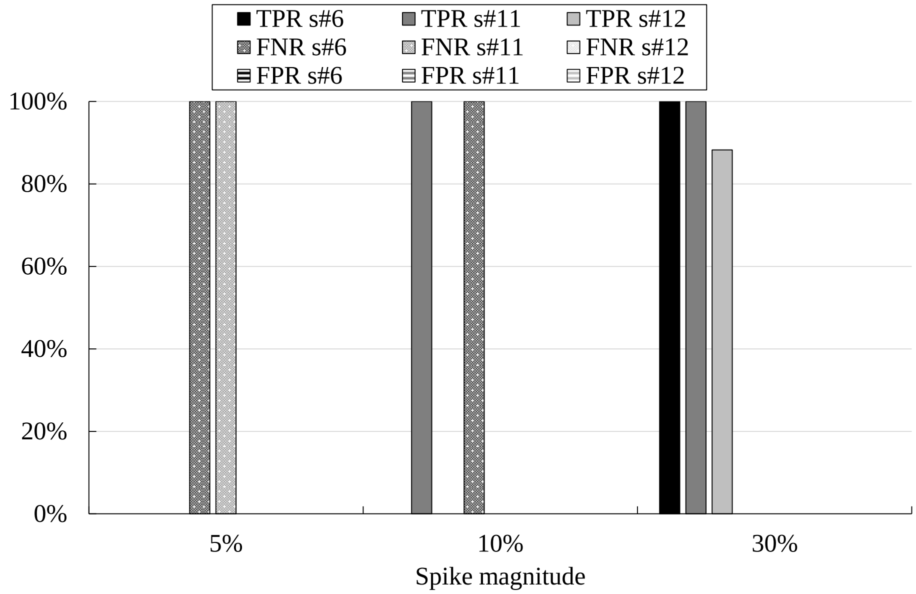


Figure 69 - Spike fault results of *scenario 1*

For the sake of brevity, the results for scenario 2 are not reported. In fact, it was found out that the simultaneous presence of spikes in sensors #6, #11 and #12 does not significantly influence the fault detection capability. This means that the virtual sensor predicted by the BHM remains robust also if multiple faulty sensors are included within the pool.

The results for scenario 3, which is clearly the most challenging scenario considered in this theis, are documented in Fig. 70, which shows TPR, FNR and FPR for 5 %, 10 % and 30 % spike magnitude, respectively.

It can be seen that BHM demonstrates good consistency in the prediction of the virtual sensor. In fact, the maximum deviation between the virtual sensor trend obtained by using raw data and its trend obtained by implanting all the faults of scenario 3 is approximately 3 %, 6 % and 11 % for 5 %, 10 % and 30 % spike magnitude, respectively. Therefore, it is always lower than the corresponding implanted spike magnitude.

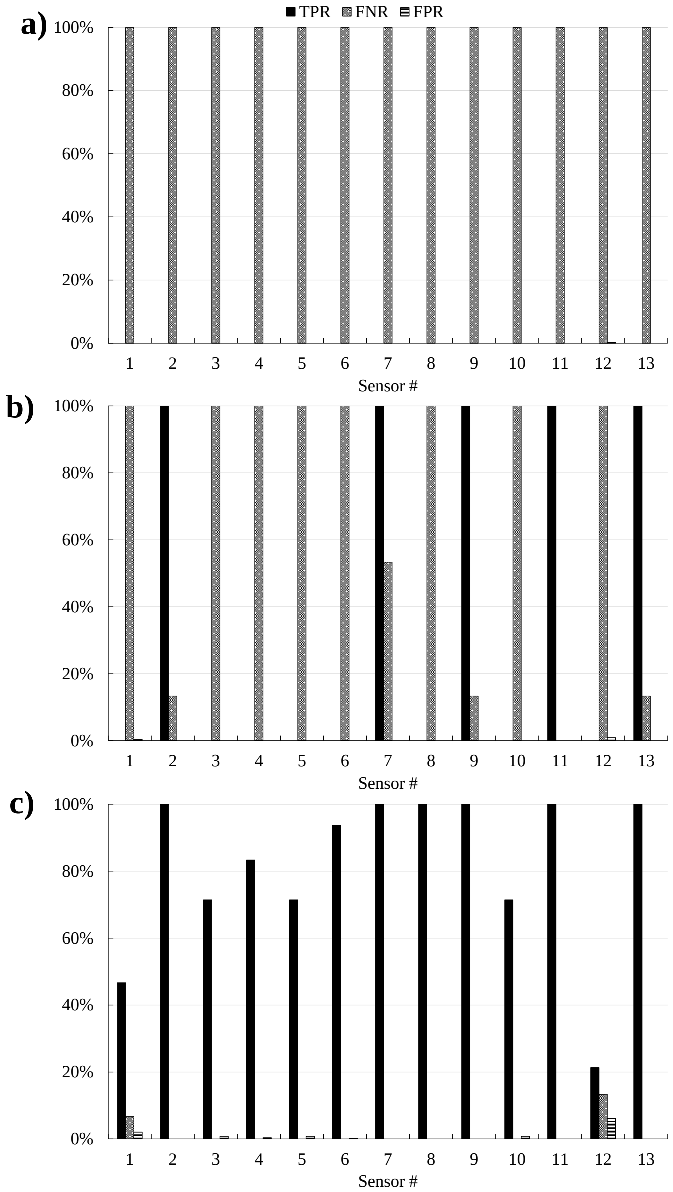


Figure 70 - Spike fault results of *scenario 3* - a) 5 % spike magnitude; b) 10 % spike magnitude; c) 30 % spike magnitude

The comments made for scenario 1 and scenario 2 also stand for scenario 3. In fact, if *k* = 5 % FNR is equal to 100 %. The results obtained with 10 % spike magnitude can be commented in a similar fashion.

Otherwise, when the implanted faults are significantly higher than the standard deviation across the pool of sensors, the performance of BHM are more uniform among both the different sensors. In fact, very few spikes are flagged as reliable for sensor #1 and sensor #12 (which are characterized by the lowest measured values), as demonstrated in Fig. 70. The average TPR is approximately equal to 80 %.

Bias fault. The results for scenario 1 are summarized in Fig. 71 which reports the TPR, the FNR and FPR values for the three sensors (#6, #11 and #12) with implanted bias faults. As made above, the TPR, FNR and FPR values have to be analyzed in a comprehensive view. Sensor #6 is characterized by satisfactory results only if fault magnitude is at least equal to 30 %; otherwise, high FNR values are obtained. Instead, with regard to sensor #11, the number of true positives detected by the virtual sensor is less influenced by both the tuning of the fault detection criterion and fault magnitude. Finally, faults implanted in sensor #12 (of which the time series includes the lowest measured values) are successfully detected only if their magnitude is very high and the tuning of the fault detection criterion avoids false positives.
The FPR values are null independently of bias magnitude.

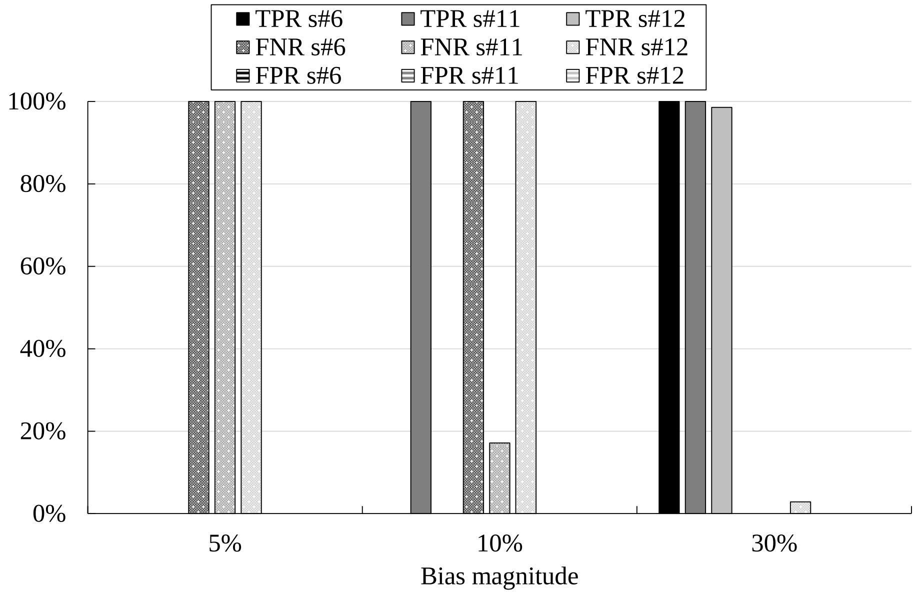


Figure 71 - Bias fault results of *scenario 1*

Moreover, it should be noted that there are not substantial differences from the corresponding scenario 1 of spike faults. This means that, if a single faulty sensor is included within the pool, the virtual sensor predicted by BHM is still characterized by consistency and robustness, i.e., it is independent of the number of faulty observations included within the moving window at time *t* used to estimate the parameters for virtual sensor generation at time *t*+1.

In fact, while in the case of spike faults the maximum number of faulty observations within the moving window is equal to 5, two implanted bias faults have a length that is higher than the dimension of the moving window; thus, all the data within the window are faulty and this occurs for several consecutive time points.

As for spike faults, the results of scenario 2 are not shown in this thesis for the sake of brevity. However, with regard to bias faults, it is worth highlighting that the FNR exceeds 40 % for sensor #6 and for sensor #12 at fault magnitude equal to 30 %.

This means that the simultaneous presence of three faulty sensors and high fault magnitude implies a more challenging fault detection by means of BHM virtual sensor. This effect was not highlighted in the results about spike faults in scenario 2. An increase of the number of faulty sensors within the pool further highlights this different behavior between spikes and bias faults, especially for high values of fault magnitude.

Figure 72 shows the results obtained for scenario 3 where all the sensors of the pool are affected by bias faults. With regard to 30 % fault magnitude (see Fig. 72c), it can be noted that a significant decrease of the average TPR value across sensors is obtained, which is approximately 30 % for k = 3. In fact, the TPR values exceed 60 % for only one sensor, i.e. sensor #11, while in scenario 3 of spike faults this condition was verified for most of the sensors k = 3. Moreover, the FNR is extremely high, while in the case of spike faults the FNR values are equal to zero for all sensors. This is explained by the fact that, in scenario 3 with bias faults, the maximum deviation between the virtual sensor trend obtained by using raw data and its trend obtained by implanting the faults in all the sensors of the pool is comparable to bias magnitude.

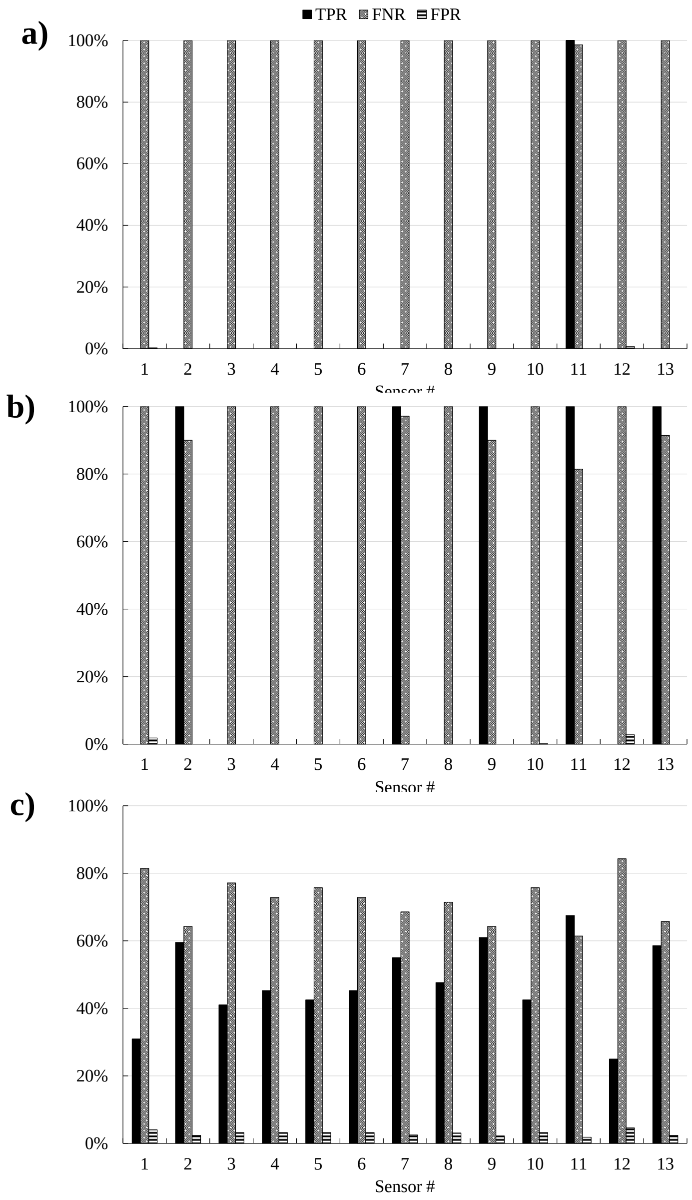


Figure 72 - Bias fault results of scenario 3 - a) 5 % bias fault magnitude; b) 10 % bias fault magnitude; c) 30 % bias fault

4. DETECTION OF UNIT OF MEASURE INCONSISTENCY

A further challenge related to acquisition and handling reliable measurements is provided by the occurrence of UMIs.

UMIs are usually detected by means of visual inspection or by analyzing the signal values with respect to minimum and maximum thresholds in which the signal usually ranges, i.e., acceptability thresholds. However, this procedure may lack of robustness and generality by varying UOM type and scale. In fact, in some cases, the measured value affected by UMI may lie within the acceptability thresholds tuned in a different UOM; as a result, UMI may not be detected. In addition, a further challenge linked to the acceptability thresholds relies on the assessment of the most reliable and suitable range of the physical quantity, which may require an in-depth and prolonged analysis. Thus, to avoid false positives and false negatives, the influence of the acceptability thresholds for UMI detection has to be reduced. Finally, from an industrial perspective, UMI should be automatically detectable, preferably as soon as possible.

It is worth noting that UMI detection is strictly correlated to sensor diagnosis. In fact, because of UMI occurrence, a sensor may be incorrectly labeled as faulty because of erroneous UOM association. Thus, an effective UMI detection could also enhance the detection of sensor faults.

According to the comments mentioned above, this thesis is also addressed to the detection of UMI issues, whose characteristics allow to differentiate them from sensor faults (see Section 2.4).

Thus, this thesis investigates the capability of ML learning classifiers, so that the above-mentioned requirements may be met. In fact, ML is a data-driven approach that handles data in order to generate a model, which is subsequently employed to analyze and classify further measurements. ML classifiers usually have general validity, so that they can be exploited for analyzing different physical quantities. Secondly, ML classifiers may process a large amount of data in a limited timeframe. This second requirement is crucial towards the big data analysis.

This thesis investigates only supervised ML classifiers since they classify unlabeled data.

In the literature, the effectiveness of supervised ML approaches is documented in different research fields, also for GT applications.

Supervised ML classifiers are usually divided into two classes, i.e., parametric and non-parametric classifiers. The former classify data according to the statistical probability distribution of each class [84], whereas the latter are exploited when the probability density function of data is unknown [84].

Naïve Bayes, Decision Tree and SVM are the most popular parametric classifiers. Instead, K-Nearest Neighbors, Logistic Regression and Artificial Neural Network belong to the non-parametric class.

In addition, two classification strategies may be encountered: eager learners (e.g., Naïve Bayes) use data for training for constructing a generalization model; such a model is used to classify data for testing. Instead, lazy learners (e.g., K-NN) do not construct a model by means of the labeled data, which instead are only stored [85].

In line with the literature survey reported in Section 1, three supervised ML classifiers, namely Support Vector Machine (SVM), Naïve Bayes and K-Nearest Neighbors (K-NN) are investigated. In fact, they are the most popular supervised ML classifiers, which proved their effectiveness in several research fields.

4.1 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised non-parametric algorithm that was originally developed for pattern ML models. SVM is an eager learner; thus, data whose label is unknown, i.e., unlabeled data, are classified as follows.

Initially, the SVM model is trained by means of labeled data, i.e., their class is known. Then SVM model effectiveness is tested by classifying unlabeled data.

Thanks to its classification performance, SVM is popular in several research fields, as well as in engineering [86, 87], informatics [88] and remote sensing [89].

In fact, according to a literature survey, SVM generally guarantees higher classification accuracy than other methodologies, e.g., neural networks, even in case of a low number of training data. In addition, overfitting problems are usually avoided [89].

SVM is generally employed for binary classification problems, i.e., data referring to two classes only. As sketched in Fig. 73, SVM is based on the principle of structural risk minimization, by generating two parallel hyperplanes (i.e., margin width) that maximize the distance between the two classes; the SVM hyperplane lies halfway the margin width. Data for training that lie on the margin width are the so-called support vectors.

As highlighted in [57, 59], SVM classification can be carried out by means of both linear and non-linear kernel functions. Radial Basis Function (RBF), sigmoidal and polynomial functions are the most popular non-linear SVM approaches. The comparison of non-linear kernel functions is the topic of some studies aimed at identifying the most efficient approach for data classification.

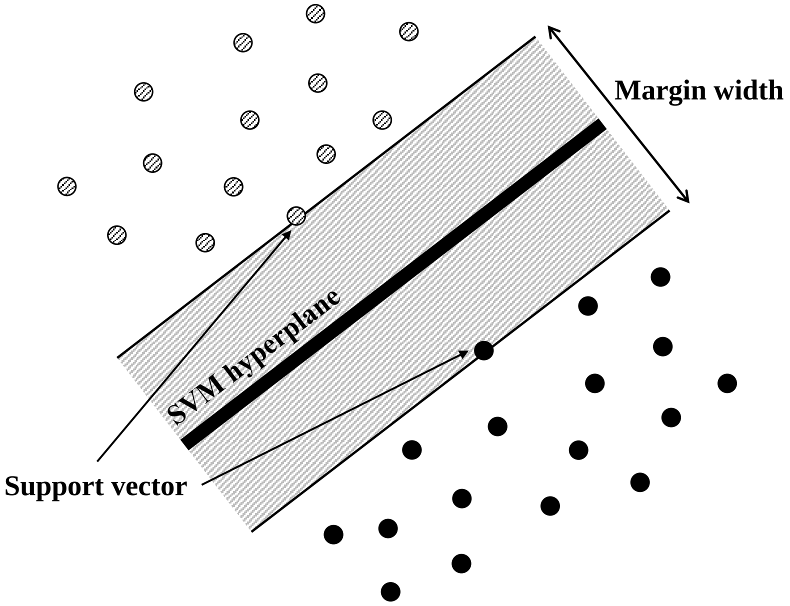


Figure 73 – Linear SVM

For example, the study [59] stated that RBF provided high classification efficiency, even though the polynomial kernel with second degree also achieved encouraging results. However, the computational time required by non-linear kernel functions may not be negligible.

Thus, the same study [59] concluded that the linear kernel function is preferable if classification results have to be quickly obtained. The RBF accuracy was also confirmed by Zhou *et al.* in [90].

In line with the literature survey, this thesis investigates the effectiveness of both linear and RBF SVM approaches to assess the most suitable SVM approach to detect UMIs, by evaluating benefits and drawbacks of each approach.

Among the different non-linear kernel functions, only the RBF approach is investigated, since its accuracy was demonstrated in several studies. Instead, the linear approach may provide promising results since pairs of UOMs can be linearly separated. In addition, the linear approach generally requires lower computational time than the RBF approach. The influence of several SVM parameters, e.g., the amount of data for training and number of classes, is also evaluated.

4.1.1. Kernel functions

In an SVM with *linear* kernel function, training data are separated by means of a linear hyperplane (see Fig. 73). Thus, linear SVM is recommended when datasets are linearly separable, as in case of UOM selection, and if multiple features are involved; in addition, the training time is lower than non-linear SVM.

Instead, a *Radial Basis Function* (RBF), such as all non-linear kernel functions, is recommended if training data are not linearly separable.

4.1.2. Multi-class classification

SVM is also employed to handle multi-class problems. In fact, thanks to the *divide-and-conquer* paradigm, the original multi-class problem is divided into several sub-problems that can be solved by binary classifiers that divide data into a two-class problem. In such a manner, the classification performance is increased.

As stated by Krawczyk *et al.* [91], the One-vs-All (OVA) and One-vs-One (OVO) decomposition are the most popular strategies to exploit the *divide-and-conquer* paradigm, thanks to their simplicity and accuracy.

The main difference between the OVA and OVO decomposition consists in the logic of the multi-class problem decomposition.

OVA decomposition. In OVA decomposition, also known as "One Against All" [92], a c-class problem is divided into c binary sub-problems, i.e., one per each class. Thus, each classifier is aimed at separating one class from all the others.

All classifier outputs are stored in one score-vector P; the element p_i ($p_i \in [0,1]$) represents the confidence (i.e., the posterior probability [93]) that samples are labeled by classifier i.

The data used for testing are labeled by the classifier that provides the highest confidence [91].

OVO decomposition. In the OVO decomposition, also known as "One Against One", "all pairs", "pairwise coupling", or "round robin" [92], a c-class classification problem is divided into $c(c-1)/2$ binary classification problems, by investigating all pairs of classes. Each classifier is trained by one pair of classes, whereas the rest is ignored. In such a manner, $c(c-1)/2$ classifiers are trained.

Then, outputs of all the binary classifiers are combined in score-matrix P. By considering the classifier that distinguishes class i from class j ($i \neq j$), p_{ij} ($p_{ij} \in [0,1]$) is the confidence (i.e., the posterior probability [93]) that a given sample is labeled in class i. Thus, $p_{ji} = 1 - p_{ij}$.

Detailed information regarding the calculation of p_{ij} are provided by Liu *et al*. [94]. The tested samples are labeled according to the score-matrix [94]; in particular, each classifier assigns data to one of its two training classes and, finally, the most voted class is labeled.

OVA vs OVO decomposition: benefits and drawbacks. Some literature studies investigated the most suitable decomposition approach on the basis of the research field and data features, by highlighting advantages and drawbacks of both approaches.

For example, López and Maldonado [95] stated that, according to their literature review, the OVA decomposition proves to be an efficient approach to handle multi-class problems. However, some class imbalances may be introduced if a class includes the majority of data, whereas only a few data refer to the remaining classes. Thus, data may be poorly classified if classes are not equally distributed [95, 96].

Krawczyk *et al.* [91] stated that OVO decomposition is generally more accurate than OVA, since simpler sub-problems are handled and imbalanced datasets are not created. Moreover, according to Milgram *et al.* [92], OVO decomposition may be preferable if a very large number of classes is available.

Training time is one of the major drawbacks of the OVO decomposition, since a higher number of classifiers has to be trained. In addition, as stated in [91, 93, 94], a further disadvantage relies on the fact that some trained classifiers are non-competent. In fact, each classifier is trained by means of a given pair of classes. Thus, when unknown data have to be labeled, each classifier votes in favor of one of the two classes involved in its training, even though the data used for testing may not belong to either of them.

Thus, both OVA and OVO decomposition can be effective, though their investigation is still an ongoing topic for research. In fact, OVA and OVO exploitation is strongly related to the features, i.e., data parameters involved in the training phase, of data involved in the multi-class problem.

In this thesis, the following SVM approaches are investigated:

- approach RBF OVO (RO);
- approach RBF OVA (RA);
- approach linear OVO (LO);
- approach linear OVA (LA).

4.2 NAÏVE BAYES

Naïve Bayes classifier is one of the most popular eager supervised classifier used for classification purposes.

Naïve Bayes classifier is based on Bayes' theorem and assumes that features are mutually independent. In practice, the independency assumption may be violated, but it allows to significantly simplify the classification task, without a dramatic reduction of prediction accuracy and robustness [45, 97].

Naïve Bayes classifier assigns each unlabeled vector $\boldsymbol{Y} = \{y_1, \ldots, y_c\}$ to the most probable class, i.e., that with the highest posterior probability (or conditional probability), calculated as in Eq. (25).

$$P(C_k|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|C_k)\, P(C_k)}{P(\boldsymbol{Y})} \tag{25}$$

Thus, the conditional probability that vector $\boldsymbol{Y}$ belongs to class C_k is correlated to the likelihood (or generative model) $P(\boldsymbol{Y}|C_k)$ and the class prior probability $P(C_k)$.

Class prior probability is calculated by means of the training phase, on the basis of available data for training.

As prior probability of a given class is the ratio between the number of data that belong to that class and the total number of data composing the training set, if balanced datasets are available, all classes are characterized by the same class prior probability. Instead, the likelihood is calculated on data for testing, by accounting for characteristics of data for training.

Based on the conditional independency assumption, likelihood $P(\boldsymbol{Y}|C_j)$ can be calculated separately for each variable, as in Eq. (26), by reducing a multidimensional problem to a one-dimensional problem:

$$P\left(Y_i \middle| C_j\right) = P\left(Y_i \middle| C_j, Y_{i+1}, \ldots, Y_c\right) \tag{26}$$

so that Eq. (25) can be written as follows (Eq. (27))

$$P\left(c_j \middle| \boldsymbol{Y}\right) = \frac{P\left(c_j\right) \prod_{i=1}^{c} P\left(\boldsymbol{Y}_i \middle| c_j\right)}{P(\boldsymbol{Y})} \tag{27}$$

Finally, according to Eq. (28), the predicted label is the one that maximizes the numerator of Eq. (27), since the marginal probability $P(\mathrm{Y})$ (see denominator of Eq. (27)) is constant [97].

$$c_j = \arg\max P(c_j) \prod_{i=1}^{c} P(Y_i|c_j) \tag{28}$$

Several Naïve Bayes classifiers exist; they differ from each other by the assumption of the likelihood distribution $P(Y_i \mid C_j)$. In the literature, the Gaussian Naïve Bayes, Bernoulli Naïve Bayes and Multinomial Naïve Bayes are the most popular Naïve Bayes classifiers.

In this thesis, as made in [97], a Gaussian Naïve Bayes classifier is assumed. Thus, the likelihood of x_i exhibits a normal distribution (see Eq. (29))

$$P(Y_i|C_j) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \bar{Y}_k}{\sigma_k}\right)^2\right] \tag{29}$$

, where σ_k and $\bar{\mathrm{Y}}_k$ are the standard deviation and mean value of class k, respectively. It is worth noting that, for each class, standard deviation and mean value are calculated for each feature.

Naïve Bayes classifier is broadly exploited for several real-world applications, such as spam filtering and document classification, as well as for diagnostics purposes. For example, Aralikatti *et al.* [98] exploited the Naïve Bayes classifier to classify the faulty condition of a carbide tool.

Niazi *et al.* [61] applied a Naïve Bayes classifier to detect hotspot in solar photovoltaic modules. Experimental results achieved an accuracy higher than 94 %.

Recently, Naïve Bayes classifier has also been applied for detecting and classifying faults occurring in transmission lines [45, 99]. For example, Aker *et al.* [99] developed a novel approach that combines the Discrete Wavelet Transform and Naïve Bayes classifiers; as a result, the novel approach outperformed alternative classifiers, e.g., Multi-Layer Perceptron Neural Network.

Instead, da Silva *et al.* [45] designed a Naïve Bayes classifier aimed at detecting multiple fault scenarios occurring in transmission lines, by reaching a classification accuracy equal to 95 %.

Benefits and drawbacks. One of the main benefits of the Naïve Bayes classifier consists in the fact that it can be exploited when both limited and very large datasets are available. In fact, the classification model is easily built, and its parameters are not estimated by means of complex iterations [84].

In spite of its simplicity, the Naïve Bayes classifier can efficiently handle complex real-world problems, even though an explicit model for its deployment cannot be provided [84].

4.3. K-NEAREST NEIGHBORS

K-Nearest Neighbor (K-NN) classifier is a non-parametric classifier, i.e., it does not assume a pre-defined distribution of data [46]. K-NN classifies each unlabeled data as the most frequent label among its *K* nearest labeled neighbors, which are identified by means of the Minkowski distance, calculated as in Eq. (30).

$$d = \left(\sum_{i=1}^{c} \left| x_{\mathrm{i}} - y_{\mathrm{i}} \right|^{l} \right)^{\frac{1}{l}} \tag{30}$$

In Eq. (30), x_{i} and y_{i} are the coordinates of each unlabeled and labeled data in the *c*-dimensional space, respectively. Based on the value of *l*, different distance definitions are encountered. For example, if *l* is set equal to 1, *d* represents Manhattan distance, whereas Eq. (30) is Euclidean distance if *l* is equal to 2.

Benefits and drawbacks. The K-NN classifier is generally defined as a lazy learner, since an explicit training step is not performed [100]: labeled data are only stored and a proper model is not created. Thus, time consumption for "training" the classifier is minimized.

As stated by Zhang *et al.* [62] and Sanchez *et al.* [46], the K-NN classifier is exploited in several application fields thanks to its conceptual simplicity, which makes it also suitable for handling multi-classification problems.

Despite its simplicity, the K-NN classifier proved to be highly competitive, since its accuracy can be equal or even higher than that more complex classifiers [101]. However, the sensitivity of the K-NN classifier to the *K* value is one of the major drawbacks and challenges related to the application of such classifier. To this purpose, several studies were addressed to identify the optimal *K* value for accuracy maximization.

In general, 1-NN (i.e., $K = 1$) is considered a promising classifier, such that it has been considered hard to be beaten for classification purposes [102].

However, it may provide misleading results, since the label is assigned by means of only the nearest neighbor. Thus, the reliability of the 1-NN classifier may be compromised if outliers are included within the dataset. Instead, according to [100], given that N_{tr} is the number of available labeled data, K may be set equal to $\mathrm{N}_{\mathrm{tr}}^{0.5}$ if the dataset is larger than 100 samples. However, this rule of thumb has been proved to not be suitable for all datasets [100].

Finally, the K-NN classifier may be significantly challenged if few labeled data are available with respect to the number of features [46] or in case that unbalanced datasets are considered [62].

In the literature, several attempts have been made to increase the K-NN reliability and robustness, as well as to smooth the negative effect of outliers included in the dataset. For example, Chaudhuri [103] developed a novel definition of neighborhood, which is the basis of a novel approach (i.e., Nearest Centroid Neighbors) for capturing both the concept of proximity and symmetric placement.

In the literature, several studies exploited both the traditional and improved K-NN classifiers to diagnostics purposes.

For example, Shi *et al.* [60] applied the K-NN classifier to detect different fault types affecting rotors. The authors compared the K-NN accuracy to other classifiers, e.g., linear SVM, by stating that K-NN exhibited the highest fault diagnosis accuracy.

Aslinezhad and Hejazi [104] used the K-NN classifier for detecting and determining clearance and deformation of turbine blade tip. The authors varied the K number, by achieving acceptable results in terms of accuracy.

Finally, Moshrefzadeh [105] developed a novel approach, namely subspace K-Nearest Neughbors, for online condition monitoring and diagnosis of rolling element bearings. The author compared the diagnostics effectiveness of the novel K-NN classifiers to a SVM learner that exploits a cubic kernel function and the OVO approach. Regarding the subspace K-NN, only the nearest neighbor is considered, i.e., $K = 1$. According to results, the novel K-NN classifier outperformed SVM.

Thus, according to the literature survey, this thesis evaluates the effectiveness of the K-NN classifier by considering three different K values.

By assuming that N_{tr} is the number of labeled data sampled in c different UOMs, among which the K neighbors are selected:

(i) $K = 1$;

(ii) $K = (N_{\mathrm{tr}}/c)^{0.5}$;

(iii) $K = N_{\mathrm{tr}}^{0.5}$.

Independently of the K number, neighbors are identified by means of the Euclidean distance.

It has to be mentioned that in K-NN the posterior probability of a given class is the ratio between the number of the K neighbors that belong to that class and the K value. Thus, when $K = 1$, confusion matrix and posterior probability are numerically equal [106].

4.4 IMPROVED NEAREST NEIGHBOR

In this thesis, a novel classification methodology is tuned with the aim to overcome the two main drawbacks of the K-NN classifier, so that (i) the setup of the K parameter can be avoided and (ii) the negative effect of outliers, i.e., out of range values, can be reduced.

It is worth noting that the novel classifier significantly differs from the methodology developed by Chaudhuri [103], namely Nearest Centroid Neighbors, which is instead based on a novel definition of neighborhood proposed by the same author. Such novel definition allows to classify one unlabeled data, i.e., x, by considering the K neighbors that are (i) near to x and (ii) symmetrically placed around x. According to the author, the second requirement is guaranteed by calculating the mean, i.e., centroid, of the K neighbors.

The Nearest Centroid Neighbors methodology proposed by Chaudhuri identifies the K neighbors as follows.

The first neighbor of x is its nearest labeled data.

Then, the remaining $(K - 1)$ neighbors are found by means of an iterative procedure. For identifying the second neighbor, the centroid of each labeled data and the first neighbor is calculated.

Among all labeled data, the second neighbor is the one that places the centroid as near as possible to the unlabeled data x.

Then, the procedure is repeated until all K neighbors are identified.

However, both drawbacks of the K-NN classifier may not be overcome. For example, the K value has to be set.

In order to overcome the two above-mentioned drawbacks of the K-NN classifier, the Improved Nearest Neighbor methodology developed in this thesis comprises three straightforward steps:

(i) the centroid of each UOM class is calculated. To this aim, data for training are gathered based on their label and, for each class, the mean value of each feature is calculated. Such mean values define the coordinates of centroids;

(ii) the distance between each data for testing and each centroid is calculated. In order to provide a rigorous comparison with the K-NN classifier, the Euclidean distance is computed in this thesis;

(iii) each data for testing is labeled as its nearest centroid.

For the sake of clarity, Fig. 74 depicts the principle of operation of the novel methodology for a two-class classification problem, i.e., $c = 2$, by accounting for two features, i.e., F_1 and F_2. Light-blue and dark-blue symbols represent labeled data that belong to the first and second class, respectively.

After that centroid of each class, reported in red, is calculated, the distance d between the unlabeled data (yellow symbol in Fig. 74) and each centroid is computed.

The centroid of the class that minimizes the distance to the unlabeled data provides the label to the field data. According to this procedure, the unlabeled data reported in Fig. 74 belong to class #1.

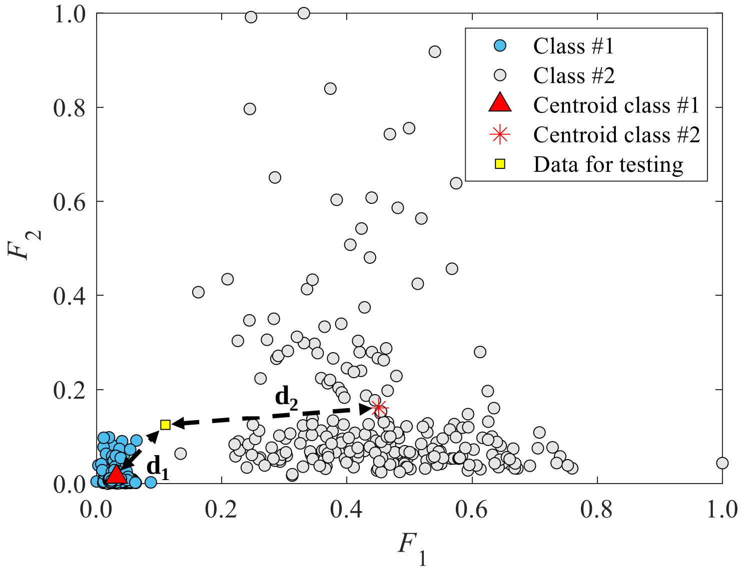


Figure 74 – Improved Nearest Neighbor methodology for two-class classification problem

Benefits and drawbacks. The Improved Nearest Neighbor methodology is a straightforward classifier characterized by the same benefits of the K-NN classifier, e.g., low computational time, but K-NN drawbacks are softened.

For instance, similarly to 1-NN, the Improved Nearest Neighbor methodology does not require to tune the K parameter.

Though the novel methodology classifies data according to one single point, it may be more robust than 1-NN, since the label is provided by one centroid, which accounts for the distribution of all data for training. Thus, the negative effect of outliers may be smoothed.

As in 1-NN, confusion matrix and posterior probability are numerically equal.

4.5 TRAINING AND TESTING DATA

Regardless of the ML classifier, Fig. 75 sketches the general procedure adopted to identify the data used for the training and testing phases.

It has to be mentioned the fact that in Fig. 75 the word “training” is improperly used for the K-NN and Improved Nearest Neighbor classifiers; in fact, a proper model is not built. For these models, the training phase consists in the detection of neighbors.

First, all field data of each GT must be preliminarily labeled by means of the supposed label. Then, labeled field data can be filtered by means of rough acceptability thresholds, which depend on the considered physical quantity. All data exceeding the acceptability thresholds (i.e., out of range values) will be only used for testing since they may be candidate UMIs.

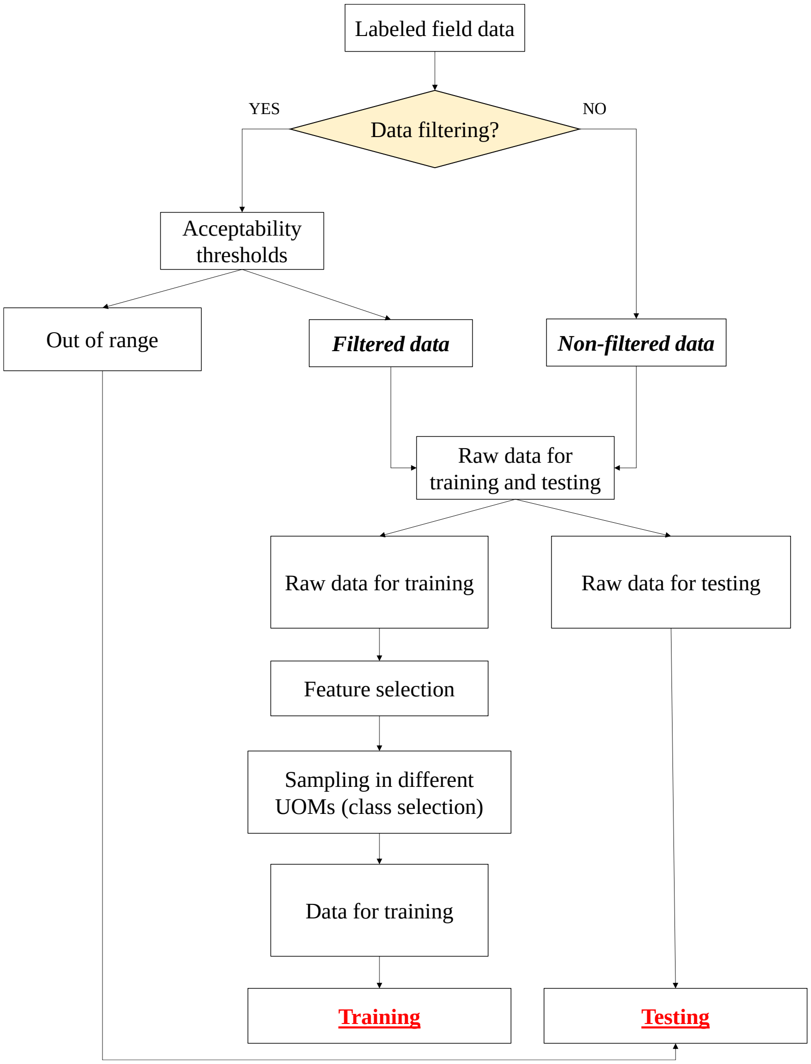


Figure 75 - Flowchart of the procedure for identifying training and testing data

Instead, data lying within acceptability thresholds, namely *filtered data*, will be used for both training and testing the classifiers. Otherwise, if labeled field data are not filtered by means of acceptability thresholds, all data composing the dataset, namely *non-filtered data*, are involved in both training and testing phases.

A fraction of *filtered data* or *non-filtered data* is selected to train the classifier ("Raw data for training"), whereas the remaining data will be tested ("Raw data for testing"). After that, the *features*, i.e., data parameters involved in the training phase, are chosen.

Then, raw data for training are sampled in different UOMs (so, they are converted to different UOMs), so that the *class* number, i.e., c ("Sampling in different UOMs" step), is defined. It has to be noted that sampling in different UOMs make the different classes inherently balanced.

Thus, training is carried out by means of part of *filtered* or *non-filtered* data and the same data converted into the candidate UOMs. In this thesis, N_{tr} represents the number of data that train the SVM and Naïve Bayes classifiers ("data for training"). Such a value is also exploited by K-NN for the K value set up ($K = N_{\mathrm{tr}}^{0.5}$ and $K = (N_{\mathrm{tr}}/c)^{0.5}$). It has to be highlighted that the ratio N_{tr}/c corresponds to the number of "Raw data for training".

The remaining *filtered* or *non-filtered* data and out of range values, if any, are tested. Thus, unlike the training phase, only labeled field data with the original UOM are employed to test the SVM approach reliability.

4.6. FIELD DATA

The analyses carried out in this thesis consider one dataset that refers to a pressure sensor acquired from Siemens GTs in operation.

The dataset accounts for thirty GTs of the same type, which run in eleven different sites. As shown in Fig. 76, multiple GTs usually run in the same site.

Labeled field data were derived from steady-states of which the mean value, the standard deviation, the minimum value and the maximum value were calculated over 60 consecutive time points. Thus, GT behavior is provided in the same operating conditions and coherent comparisons can be performed.

Because of confidentiality reasons, Fig. 76 shows the maximum variation of the physical quantity under analysis, by providing the nondimensional values calculated with respect to the maximum acceptability threshold. The minimum and maximum thresholds define the expected range of variation of the physical quantity. All measurements, 2721 in total, were originally labeled as *kPa absolute*. As it can be seen in Fig. 76, the majority of data lie in the range 0.65 - 1.0, i.e., the nondimensional acceptability thresholds of the considered GT type.

Conversely, approximately 13 % of data (acquired from GT #28, #29 and #30, i.e., the GTs installed in Site #11) is significantly lower than the minimum threshold. Out of range values are approximately two orders of magnitude lower than the minimum acceptability threshold. A further data inspection made by Siemens confirmed the UMI occurrence and the correct UOM, i.e., *bar absolute*, was subsequently assigned, since the minimum and maximum values stored in Site #11 were two orders of magnitude lower than that of the other sites.

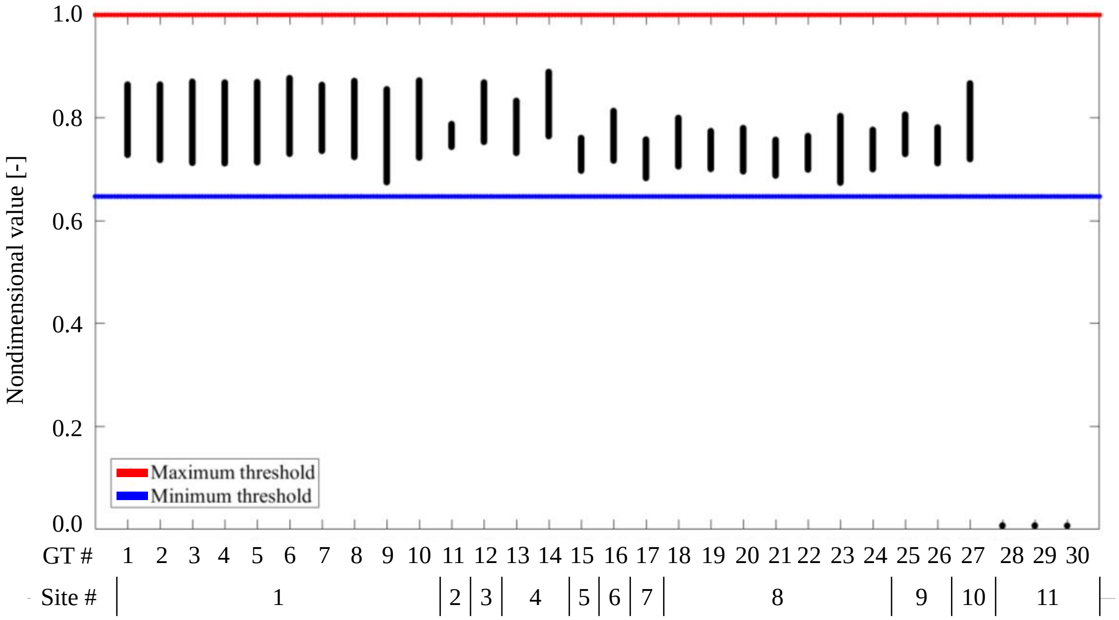


Figure 76 - Dataset composition

4.7. PERFORMANCE ANALYSIS

Several analyses are carried out to evaluate capability and robustness of all ML classifiers.

In fact, the classifier performance may be affected by several parameters, i.e., field data reliability, characteristics of data for training and *c* number.

For three (i.e., SVM, Naïve Bayes and K-NN) out of four classifiers, the dedicated tools available in the Matlab® environment are used. Instead, for the Improved Nearest Neighbor methodology, a dedicated routine is developed.

In this thesis, the number of features is set equal to 2, i.e., mean value and standard deviation of data. The influence of the number of features is discussed in Manservigi *et al*. [107].

The effect of each parameter under analysis is investigated as follows.

4.7.1. Field data reliability

Two different analyses are carried out to evaluate the robustness of each classifier with respect to data quality. In fact, classifier performance could be significantly affected by the acceptability threshold values, since they discriminate between filtered data and out of range values since the beginning.

First, acceptability thresholds are set so that filtered data only are employed to train the classifiers; in such a manner, out of range values are added to the testing data.

Secondly, acceptability thresholds are not set. Thus, all labeled field data are used for training and testing.

4.7.2. Training data

Two different analyses are carried out to evaluate the robustness of each classifier.

Training by using 10 % of the data. For each GT, 10 % of filtered or non-filtered data are employed to train the classifier, while the remaining 90 % of data are tested. This analysis is aimed at challenging classification prediction by using a limited training dataset. Further analyses dealing with the effect of the amount of data employed for training the classifier are discussed in Manservigi *et al*. [107].

Site cross-validation. This analysis is addressed to evaluate the robustness of each classifier at labeling data acquired from a novel site, i.e., whose data are not employed in the training phase. This analysis is carried out as follows. Among filtered or non-filtered data, all data acquired from one site in turn are tested, while all data of remaining sites train the classifiers. Thus, if filtered data are employed, data acquired from two sites (i.e., Site #11 and one additional site in turn) are tested. Conversely, only one site is tested if non-filtered data are accounted for.

Number of classes. This thesis investigates the capability of each classifier to identify the true UOM label by varying the c number. Two analyses are carried out as follows.

Absolute and gauge UOMs. First, starting by the assumed UOM, i.e., *kPa absolute* for all sites, training data are converted into 11 further UOMs, i.e., *kPa gauge* and mmH_2O, *mbar*, inH_2O, *kPa*, *psi*, *bar*, as both absolute and gauge values. Thus, the correct label has to be identified among twelve classes.

Figure 77 depicts the twelve available UOMs, by highlighting the conversion factors between *kPa absolute* and the other absolute UOMs. Such conversion factors have general validity regardless of the considered dataset. Conversely, the relationship between absolute and gauge

UOMs depends on the dataset under analysis, though gauge values always lie on the right side of the respective absolute UOM. In Fig. 77 the position of gauge values is representative of the actual correlation between each absolute and its corresponding gauge UOM.

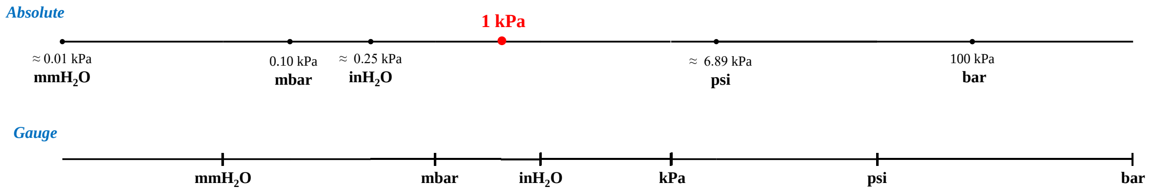


Figure 77 - UOM conversion factors

It is worth noting that all classifiers are significantly challenged. In fact, for the considered dataset, the conversion factor between *kPa absolute* and *inH$_2$O gauge* is equal to just 1.2. Moreover, another challenge is represented by the use of non-filtered data, i.e., incorrectly labeled data, for training the classifiers.

Absolute UOMs. In the second step of the analysis, only the six absolute UOMs are considered. Subsequently, among them, the UOM that is assigned with the lowest posterior probability is deleted, by reducing the *c* number, until the most probable four classes are detected.

In this case, the identification of the correct UOM is mostly challenged in the case of *mbar vs. inH$_2$O* and *inH$_2$O vs. kPa*, since the scale factors are equal to 2.5 and 4 (see Fig. 77), respectively. In addition, each classifier is further challenged by using non-filtered data for training.

4.7.3. Computational time

For the sake of industrial attractiveness, the computational time to train each classifier is also accounted for to identify the preferable ML technique.

4.7.4. Classification performance

The classification capability is evaluated by accounting for:

(i) for each GT, the amount of data that are equally labeled. In fact, if a UMI occurred, all data collected by a given GT have to be labelled in the same class. This piece of information is conveyed by the confusion matrix. The confusion matrix also provides the classification accuracy, i.e., rate of data that are successfully labeled;

(ii) the posterior probability that a given datum is classified in a given UOM class. For each datum, the sum of the posterior probability of all classes has to be 100 %;

(iii) for each GT, a different amount of data is available.

4.8. RESULTS AND DISCUSSION

For each ML classifiers, this Section discusses the results of the analyses with (i) twelve UOMs including absolute and gauge pressures and (ii) six and four absolute pressures.

When twelve UOMs are accounted for, results referring to the site cross-validation are also reported. In fact, this is the most challenging case study since a larger number of classes and greater proximity between subsequent UOMs are taken into account.

With regard to the case study only dealing with the absolute UOMs, even though the number of classes could be further reduced, the minimum number of classes evaluated in this thesis is equal to four. In fact, as discussed below, the posterior probability with which the UOM class is identified is generally very high by using four UOMs.

Moreover, it is worth pointing out that, independently of the *c* number, the training phase is performed by means of 10 % of both filtered and non-filtered data.

As discussed in Section 4.6, the field dataset comprises two true UOMs, i.e., *kPa absolute* for Site #1 through #10 and *bar absolute* for Site #11. For the sake of clarity, for each true UOM, the average posterior probability and the average classification accuracy are reported. In such a manner, overall rules of thumb can be grasped.

The complete confusion matrices are only reported for the direct comparison of the optimal ML classifiers. Instead, in Paragraphs 4.8.1. through 4.8.4 only the rate of correctly labeled data, i.e., classification accuracy, is shown. In both cases, classification accuracies higher than 95 % are highlighted in black.

For the sake of clarity, full and dotted bars refer to absolute and gauge pressures, respectively.

Different combinations of field data reliability, training and testing data and number of classes are discussed in the following.

4.8.1. Support Vector Machine

Absolute and gauge pressures (twelve UOMs): training with 10 % of filtered data. As can be seen in Table 10 and Fig. 78, both OVO approaches (i.e., RO and LO) prove to successfully detect the true UOM. The classification accuracy reported in Table 10 points out that the large majority of data (greater than 95.5 %) is classified within the true class in spite of the challenging low conversion factor between *kPa absolute* and *inH$_2$O gauge*.

Table 10 – Classification accuracy for twelve UOMs, training with 10 % of filtered data

Site	True UOM	RO	RA	LO	LA
Site #1 through #10	*kPa absolute*	95.5%	81.0 %	98.3 %	81.6 %
Site #11	*bar absolute*	99.7 %	0.0 %	99.7 %	0.0 %

Posterior probability reported in Fig. 78 confirms the outstanding result. In fact, for Site #1 through #10, the true label (i.e., *kPa absolute*) is assigned on average with 87 % and 82 % of posterior probability by RO and LO, respectively.

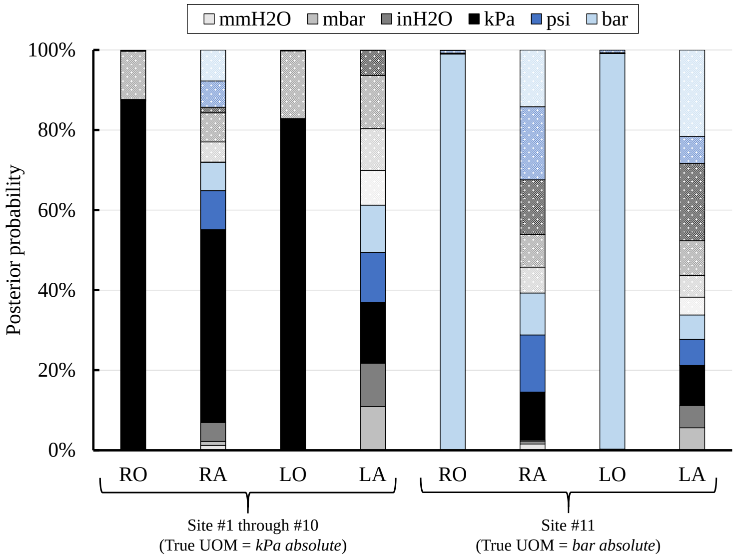


Figure 78 - Posterior probability for twelve UOMs, training with 10 % of filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

Because of the challenging conversion factor, *inH$_2$O gauge* proves to be the second-most probable UOM. In fact, 4.0 % and 1.1 % of data are labelled as *inH$_2$O gauge*, though its posterior probability is on average equal to 12 % and 17 % for RO and LO, respectively. Instead, posterior probabilities of other UOMs are negligible.

These noticeable results are confirmed also for Site #11, whose 99.7 % of data are classified as *bar absolute* with the 99 % of posterior probability by means of both OVO approaches.

Instead, OVA approaches exhibit scattered and misleading results, especially for Site #11, in which approximately the 99 % of data are labelled as *psi gauge* (RA) or *bar gauge* (LA).

Absolute and gauge pressures (twelve UOMs): training with 10 % of non-filtered data. This analysis is performed to evaluate the response of SVM approaches if incorrectly labelled data (i.e., out of range values acquired from Site #11) are employed to train the model. As can be grasped from Table 11 and Fig. 79, results obtained by means of filtered data are also generally confirmed by using non-filtered data. In fact, RO and LO clearly identify the true UOM, which is detected with the highest posterior probability.

Table 11 – Classification accuracy for 12 UOMs, training with 10 % of non-filtered data

Site	True UOM	RO	RA	LO	LA
Site #1 through #10	*kPa absolute*	97.4 %	83.6 %	90.4 %	69.6 %
Site #11	*bar absolute*	99.1 %	0.0 %	98.8 %	0.0 %

As expected, classification performance is affected by training data reliability. In fact, as can be seen, results in Fig. 79 are more scattered than in Fig. 78 and posterior probabilities of incorrect UOMs increase. As a result, true UOMs are identified with lower posterior probabilities.

For instance, for Site #1 through #10 the *kPa absolute* label is assigned with 66 % of posterior probability by means of RO, while this rate is approximately equal to 78 % for Site #11. Instead, the true UOM is provided by means of LO with 45 % and 70 % of posterior probability for Site #1 through #10 and Site #11, respectively.

Thus, with respect to the analysis by means of filtered data only, LO is more affected by the reliability of data used for training the SVM approaches than RO.

In addition, based on the range of posterior probability variation, RO may be less dependent on the tested site than LO. However, the large majority of data (Table 11), i.e., up to 97.4 % for RO and 90.4 % for LO, are correctly labeled.

OVA approaches generally confirm a poorer classification performance than RO and LO, both in terms of posterior probability and classification accuracy, especially for identifying the UOM of Site #11.

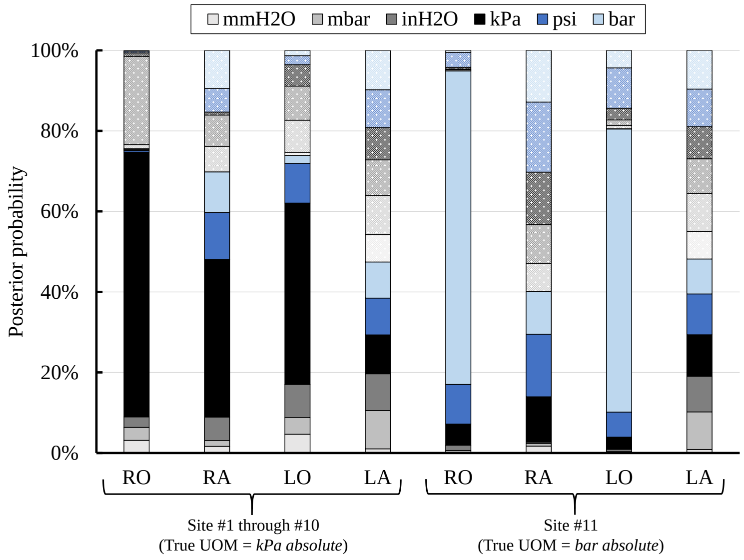


Figure 79 - Posterior probability for twelve UOMs, training with 10 % of non-filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of filtered data. Since filtered data are employed, data for training refer to nine sites selected from the ten *kPa absolute* sites (i.e., Site #1 through #10), while the remaining *kPa absolute* site and Site #11 are tested.

All site combinations are investigated and the average results in terms of classification accuracy and posterior probability are reported in Table 12 and Fig. 80.

Table 12 – Classification accuracy for 12 UOMs: site cross-validation by means of filtered data

Site	True UOM	RO	RA	LO	LA
Site #1 through #10	*kPa absolute*	94.2 %	83.7 %	97.4 %	76.2 %
Site #11	*bar absolute*	100.0 %	0.0 %	100.0 %	0.0 %

Since each site accounts for a different number of data, each training combination is performed by means of a different rate of data, which varies from 62 % to 86 % of the data included in the dataset. As a result, this analysis further strengthens the classifier reliability vs. amount of training data.

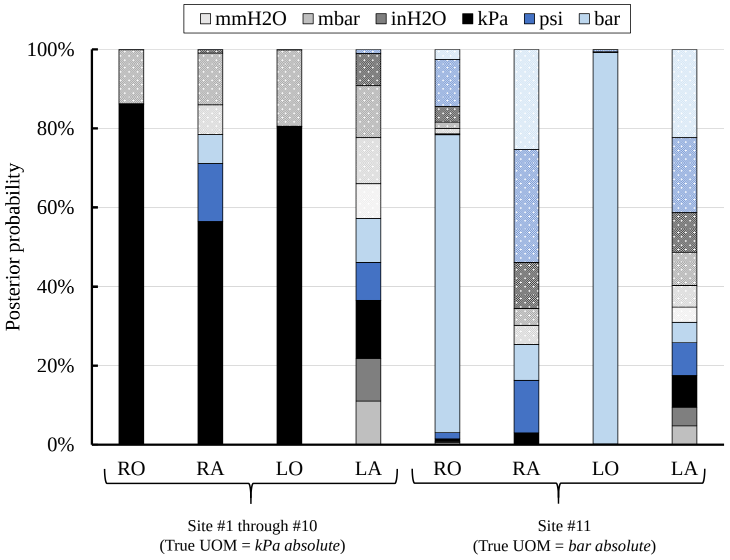


Figure 80 - Posterior probability for twelve UOMs:
site cross-validation by means of filtered data
(full bar: absolute UOMs, dotted bar: gauge UOMs)

By comparing Table 10 *vs*. Table 12 and Fig. 78 *vs*. 80, the performance of SVM approaches is generally confirmed. As a general comment, RO and LO exhibit significantly better results than RA and LA.

Regarding Site #1 through #10, the large majority of data (at least 94 %) are correctly labeled and the *kPa absolute* label is assigned with at least 80 % of posterior probability by means of the OVO approaches.

Regarding Site #11, RO provides the true label by means of a lower posterior probability than in Fig. 78, i.e., approximately equal to 75 %, even though all tested data are classified within the correct class. Instead, the posterior probability of the *psi gauge* class (i.e., the closest class to *bar absolute*) increases. Thus, as a general comment, the RO reliability is confirmed.

Finally, scattered and poorer results are once again provided by means of the OVA approaches.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of non-filtered data. As can be seen through Table 13 and Fig. 81, the assignment of the true UOM by means of RO is slightly dependent on the site under analysis. At least 93 % of data are correctly identified, while the posterior probability with which the *kPa absolute* and *bar absolute* labels are provided are in the range from 65 % (Site #11) to 70 % (Site #1 through #10).

Table 13 – Classification accuracy for 12 UOMs: site cross-validation by means of non-filtered data

Site	True UOM	RO	RA	LO	LA
Site #1 through #10	*kPa absolute*	92.8 %	85.1 %	86.3 %	7.0 %
Site #11	*bar absolute*	100.0 %	0.0 %	100.0 %	0.0 %

Conversely, the true class identification by means of LO proves to be more dependent on the considered site. In fact, approximately 86 % of *kPa absolute* data (i.e., Site #1 through #10) are correctly identified by means of an average posterior probability that is equal to 44 %, while all tested data acquired from Site #11 are correctly labeled with 99 % of posterior probability. RA and LA detections are not reliable, especially for Site #11.

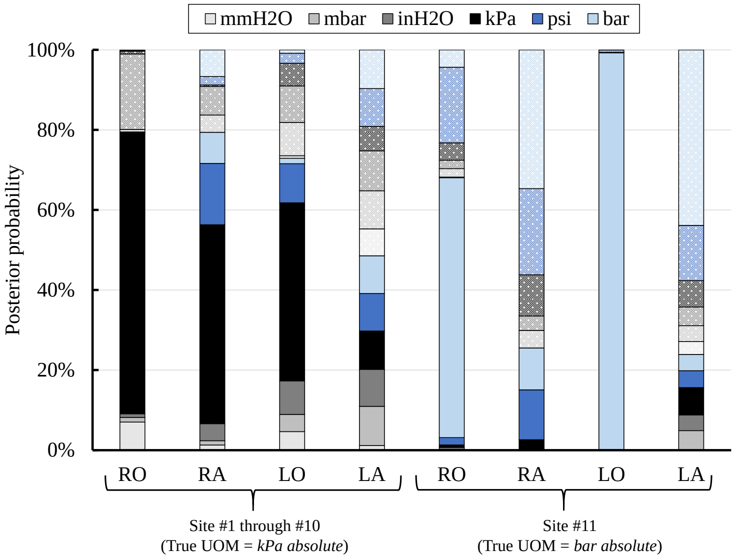


Figure 81 - Posterior probability for twelve UOMs: site cross-validation by means of non-filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

Absolute pressures (six UOMs): training with 10 % of filtered data. If absolute pressures are only taken into account RO, RA and LO provide similar results (see Table 14 and Fig. 82).

In fact, RO, RA and LO correctly identify all data (see Table 14).

Table 14 – Classification accuracy for six absolute UOMs: training with 10 % of filtered data

Site	True UOM	RO	RA	LO	LA
Site #1 through #10	*kPa absolute*	100.0 %	100.0 %	100.0 %	0.0 %
Site #11	*bar absolute*	100.0 %	100.0 %	100.0 %	100.0 %

In addition, the UOM that was originally assigned, i.e., *kPa*, is confirmed for Site #1 through #10 with approximately the 99 % of probability, whereas Site #11 is correctly labeled in the *bar* class. RA also provides almost correct results, since the correct UOM is selected, though the posterior probability is lower than in RO and LO (from 52 % to 78 % for Site #11 and the other sites, respectively). In this case, the *psi* and *inH*$_2$*O* posterior probabilities increase, but, as shown in Fig.82, these UOMs are the closest classes to *kPa*. Thus, the result proves to be physically sound. Instead, the correct UOM is poorly assigned by means of LA, since all UOMs, especially for Site #1 through #10, are almost equally probable. In particular, all data are labeled as *inH*$_2$*O*.

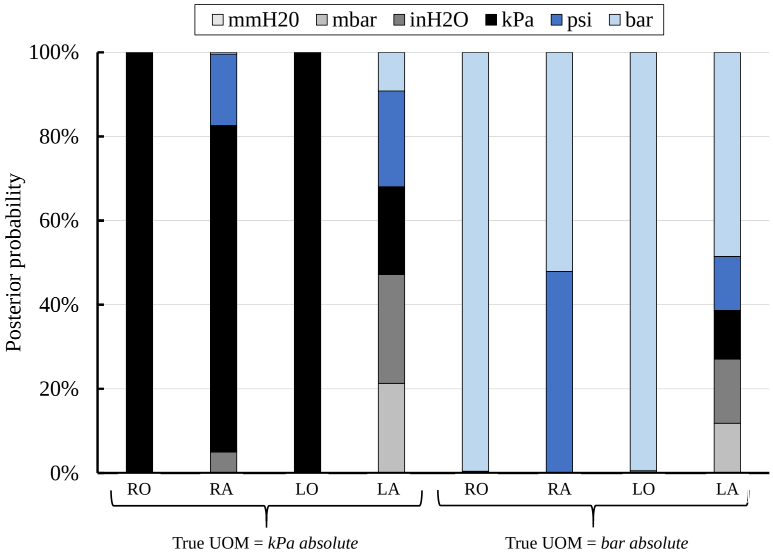


Figure 82 - Posterior probability with six absolute UOMs: training with 10 % of filtered data

Absolute pressures (four UOMs): training with 10 % of filtered data. Based on the results obtained by considering six absolute pressures, the *mmH*$_2$*O* class is assigned with the lowest posterior probability for all the considered approaches. Thus, in the test that accounts for only five absolute UOM classes, the *mmH*$_2$*O* class is deleted. In this case, RA classification significantly improves, by achieving the RO and LO posterior probability, i.e., approximately the 99 %. Instead, LA results are scattered. According to the results, *mbar* class is subsequently deleted. The results that refer to the four-class classification, i.e., *bar*, *psi*, *kPa* and inH_2O, are shown in Table 15 and Fig. 83.

As can be seen by means of Table 15, the reduction of number of classes allows to significantly improve classification capability of all SVM approaches. In fact, at least 99.7 % of data are classified by means of the true label.

Table 15 – Classification accuracy for 4 absolute UOMs: training with 10 % of filtered data

Site	True UOM	RO	RA	LO	LA
Site #1 through #10	*kPa absolute*	100.0 %	100.0 %	100.0 %	100.0 %
Site #11	*bar absolute*	100.0 %	100.0 %	100.0 %	99.7 %

Such univocal response is confirmed by means of posterior probability (Fig. 83). In fact, the posterior probability of the true UOM reaches 100 % for RO, RA and LO. As a general comment, LA performance significantly increases, even though the *psi* posterior probability ranges from 11 % to 46 % on average. This result can be explained by observing that *psi absolute* is the intermediate UOM between *bar absolute* (the correct UOM of Site #11) and *kPa absolute* (the correct UOM of all the other sites).

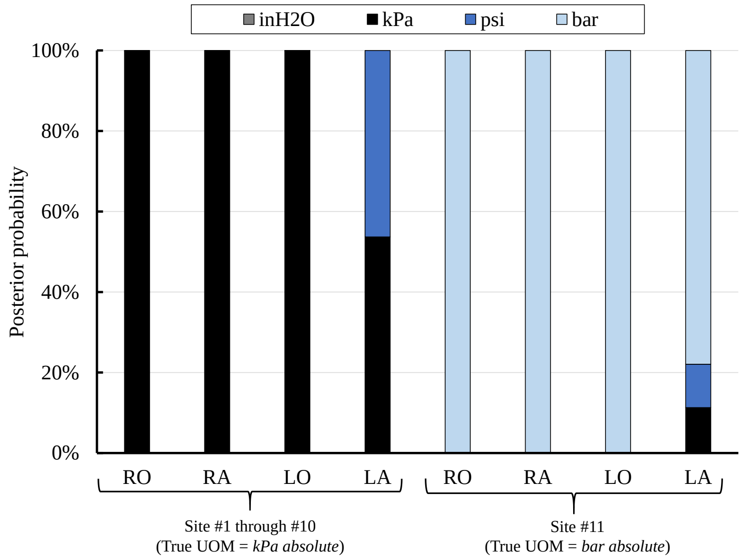


Figure 83 - Posterior probability with four absolute UOMs:
training with 10 % of filtered data

Absolute pressures (six UOMs): training with 10 % of non-filtered data. Table 16 confirms the classification results reported in Table 14, with the exception of LA, which classifies all data collected by Site #1 through #10 as *mbar* instead of *inH₂O*.

Table 16 – Classification accuracy for six absolute UOMs: training with 10 % of non-filtered data

Site	True UOM	RO	RA	LO	LA
Site #1 through #10	*kPa absolute*	100.0 %	100.0 %	100.0 %	0.0 %
Site #11	*bar absolute*	100.0 %	100.0 %	100.0 %	100.0 %

Figure 84 summarizes UOM classification when the SVM training is carried out by means of non-filtered data, i.e., also including out of range values. Thus, this test represents a more challenging scenario for UOM classification. As a matter of fact, UOM assessment is provided by means of lower posterior probabilities than the corresponding values reported in Fig. 82. However, the correct UOM is clearly provided by RO by means of a posterior probability that always exceeds 83 %.

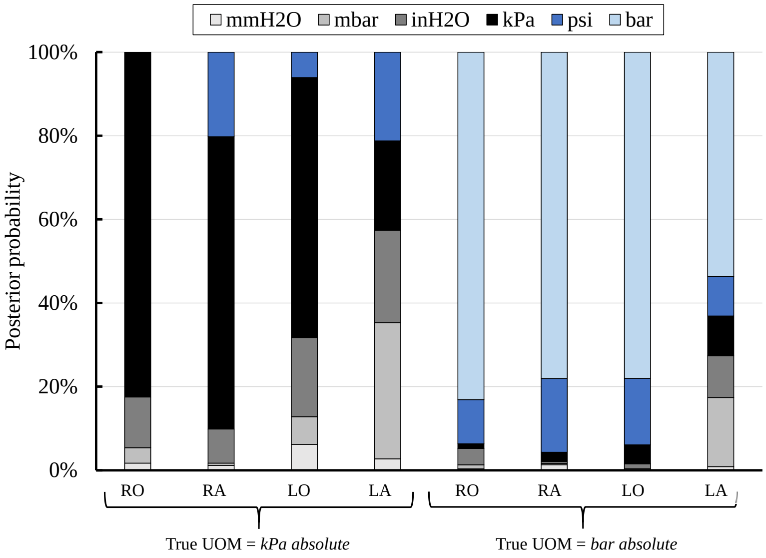


Figure 84 - Posterior probability with six absolute UOMs: training with 10 % of non-filtered data

A further interesting result is conveyed by the UOMs that RO selects. In fact, for Site #1 through #10, *in*H_2O, *mbar* and *mm*H_2O (i.e., pressure UOMs placed at the left-hand side of *kPa absolute*) are assigned by means of a posterior probability that increases passing from *mm*H_2O to *in*H_2O. Conversely, even though SVM training is carried out by involving bar data that were erroneously assumed to be *kPa*, the psi and bar posterior probability is equal to 0 %. This outcome could be explained by comparing the order of magnitude of each UOM (see Fig. 77). In fact, because of data collected in Site #11, the erroneous supposed *kPa* values tend to overlap the UOM provided by the other sites, i.e., *bar*, and the other UOMs are shifted towards lower values. A similar behavior is confirmed in RA, LO and LA, in which only *bar* posterior probability is always equal to 0 %.

As observed in the analyses carried out by means of twelve UOMs, LO is the SVM approach that is mainly affected by data quality. In fact, even though the correct UOM class is provided, the posterior probability is approximately equal to 78 % in Site #11 and 62 % in the other sites. RA results are slightly better than LO. Finally, LA classification is not univocally defined.

Then, the class number has been decreased, by first deleting *mm*H_2O and then *mbar*. As a result, the classification effectiveness significantly enhances. In fact, when only four classes are selected, i.e., *bar*, *psi*, *kPa* and *in*H_2O, the probability with which the true label is provided ranges from 72 % to 98 % for RO, RA and LO. Instead, LA confirms the poorest classification capability, even though the correct UOM is conveyed with 56 % of posterior probability at maximum.

Discussion. According to the results obtained in this Section, OVO approaches prove to be much more effective than OVA approaches. In fact, RO and LO correctly detect UMI occurrence even in the case that the true UOM has to be identified among twelve classes, whose conversion factors may be also very close to one. Instead, OVA approaches usually provide more scattered results in terms of posterior probability, especially when the number of classes increases. This outcome can be explained by the fact that OVA approaches generate imbalanced datasets.

Thus, outcomes achieved in this thesis agree with the results obtained by Krawczyk et al. [91]. Conversely, the assumption made by Milgram *et al*. [92], which suggests employing OVO decomposition when a very large number of classes is available, is partially verified. In fact, even though the most probable UOMs were identified, the number of available classes considered in this thesis is not higher than twelve.

By varying the reliability of data used for training SVM classifiers, RO demonstrates to be the most suitable approach for UMI detection, also if erroneous labeled data are included within the training dataset. However, as demonstrated by varying the number of classes, LO effectiveness can be significantly enhanced by reducing the number of available classes. Instead, RA reaches the RO performance only when the number of classes is reduced.

It is worth noting that RO requires the highest computational time, which ranges from a few seconds to approximately 3 minutes. Such a difference in computational time may seem negligible; however, it can significantly affect SVM application if a large amount of data is analyzed.

Thus, the analyses carried out in this thesis are confirmed twice, since UMI is correctly detected and the true UOM is successfully provided.

4.8.2. Naïve Bayes

This paragraph discusses classification results, reported in Fig. 85 through 89, provided by means of the Naïve Bayes classifier. Each figure documents both Posterior Probability (PP) and Confusion Matrix (CM) for all sites.

Absolute and gauge pressures (twelve UOMs): training with 10 % of filtered data. Figure 85 shows field data classification provided by the Naïve Bayes classifier.

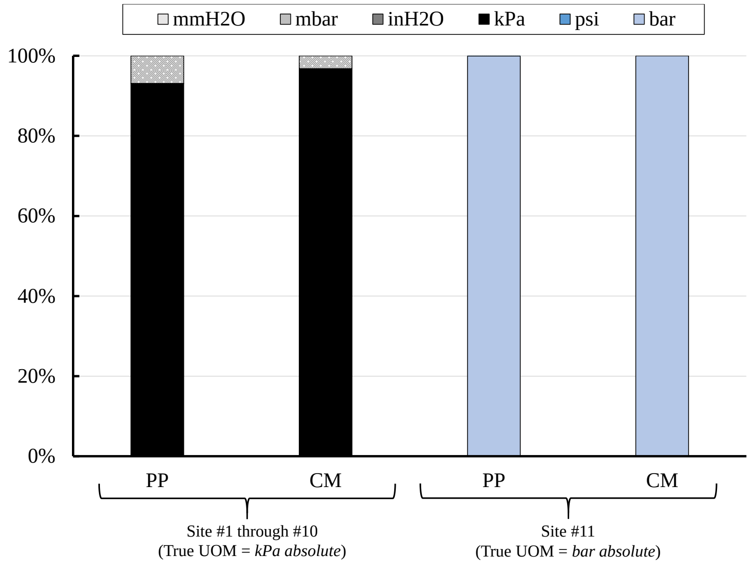


Figure 85 - Posterior probability and confusion matrix for twelve UOMs, training with 10 % of filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

As can be grasped, the true UOM is correctly assigned. In fact, on average, 97 % of data acquired from Site #1 through #10 are correctly labeled as *kPa absolute*, with 93 % of posterior probability.

The *inH$_2$O gauge* label only slightly challenges the classifier. In fact, only a few data, i.e., approximately equal to 3 %, are erroneously labeled, so that the average posterior probability of *inH$_2$O gauge* is equal to 7 %. It has to be highlighted that posterior probability of remaining UOMs is equal to zero.

Thus, even though several classes are available, the Naïve Bayes classifier allows to focus the identification of the correct label between only two alternatives.

Finally, all unlabeled data acquired from Site #11 are unequivocally classified as *bar absolute*, by means of a posterior probability equal to 100 %.

Absolute and gauge pressures (twelve UOMs): training with 10 % of non-filtered data. Similarly to SVM, more scattered results are obtained when non-filtered are involved within the training phase. In fact, approximately 94 % of data (see Fig. 86) acquired from Site #1 through #10 are correctly labeled as *kPa absolute*, with 84 % of posterior probability on average, while posterior probability of the remaining UOMs slightly increase with respect to the filtered data analysis (Fig. 85).

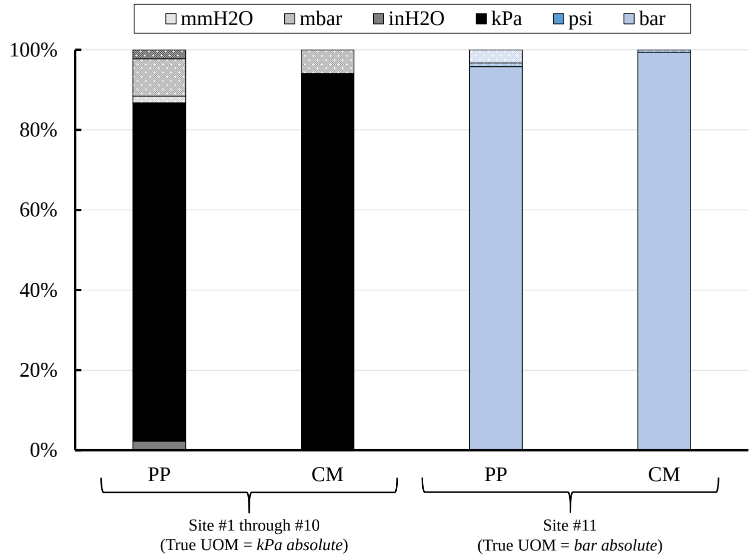


Figure 86 - Posterior probability and confusion matrix for twelve UOMs, training with 10 % of non-filtered data
(full bar: absolute UOMs, dotted bar: gauge UOMs)

For example, *inH$_2$O gauge* and *kPa gauge* posterior probabilities are equal to 9 % and 2 %, respectively.

Site #11 proves to be less affected by reliability of training data than the other sites. In fact, the *bar absolute* label is assigned to 99 % of data with 96 % of posterior probability, while *psi gauge* and *bar gauge*, i.e., the nearest UOMs to *bar absolute*, are rarely accounted for.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of filtered data. Results obtained by means of training with 10 % of filtered data (Fig. 85) are confirmed by the site cross-validation, by proving the robustness of the Naïve Bayes classifier even when the rate of training data is varied. In fact, Fig. 87 shows that on average 94 % and 100 % of data are correctly labeled as *kPa absolute* and *bar absolute*, respectively, while posterior probability is in the range from 90 % to 100 %.

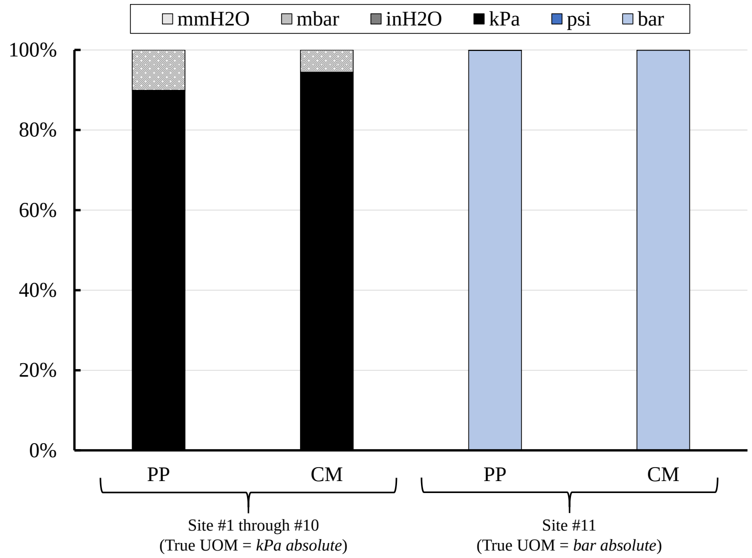


Figure 87 - Posterior probability and confusion matrix for 12 UOMs: site cross-validation by means of filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of non-filtered data. Site cross-validation carried out by means of non-filtered data confirms results obtained by using 10 % of data. In fact, as depicted in Fig. 88, approximately 92 % of data collected by Site #1 through Site #10 are correctly labeled, by means of 83 % of posterior probability.

As can be grasped by comparing Fig. 88 to Fig. 87, classification accuracy is only slightly affected by reliability of training data. In fact, the decrease in terms of classification accuracy and posterior probability is approximately equal to 2 % and 8 %, respectively.

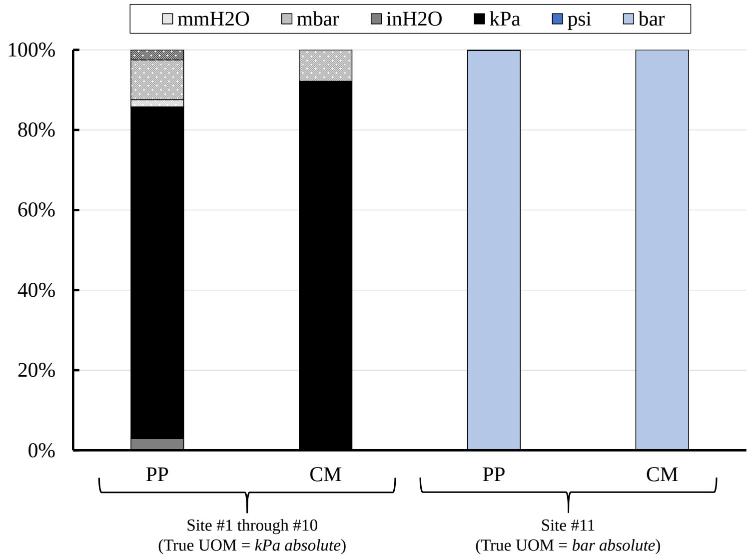


Figure 88 - Posterior probability and confusion matrix for twelve UOMs: site cross-validation by means of non-filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

Absolute pressures (six UOMs): training with 10 % of filtered data. When only absolute UOMs are considered, Naïve Bayes classifier clearly assigns the correct label to all unlabeled data with 100 % of posterior probability, independently of the site. Thus, for the sake of brevity, the respective figure is omitted.

As a result, unlike SVM, results achieved by considering five and four absolute UOMs are not discussed.

Absolute pressures (six UOMs): training with 10 % of non-filtered data. As can be seen in Fig. 89, both posterior probability and confusion matrix clearly provide the correct UOM for all sites. In fact, for Site #1 through #10, 94 % of data are correctly labeled as *kPa absolute*, with approximately the same posterior probability (i.e., 93 %). For these sites, *inH$_2$O absolute*, i.e., the nearest UOM to *kPa absolute*, is the second-most probable UOM, though only 7 % of data are labeled by means of the *inH$_2$O* absolute label.

Instead, all the tested data that refer to Site #11 are correctly identified.

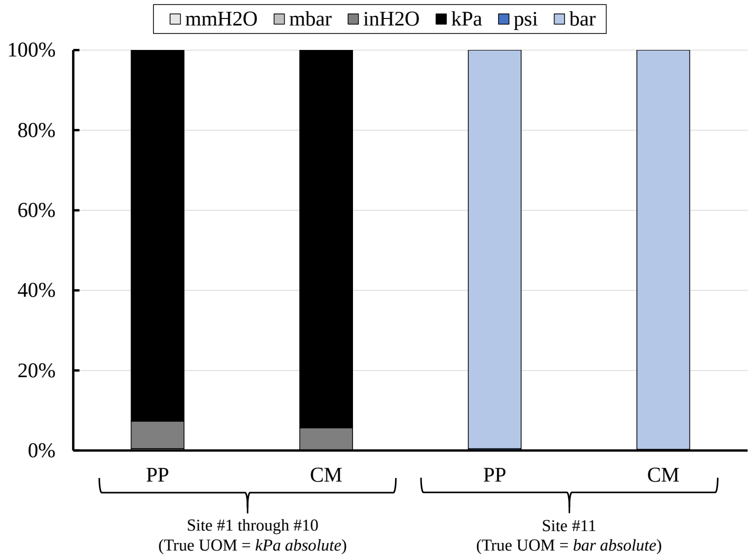


Figure 89 - Posterior probability and confusion matrix with six absolute UOMs: training with 10 % of non-filtered data

Discussion. As a general comment, Naïve Bayes classifier provides extremely positive results both in terms of classification accuracy and posterior probability. In addition, the classifier proves to be a robust methodology, since classification accuracy is only slightly affected by reliability of data for training. In fact, rate of data correctly labeled is always higher than 92 %.

Thanks to the low computational time, each analysis was concluded in 1 s.

4.8.3. K-Nearest Neighbor

This Paragraph documents classification results provided by the K-NN classifier, by varying the number of neighbors, i.e., $K = 1$, square root of the number of raw data for training ($K = (N_{tr}/c)^{0.5}$) and square root of the number of data for training ($K = N_{tr}^{0.5}$). As previously discussed, for this classifier, "data for training" are the labeled data among which neighbors can be identified.

In has to be mentioned that, independently of the given test, the K neighbors represent a negligible rate of data with respect to the "data for training". In fact, when neighbors are detected among 10 % of filtered data and $K = (N_{tr}/c)^{0.5}$, K corresponds to 1.9 % and 2.7 % of "data for training" if twelve and six UOMs are considered, respectively. Such a rate is equal to 0.5 % and 1.1 % if $K = N_{tr}^{0.5}$. Approximately the same rates are confirmed if non-filtered data are accounted for. Results of each test are summarized by means of Table 17 through 21 and Figs. 90 through 94, by reporting a brief snapshot of classification accuracy and posterior probability.

Absolute and gauge pressures (twelve UOMs): training 10 % of filtered data. Independently of the number of neighbors, the K-NN classifier provides satisfactory results (see Table 17 and Fig. 90).

In fact, the classifier selects two UOM classes only, i.e., *kPa absolute* and *inH₂O gauge*, i.e., the nearest class to *kPa absolute*, for Site #1 through #10.

As reported in Table 17, the rate of data that are correctly labeled ranges from 77 % (i.e., when $K = N_{tr}^{0.5}$) to 94 % (i.e., when $K = (N_{tr}/c)^{0.5}$). Instead, when only one neighbor is accounted for, 89 % of data are successfully identified (Table 17). As a consequence, the rate of data classified as *inH₂O gauge* can be as high as 23 %.

Table 17 – Classification accuracy for twelve UOMs: training with 10 % of filtered data

Site	True UOM	$K = 1$	$K = (N_{tr}/c)^{0.5}$	$K = N_{tr}^{0.5}$
Site #1 through #10	*kPa absolute*	89.4 %	94.4 %	77.3 %
Site #11	*bar absolute*	100.0 %	100.0 %	99.7 %

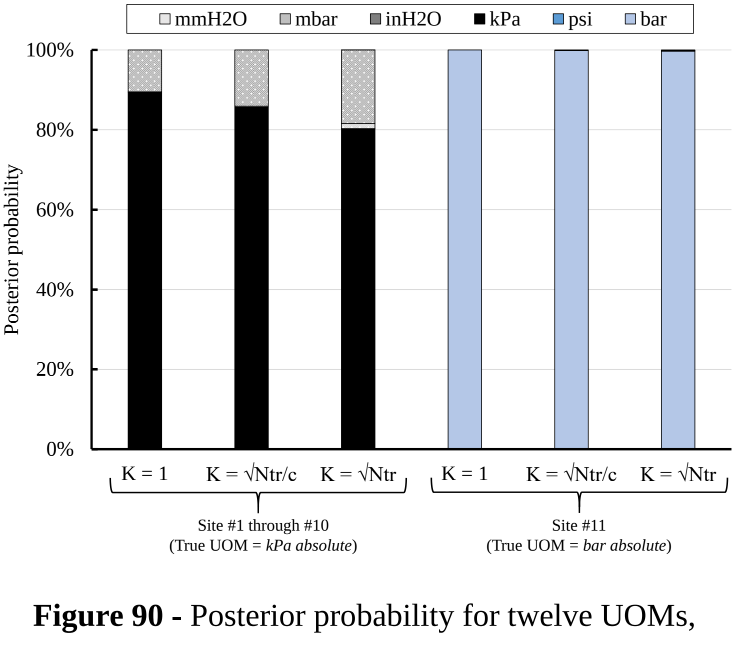


Figure 90 - Posterior probability for twelve UOMs, training with 10 % of filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

These results are confirmed by means of Fig. 90, which reports the average posterior probability for each K-NN classifier.

As can be seen, the correct label is provided by means of a posterior probability that slightly decreases by increasing the number of neighbors, from 89 % ($K = 1$) to 80 % ($K = N_{\mathrm{tr}}^{0.5}$).

For Site #11, all data for testing are correctly labeled with 100 % of posterior probability.

Absolute and gauge pressures (twelve UOMs): training with 10 % of non-filtered data. Table 18 and Fig. 91 summarize classification accuracy and posterior probability matrix when non-filtered data are employed. Regarding Site #1 through #10, both posterior probability and classification accuracy exhibit the same behavior shown in Table 17 and Fig. 90. In fact, as can be grasped by means of Fig. 91, the rate of correctly labeled data varies in the range from 76 % to 89 %, while posterior probability decreases from 82 % to 71 %, by increasing the number of neighbors.

Table 18 – Classification accuracy for 12 UOMs, training with 10 % of non-filtered data

Site	True UOM	$K = 1$	$K = (N_{\mathrm{tr}}/c)^{0.5}$	$K = N_{\mathrm{tr}}^{0.5}$
Site #1 through #10	*kPa absolute*	81.5 %	89.5 %	76.2 %
Site #11	*bar absolute*	44.3 %	54.4 %	99.7 %

Unlike Fig. 90, three UOMs are selected, i.e., *kPa absolute*, *inH$_2$O gauge* and *mmH$_2$O absolute*. This result relies on the fact that *inH$_2$O gauge* is the nearest UOM class to the correct label, i.e., *kPa absolute*. In addition, when data acquired from Site #11, i.e., incorrectly assumed as *kPa absolute*, are sampled in different UOMs, data incorrectly labeled as *mmH$_2$O absolute* overlap the region on which the *kPa absolute* label lies. In fact, as shown in Fig. 77, the scale factors between (i) *bar absolute* and *kPa absolute* and (ii) *kPa absolute* and *mmH$_2$O absolute* are approximately the same. Thus, in the same region three different UOMs are approximately overlapped.

More in detail, roughly 9 % of data are labeled as *mmH$_2$O absolute* when K is set equal to 1, while this rate decreases by increasing the number of neighbors. Instead, posterior probability is almost constant and equal to 9 %. Regarding the *inH$_2$O gauge* label, both posterior probability and rate of labeled data are equal to 9 % if $K = 1$, i.e., as well as the *mmH$_2$O absolute* label, while posterior probability and rate of labeled data can be as high as 18 % and 24 %, respectively, when *K* increases.

Finally, posterior probability of the *mbar gauge* label slightly increases by increasing the number of neighbors. In fact, as highlighted in Fig. 77, *mbar gauge* is the second-closest UOM to *kPa absolute*.

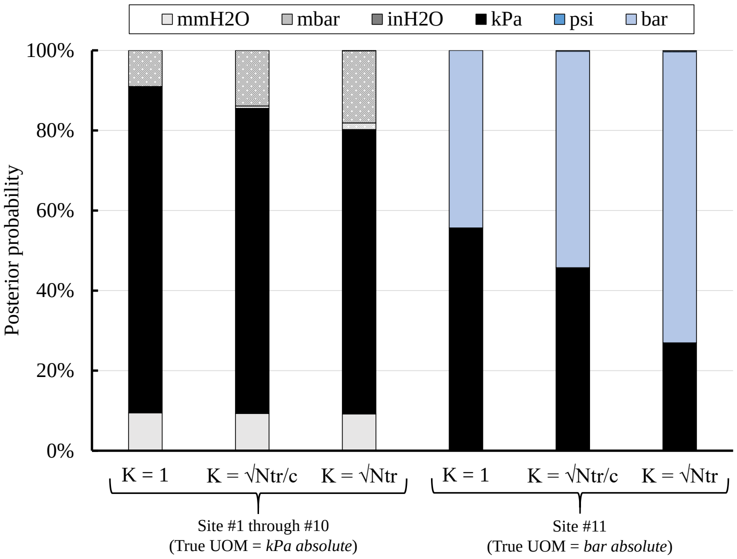


Figure 91 - Posterior probability for twelve UOMs,
training with 10 % of non-filtered data
(full bar: absolute UOMs, dotted bar: gauge UOMs)

Regarding Site #11, more scattered results are obtained, and drawbacks correlated to the 1-NN classifier are emphasized. In fact, 56 % of data are erroneously labeled as *kPa absolute*, while remaining data are classified as *bar absolute*. This result can be explained by the fact that, because of the original label of raw data, both the *kPa absolute* and *bar absolute* labels coexist within the same region. However, the K-NN classification capability significantly enhances when a higher number of neighbors is considered, since the negative effect of out of range values is smoothed. In fact, as shown in Table 18, approximately all data used for testing are correctly labeled and the posterior probability increases up to 73 %, when $K = N_{\mathrm{tr}}^{0.5}$.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of filtered data. Comments regarding training carried out by means of 10 % of filtered data are confirmed by means of the site cross-validation. In fact, as shown in Table 19, for Site #1 through #10, maximum rate of correctly labeled data (i.e., 94 %) is achieved.

Table 19 – Classification accuracy for 12 UOMs: site cross-validation by means of filtered data

Site	True UOM	$K = 1$	$K = (N_{\mathrm{tr}}/c)^{0.5}$	$K = N_{\mathrm{tr}}^{0.5}$
Site #1 through #10	*kPa absolute*	91.7 %	94.1 %	89.5 %
Site #11	*bar absolute*	100.0 %	100.0 %	100.0 %

In addition, posterior probability is characterized by a decreasing trend by increasing the value of the K parameter (Fig. 92).

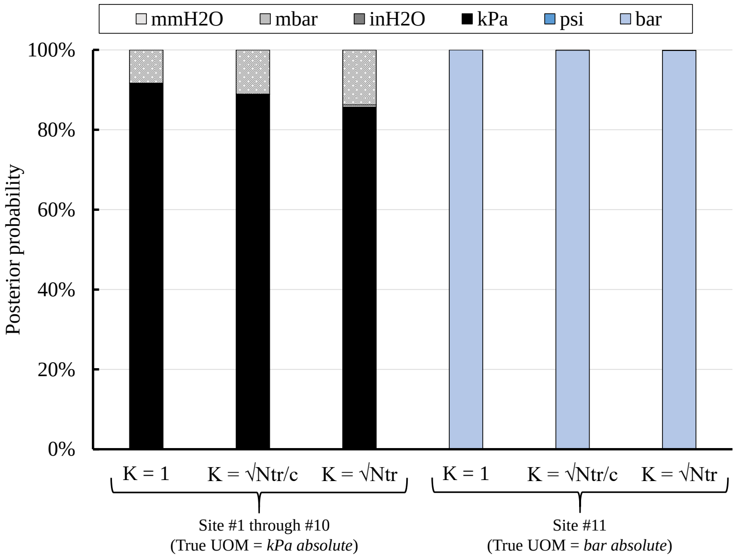


Figure 92 - Posterior probability for twelve UOMs: site cross-validation by means of filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

In addition, the maximum posterior probability (i.e., 92 %) is obtained when $K = 1$.

However, it is worth noting that the effect of number of neighbors on posterior probability and classification accuracy is more smoothed than in Table 17 and Fig. 90, since their variation is not higher than 4 %.

Instead, all data used for testing Site #11 are correctly labeled and posterior probability is equal to 100 %.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of non-filtered data. Results reported in Table 20 and Fig. 93 qualitatively confirm the outcomes obtained by means of the previous analyses. In fact, for Site #1 through #10, K-NN selects the same three UOMs identified by the analysis that exploits 10 % of non-filtered data for training the classifier (Table 18 and Fig. 91).

Table 20 - Classification for 12 UOMs: site cross-validation by means of non-filtered data

Site	True UOM	$K = 1$	$K = (N_{tr}/c)^{0.5}$	$K = N_{tr}^{0.5}$
Site #1 through #10	*kPa absolute*	78.6 %	91.9 %	84.6 %
Site #11	*bar absolute*	100.0 %	100.0 %	100.0 %

An interesting result is represented by the posterior probability of the *mmH₂O absolute* label, which is roughly equal to 13 %, regardless of the value of the K parameter.

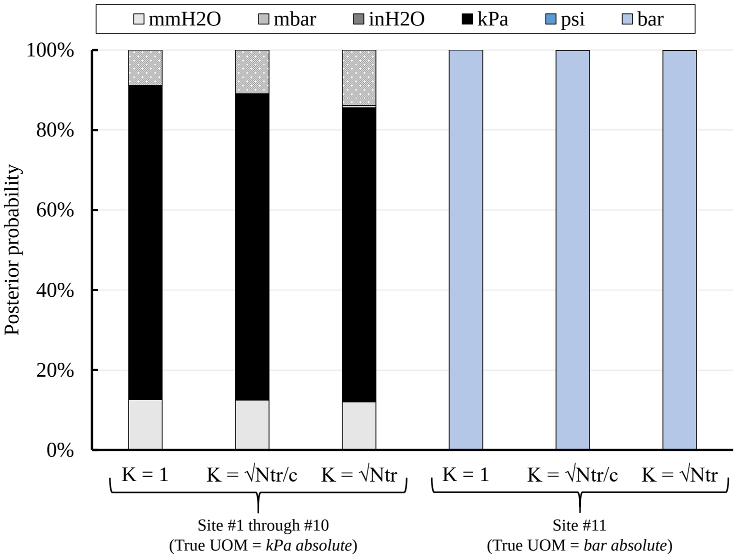


Figure 93 - Posterior probability for twelve UOMs:
site cross-validation by means of non-filtered data
(full bar: absolute UOMs, dotted bar: gauge UOMs)

Such analysis exactly reflects the composition of the dataset, in which out of range values are roughly equal to 13 %. Classification accuracy ranges from 79 % (i.e., $K = 1$) to 92 %, while posterior probability slightly varies from 77 % to 79 %.

Instead, Site #11 classification by means of site cross-validation significantly differs from the one obtained by exploiting 10 % of non-filtered data (Table 18 and Fig. 91). In fact, the current analysis allows to correctly label all data. This result can be explained by the fact that site cross-validation consists of testing all data acquired from a given site. Thus, when Site #11 is tested, neighbors are identified only within correctly labeled data. This effect allows the maximization of classification accuracy.

Absolute pressures (six UOMs): training with 10 % of filtered data. As outlined in the corresponding analysis carried out by means of the Naïve Bayes classifier, all data are correctly labeled.

This is an expected result, since the *in*H_2O *gauge* label in not accounted for. Posterior probability confirms the outstanding results, since it is equal to 100 % for all sites. Thus, for the sake of brevity, the corresponding table and figure are omitted.

For this reason, results achieved by considering five and four absolute UOMs are not reported in this thesis.

Absolute pressures (six UOMs): training with 10 % of non-filtered data. The comparison between analyses that include (i) twelve UOMs (Table 18 and Fig. 91) and (ii) six absolute UOMs (Table 21 and Fig. 94) highlights that if gauge UOMs are removed, the large majority of data that were labeled as *in*H_2O *gauge* (Site #1 through #10) are labeled as *kPa absolute*, so that up to 100 % of data are correctly classified.

Table 21 – Classification accuracy for 6 absolute UOMs, training with 10 % of non-filtered data

Site	True UOM	$K = 1$	$K = (N_{\mathrm{tr}}/c)^{0.5}$	$K = N_{\mathrm{tr}}^{0.5}$
Site #1 through #10	*kPa absolute*	90.3 %	99.4 %	99.8 %
Site #11	*bar absolute*	48.9 %	54.1 %	100.0 %

Posterior probability increases up to 90 %, proving to be almost independent of the number of neighbors.

Conversely, classification accuracy for Site #11 is generally confirmed, though the posterior probability increases and varies in the range from 49 % to 74 %. The most promising results are obtained when the maximum number of neighbors is considered, since the true label is always provided with an average posterior probability equal to 74 %.

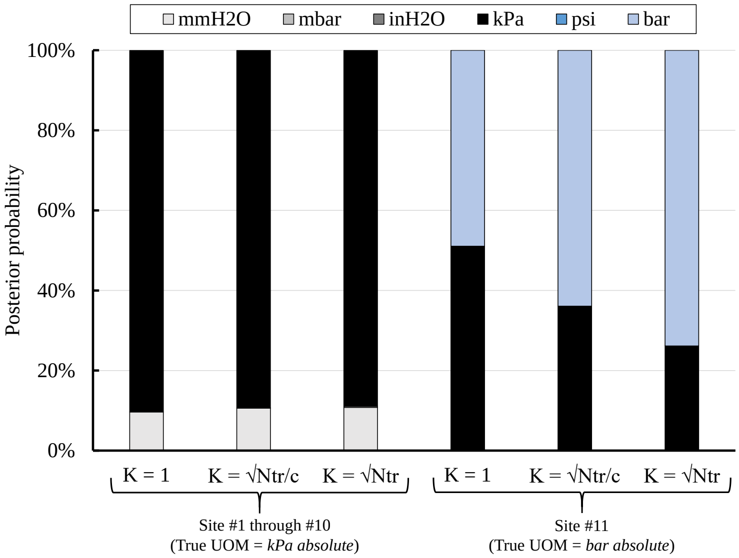


Figure 94 - Posterior probability for six absolute UOMs, training with 10 % of non-filtered data

4.8.4. Improved Nearest Neighbor

This Paragraph of the thesis discusses classification results provided by means of the novel Improved Nearest Neighbor methodology. Unlike SVM, Naïve Bayes and K-NN, results are only briefly commented; no figures are reported since posterior probability and confusion matrix correspond.

Absolute and gauge pressures (twelve UOMs): training with 10 % of filtered data. The novel methodology and the best K-NN classifier, i.e., the one that maximizes classification accuracy, are comparable. In fact, classification accuracy is as high as 91 % for Site #1 through Site #10, while roughly 9 % of data are incorrectly labeled as *inH*$_2$*O gauge*. Instead, the *bar absolute* label is provided to all data acquired from Site #11.

Absolute and gauge pressures (twelve UOMs): training with 10 % of non-filtered data. This test highlights benefits of the Improved Nearest Neighbor methodology with respect to the K-NN classifier. It has to be reminded that, as reported in Table 18 and Fig. 91, the K-NN classifier is especially challenged at classifying data acquired from Site #11. In fact, only 54 % of data were correctly labeled, though the effectiveness of the classifier improved by increasing the number of neighbors.

Instead, the Improved Nearest Neighbor methodology correctly classifies all data collected by Site #11, as well as 89 % of data of the other sites. In fact, for Site #1 through #10, only 10 % and 1 % of data are classified as *inH₂O gauge* and *mbar gauge*, respectively.

Thus, classification accuracy of the *kPa absolute* sites is approximately equal to the K-NN classifier characterized by the highest accuracy, i.e., $K = (N_{\mathrm{tr}}/c)^{0.5}$.

This outcome proves that outliers only slightly affect the classification capability of the Improved Nearest Neighbor, since the difference in terms of correctly labeled data with respect to the filtered data analysis is only equal to 2 %.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of filtered data. Site cross-validation confirms the classification accuracy provided by K-NN. In fact, for Site #1 through #11, 89 % of data are correctly labeled as *kPa absolute*, while the *bar absolute* label is associated to all data referring to Site #11.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of non-filtered data. As for the site cross-validation carried out by means of filtered data, the accuracy of the novel methodology is comparable to the accuracy of the K-NN classifier. In fact, 89 % of data are correctly labeled as *kPa absolute*, while this rate varied from 79 % to 92 % for the K-NN classifier.

In addition, the smoothed effect of outliers is confirmed. In fact, classification accuracy is as high as the one provided by using filtered data.

Finally, the *bar absolute* label is confirmed for all data acquired from Site #11.

Absolute pressures (six UOMs): training with 10 % of filtered data. The reduction of number of classes allows to enhance classification accuracy of the classifier. In fact, 99 % of data referring to Site #1 to Site #10 are correctly labeled. Thus, for these sites, the classification effectiveness is comparable to the maximum accuracy provided by K-NN. Instead, a few data (i.e., the remaining 1 %) are classified as *inH₂O absolute*. Similarly to the analysis that accounts for twelve UOMs, label of all tested data from Site #11 is correctly assigned.

As verified for the K-NN classifier, only three UOMs are selected: *kPa absolute* and *inH₂O absolute* for Site #1 through #10 and *bar absolute* for Site #11. Instead, the other UOMs, i.e., *mmH₂O absolute*, *mbar absolute* and *psi absolute*, are never accounted for. Thus, a further reduction of the number of candidate UOMs is not necessary.

Absolute pressures (six UOMs): training with 10 % of non-filtered data. If non-filtered data are used, the reduction of number of classes allows to further enhance classification accuracy. In fact, 98 % of data referring to Site #1 to Site #10 are correctly labeled. Thus, for the considered sites, the classification effectiveness is comparable to the maximum accuracy provided by the K-NN classifier. Instead, a few data, approximately 2 %, are identified as *in*H_2O *absolute*. As in the analysis that accounts for twelve UOMs, the label of all tested data from Site #11 is correctly assigned.

Discussion. This paragraph deals with a direct comparison between the Improved Nearest Neighbor and the K-NN classifiers. As expected, classification accuracy of the Improved Nearest Neighbor methodology is generally higher than that of the 1-NN; in fact, when filtered data are considered, the classification accuracy of both classifiers is roughly the same. Conversely, when non-filtered data are accounted for, the difference between the accuracy of the Improved Nearest Neighbor and accuracy of the 1-NN classifiers varies in the range from 7 % to 21 % for Site #1 through #10, while it is equal to 55 % for Site #11. These results demonstrate that the Improved Nearest Neighbor classifier is less affected by outliers than 1-NN. Instead, the comparison between the Improved Nearest Neighbor methodology and the K-NN classifier, which accounts for two different K values, reveals that the classification accuracy of the novel classifier can be up to 14 % higher than that of the K-NN that uses the maximum number of neighbors, i.e, $K = N_{\mathrm{tr}}^{0.5}$. Instead, benefits with respect to K-NN that uses $K = (N_{\mathrm{tr}}/c)^{0.5}$ are clearer when non-filtered data are employed, since in such a scenario the difference in terms of accuracy is equal to 46 %.

Thus, as general comment, the novel Improved Nearest Neighbor methodology may be considered a promising classifier for detecting UMI occurrence. In fact, outliers only slightly affect the classification capability of the Improved Nearest Neighbor methodology: as demonstrated, difference in terms of classification accuracy by using filtered and non-filtered data can be as high as 2 %.

Finally, both K-NN and the Improved Nearest Neighbor minimized computational time, both for training and testing the classifier. In fact, each analysis required 1 s.

4.8.5. Comparison of Machine Learning classifiers

This Paragraph summarizes and compares the results reported in Paragraphs 4.8.1 through 4.8.4 for identifying the optimal ML classifier to detect UMI occurrence.

The identification of the optimal ML classifier is mainly focused on four alternatives, i.e., SVM RO, Naïve Bayes (NB), K-NN by using $K = (N_{\mathrm{tr}})^{0.5}$ and the novel Improved Nearest Neighbor (INN) methodology, since they outperform the other approaches.

In fact, as discussed in Paragraph 4.8.1, among the four considered SVM approaches, i.e., RO, RA, LO and LA, RO proves to be the most suitable approach for UMI detection, also in the challenging case that erroneous labeled data are included within the training dataset.

Similarly, regarding the K-NN classifier, the most promising results are obtained when the K value is set equal to $N_{\mathrm{tr}}^{0.5}$.

In fact, the K-NN classifier coupled to the selected K value outperforms classification results achieved by means of $K = 1$ and $K = (N_{\mathrm{tr}}/c)^{0.5}$, especially when non-filtered data are accounted for.

Comprehensive confusion matrices are reported within the Appendix.

Absolute and gauge pressures (twelve UOMs): training 10 % of filtered data. The direct comparison among the four selected classifiers by using 10 % of filtered data is shown in Table A1 and Fig. 95.

As can be grasped from Table A1, SVM RO, Naïve Bayes and Improved Nearest Neighbor achieve approximately the same classification accuracy. In fact, regarding the *kPa absolute* sites, rate of correctly labelled data slightly varies from 91 % (i.e., Improved Nearest Neighbor) to 97 % (i.e., Naïve Bayes); instead, 95 % of data are correctly identified by means of SVM RO. The K-NN classifier achieves the poorest classification accuracy, though 77 % of data are properly labeled.

Posterior probability generally confirms results provided by the confusion matrix. In fact, Naïve Bayes classifier achieves the highest posterior probability, which is on average equal to 93 %, followed by the Improved Nearest Neighbor methodology (i.e., 91 %), SVM RO (i.e., 87 %) and K-NN (i.e., 80 %).

Instead, all data for testing that refer to Site #11 are univocally classified by all classifiers; in addition, posterior probability is always equal to 100 %.

Though all classifiers are trained by means of correctly labeled data, both classification accuracy and posterior probability of *kPa absolute* sites are usually lower than the ones obtained for Site #11. This is due to the fact that the scale factor between the true label of the site under analysis and its nearest UOM. In fact, as stated in Paragraph 4.7.2., conversion factor between *kPa absolute* and *inH*$_2$*O gauge*, i.e., the nearest UOM to *kPa absolute*, is only equal to 1.2, while the scale factor between *bar absolute* and *bar gauge* is roughly equal to 5. As a result, the assignment of the correct UOM is more challenging for Site #1 through #10 than for Site #11.

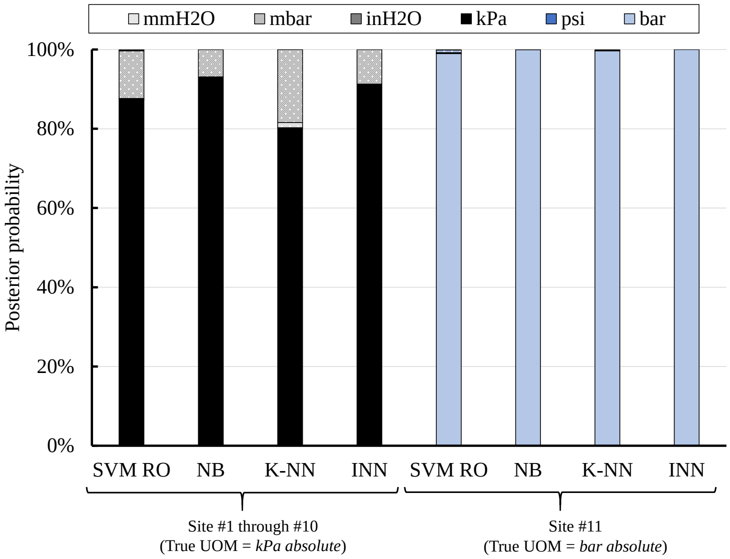


Figure 95 - Posterior probability for twelve UOMs, training with 10 % of filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

Absolute and gauge pressures (twelve UOMs): training with 10 % of non-filtered data. The comprehensive comparison of the four classifiers by using 10 % of non-filtered data is conveyed in Table A2 and Fig. 96. The comparison between the exploitation of filtered (Table A1) and non-filtered data (Table A2) reveals that all classifiers are only slightly affected by data reliability. In fact, reduction in terms of classification accuracy is not higher than 3 %, proving the robustness of each classifier.

As a general comment, the highest and homogeneous classification accuracy is provided by SVM RO, i.e., 97 % and 99 % for the *kPa absolute* and *bar absolute* sites, respectively.

Though the classification accuracy is generally confirmed, posterior probability proves to be more affected by data reliability. As a consequence, the label is correctly provided, but classifier confidence decreases.

Among all classifiers, SVM RO and K-NN exhibit the highest posterior probability decrease, which is equal to 24 % and 11 %, respectively, for Site #1 through #10 and equal to 21 % and 27 % for Site #11.

As a result, posterior probabilities provided by means of SVM RO are equal to 66 % (*kPa absolute* sites) and 78 % (*bar absolute* sites), while such rates are equal to 71 % (*kPa absolute* sites) and 73 % (*bar absolute* sites) for the optimal K-NN classifier.

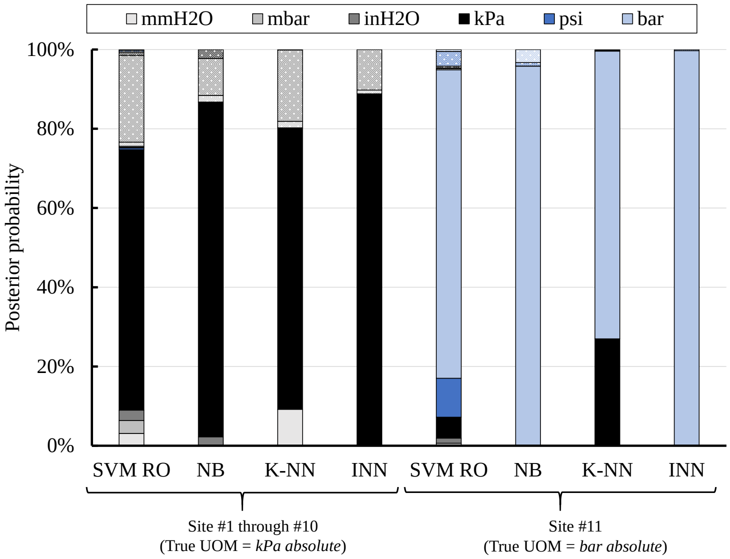


Figure 96 - Posterior probability for twelve UOMs, training with 10 % of non-filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

Instead, posterior probability calculated for the Naïve Bayes classifier only slightly decreases, up to 84 % for the *kPa absolute* sites and up to 96 % for the *bar absolute* site. Thus, the classification performance reduces by 9 % and 4 %.

Finally, the Improved Nearest Neighbor methodology exhibits the highest robustness, since posterior probability reduction in not higher than 3 %, so that the highest posterior probability is guaranteed (i.e., 89 %) for the *kPa absolute* sites.

A further interesting result, which highlights theoretical differences among the classifiers, can be derived from the comparison of posterior probabilities of all available UOMs.

As a general comment, SVM RO obtains the most scattered posterior probabilities in the considered classifiers.

For Site #1 through #10, UOMs at the left-hand side of *kPa absolute* (see Fig. 77) usually achieve higher posterior probability rates than UOMs placed at the right-hand side; this rule of thumb is valid with the exception of the *inH$_2$O gauge* label, which is the second-most probable label because of the challenging scale factor.

This interesting result may rely on the fact that the wrongly assumed *mmH$_2$O absolute* data (referring to Site #11) slightly overlap the *kPa absolute* data acquired from the other sites. Thus, the area that identifies the *mmH$_2$O absolute* label is shifted toward the *kPa absolute* label. As a consequence all UOMs located between *mmH$_2$O absolut*e and *kPa absolute,* i.e., *mbar absolute*, *inH_2O absolute* and *mbar gauge*, are directly influenced by this effect, so that their posterior probabilities are higher than that of the other UOMs. However, this outcome cannot be clearly grasped from Fig. 96 because few outliers are included within the dataset.

For the same reason, *psi absolute* and *kPa absolute* are the second and third-most probable labels, respectively, for Site #11.

Instead, for the Naïve Bayes classifier, posterior probability of each class strictly depends on the proximity of the label to the true UOM. In fact, with regard to Site #1 through #10, posterior probability of the four nearest labels to *kPa absolute*, i.e., *inH_2O absolute*, *mbar absolute*, *inH_2O gauge* and *kPa gauge*, is higher than the one of the other UOMs. Similarly, for Site #11, posterior probability of all UOMs is null, with the exception of *bar gauge* and *psi gauge*.

For the considered case study, the K-NN classifier focuses the detection of the correct UOM on a limited number of labels. In fact, for Site #1 through #10, posterior probability of four classes only is not null, i.e., *mmH_2O absolute, inH_2O gauge, mbar gauge and kPa absolute* (the true UOM). As mentioned in Section 4.8.3, this outcome can be explained by considering that wrongly assumed *mmH_2O absolute* data approximately overlap data that are correctly labeled as *kPa absolute*. In addition, the *inH_2O gauge* and *mbar gauge* labels are the nearest and second-nearest classes to the true UOM.

A similar result is confirmed for Site #11, in which the *kPa absolute* label is the only incorrect UOM whose posterior probability that is not null. In fact, wrongly assumed *kPa absolute* data overlap data correctly labeled as *bar absolute*.

The Improved Nearest Neighbor achieves similar results to the K-NN classifier, especially for Site #1 through #10. Instead, as depicted in Fig. 96, posterior probability of the true UOM is equal to 100 % for Site #11.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of filtered data. Site cross-validation generally confirms results achieved by using 10 % of filtered data for training. In fact, as shown in Table A3 and Fig. 97, classification accuracy and posterior probability are always higher than 89 % and 86 %, respectively, with the exception of posterior probability of Site #11 provided by means of SVM RO, which is equal to 75 %.

Absolute and gauge pressures (twelve UOMs): site cross-validation by means of non-filtered data. Results of the site cross-validation carried out by means of non-filtered data (Table A4 and Fig. 98) confirm results provided by training the classifiers by means of 10 % of non-filtered data. In fact, classification accuracy for Site #1 through #10 is in the range from 86 % (i.e., K-NN) to 93 % (i.e., RO) (Table A4). However, the true label is assigned by SVM

RO by means of the lowest posterior probability, i.e., 70 %, while the maximum posterior probability is provided by means of the Improved Nearest Neighbor methodology, i.e., 89 % (Fig. 98).

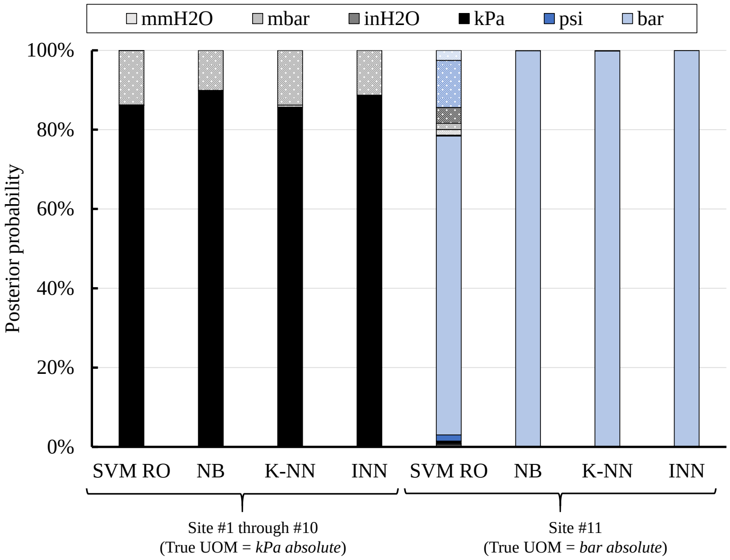


Figure 97 - Posterior probability for twelve UOMs: site cross-validation by means of filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

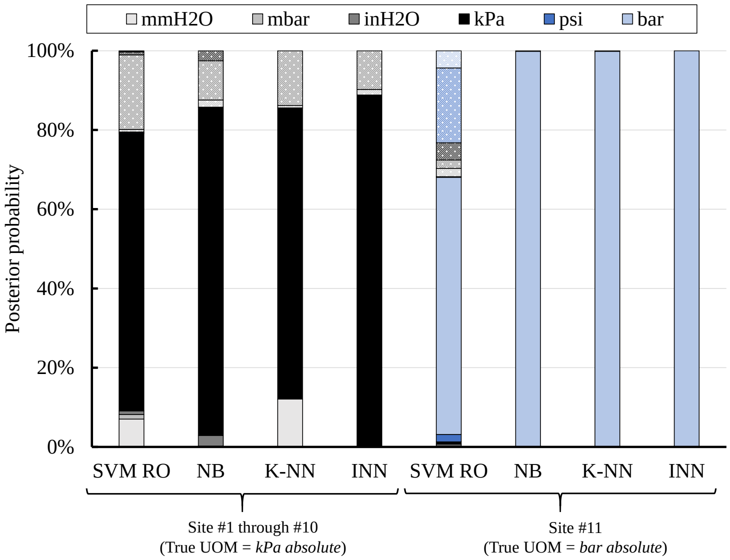


Figure 98 - Posterior probability for twelve UOMs: site cross-validation by means of non-filtered data (full bar: absolute UOMs, dotted bar: gauge UOMs)

As previously demonstrated, UMI occurred in Site #11 is generally better detected by all classifiers. This analysis confirms the general finding highlighted by means of the previous analyses. In fact, classification accuracy is only slightly affected by reliability of data for training, unlike posterior probability, which decreases by 19 %, as in SVM RO. In addition, SVM RO confirms the most scattered results in terms of posterior probability, whereas only a few classes are more probable for K-NN and Improved Nearest Neighbor classifiers.

Absolute pressures (six UOMs): training with 10 % of filtered data. If only the absolute pressures are accounted for, each classifier univocally identifies the true UOM. In fact, both posterior probability and classification accuracy are always higher than 99.5 %. For this reason, figure summarizing posterior probability and the respective confusion matrix are omitted for the sake of brevity.

Absolute pressures (six UOMs): training with 10 % of non-filtered data. As expected, both posterior probability (Fig. 99) and classification matrix (Table A4) of the true UOM are higher than the ones obtained when twelve UOMs are considered (Fig. 96 and Table A3); in fact, fewer classes are considered.

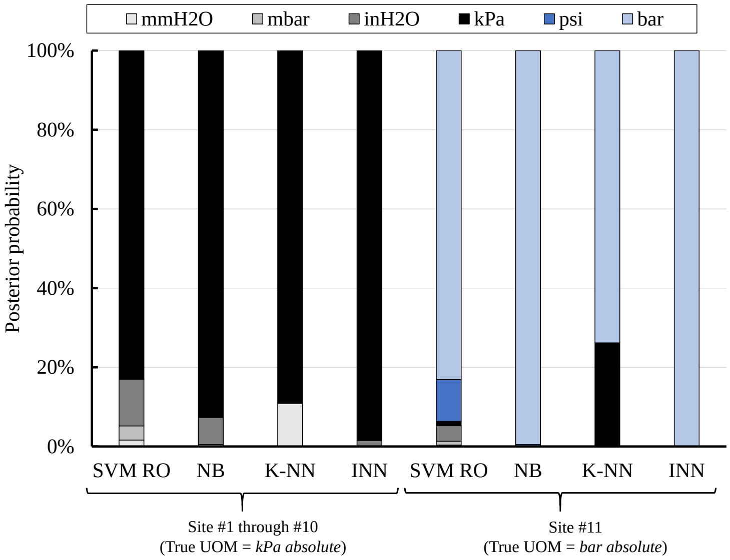


Figure 99 - Posterior probability for six absolute UOMs, training with 10 % of non-filtered data

In addition, since only absolute UOMs are investigated, the most challenging scale factor for Site #1 through #10, i.e., *kPa absolute vs.* inH_2O *gauge*, is not accounted for accordingly. As a result, the lowest classification accuracy is provided by means of the Naïve Bayes classifier, i.e., 94 % for the *kPa* label, while the classification provided by means of K-NN is characterized by the lowest posterior probability, which is equal to 74 % for Site #11.

Discussion. The detailed analyses carried out in Section 4.8 allows to identify the optimal ML classifier for detecting the UMI occurrence and subsequently providing the true UOM.

The direct comparison among the four selected classifiers reveals that Naïve Bayes classifier and the Improved Nearest Neighbor methodology generally provide the most reliable results, both in terms of classification accuracy and posterior probability. This outcome is confirmed even when non-filtered data are employed for training the classifier, by also minimizing computational time.

SVM RO may represent an effective alternative to the above-mentioned classifiers, though its classification capability can be lower than that of the optimal classifiers. In addition, computational time for both training and testing the classifier is higher than that the optimal methodologies.

Finally, in spite of the promising results achieved in this thesis, the exploitation of the K-NN classifier is discouraged for handling UMI occurrence, since its classification capability is strongly correlated to the number of selected neighbors and to the dataset composition. Thus, K-NN may not be a general approach suitable for detecting heterogeneous data, since it may require a fine tuning based on its application.

5. CONCLUSIONS

This thesis dealt with the development and validation of comprehensive methodologies aimed at evaluating the overall health state of sensors installed in gas turbines.

First, the thesis was focused on the detection and classification of sensor faults. To this aim, the I-DCIDS tool, an advanced diagnostic methodology aimed at (i) fault detection and classification and (ii) Sensor Overall Health State Analysis, was developed and validated. Unlike the previously developed DCIDS tool, the I-DCIDS tool can be employed to detect and classify seven fault categories (i.e., out of range, stuck signal, dithering, standard deviation fault, trend coherence, spike and bias), while other fault classes (e.g., drift) can also be detected and classified in one of these seven fault classes. The I-DCIDS methodology was described in detail, along with a discussion of its relevant characteristics and tuning. In addition, the I-DCIDS tool was validated by means of a sensitivity analysis performed by considering four heterogeneous and challenging field datasets affected by several fault classes, with the aim of identifying the optimal values of the configuration parameters for multi-sensor analyses (i.e., Standard Deviation Analysis, Trend Coherence Analysis, Spike Analysis and Bias Analysis).

The effectiveness of the configuration parameter values was evaluated by means of three acceptability criteria, which require detection of the faulty sensor only and the onset of the faulty event. Thanks to the sensitivity analysis, the optimal configuration parameter values were identified according to the fulfillment of the three acceptability criteria. The optimal configuration parameters were tested on an additional dataset, proving to be suitable for all types of datasets and physical quantities, even when the fault begins. This outcome makes such configuration parameters of general validity for sensor fault detection and application.

In addition, a significant amount of field data representative of different gas turbines and physical quantities (vibration, pressure, temperature, VGV position, lube oil tank level, rotational speed) were analyzed by means of the I-DCIDS tool, in order to evaluate sensor overall health state. According to results, the most critical sensor types proved to be rotational speed followed by temperature and lube oil tank level sensors. On the other side, pressure and vibration proved to be the physical quantities with the lowest rate of anomalies and noise. The anomaly classification analysis proved that most anomalies are classified as minor anomalies (isolated or serial). Instead, major anomalies were usually less frequent. Though the results are specific to the considered field data, given that the data are representative of different scenarios, it can be inferred that the analyses carried out in this thesis provide some rules of thumb about gas turbine data anomaly occurrence in field operation.

Thanks to the broad analyses carried out on experimental data, results presented in this thesis prove the diagnostic capability of the I-DCIDS tool, which is straightforward and, as demonstrated, potentially suitable for all types of measured quantities and trends over time. Moreover, thanks to its low computational time, the methodology can also be suitable for real-time applications.

Secondly, the thesis dealt with the detection of Unit of Measure Inconsistency, which is a process anomaly that makes a measured physical quantity incoherent with the assigned unit of measure.

To overcome the drawbacks of the analyses that are usually performed with the same aim, this thesis investigated the benefits of using three supervised Machine Learning techniques, i.e., Support Vector Machine (SVM), Naïve Bayes and K-Nearest Neighbors. Furthermore, a novel methodology, namely Improved Nearest Neighbor was developed with the aim to overcome K-Nearest Neighbors drawbacks. Regarding SVM, four approaches, i.e., Radial Basis Function (RBF) with One-vs-One decomposition (OVO), RBF with One-vs-All decomposition (OVA), linear OVO and linear OVA, were investigated, while the effect of the K value was analyzed for evaluating robustness and reliability of the K-Nearest Neighbor classifier.

All classifiers were tested against one experimental dataset acquired from Siemens gas turbines in operation. The dataset referred to the same pressure sensor installed on thirty Siemens gas turbines (all of the same type) running in eleven different sites.

Analyses were carried out by varying field data reliability, training and testing data and the number of classes used for training each classifier.

First, the classifiers were trained only by means of data included within acceptability thresholds. Then, both correctly labeled data and incorrectly labeled data, i.e., out of range values, were involved. In such a manner, the robustness and effectiveness of each classification technique was thoroughly evaluated even though Unit of Measure Inconsistencies were employed to train the classifiers. As expected, training data reliability generally affected the predictions.

In addition, up to twelve classes were considered, by also including challenging conversion factors for prediction.

Finally, training data were selected in two different modes. First, for each site, 10 % of data were used for training, while the remaining data were tested. Then, the classifiers were trained by excluding all data acquired from one site in turn, which was subsequently tested.

Among the considered SVM approaches, RBF OVO provided excellent results both in terms of posterior probability and classification accuracy, so that it proved to be the most effective SVM classifier for the detection of Unit of Measure Inconsistency. In fact, it proved to be more effective and robust than linear OVO even though incorrectly labeled data were also employed to train the SVM classifier.

Regarding the K-Nearest Neighbor classifier, three *K* values were accounted for. The most promising results were obtained when *K* was set equal to the square root of number of data composing the training dataset; in fact, it outperformed classification results achieved by means of the other *K* values, especially when non-filtered data were accounted for.

The detailed analyses carried out in the thesis allowed to identify the optimal machine learning classifier for detecting the UMI occurrence and subsequently providing the true UOM. In fact, the direct comparison among the four most promising classifiers, i.e., (i) SVM RO, (ii) Naïve Bayes, (iii) K-Nearest Neighbor coupled to the optimal *K* value and (iv) the Improved Nearest Neighbor revealed that the Naïve Bayes classifier and the Improved Nearest Neighbor methodology generally provided the most reliable results, both in terms of classification accuracy and posterior probability. This outcome was confirmed even when non-filtered data were employed for training the classifier, by also minimizing computational time. In fact, classification accuracy provided by means of the Naïve Bayes classifier and the Improved Neared Neighbor was always higher than 94 % and 89 %, respectively, while posterior probability was higher than 84 % and 89 %.

Instead, the classification capability of SVM RO can be lower than that of the optimal classifiers. In addition, computational time for both training and testing the classifier was higher than that the other methodologies.

Finally, the exploitation of the K-Nearest Neighbors was discouraged, since its classification capability may be strongly correlated to the number of selected neighbors and to the dataset composition and thus it may not be a general approach suitable for detecting heterogeneous data.

Future works are planned to further test the methodologies even in more challenging case studies, by considering different datasets and different sensors, a greater number of classes and a greater number of incorrectly labeled data included within the dataset. In such a manner, benefits, drawbacks and rules of thumb identified in this thesis may be further tested and validated.

ACKNOWLEDGEMENTS

I gratefully acknowledge Siemens for providing the field data used for validating the methodologies and the permission to publish the results.

I also gratefully acknowledge the Remote Diagnostics Centre Team of Siemens AG (Warwick, UK) for hosting me during my internship and for the helpful discussions and suggestions.

REFERENCES

[1] Ceschini, G. F., Gatta, N., Venturini, M., Hubauer, T., Murarasu, A., 2017, "A Comprehensive Approach for Detection, Classification and Integrated Diagnostics of Gas Turbine Sensors (DCIDS)", *J. Eng. Gas Turbines Power* **140**(3), 032402 (9 pages), doi: 10.1115/1.4037964.

[2] Roumeliotis, I., Aretakis, N., Alexiou, A., 2016, "Industrial Gas Turbine Health and Performance Assessment with Field Data", *ASME Paper GT2016-57722.*

[3] Simon, D. L., Rinehart, A. W., 2016, "Sensor Selection for Aircraft Engine Performance Estimation and Gas Path Fault Diagnostics", J Eng Gas Turb Power**138**(7), 071201 (Feb 17, 2016) (11 pages)

[4] Venturini, M., Puggina, N., 2012, "Prediction Reliability of a Statistical Methodology for Gas Turbine Prognostics", *ASME J. Eng. Gas Turbines Power*, **134**(10), 101601 (9 pages).

[5] Venturini, M., Therkorn, D., 2013, "Application of a Statistical Methodology for Gas Turbine Degradation Prognostics to Alstom Field Data", *J Eng Gas Turb Power***135**(9), 091603 (10 pages), doi:10.1115/1.4024952.

[6] Hanachi, H., Liu, J., Banerjee, A., Chen, Y., 2016, "Prediction Of Compressor Fouling Rate Under Time Varying Operating Conditions", *ASME Paper GT2016-56242.*

[7] Sarkar, S., Jin, X., and Ray, A., 2011, "Data-Driven Fault Detection in Aircraft Engines With Noisy Sensor Measurements", *J. Eng. Gas Turbines Power*, **133**(8), p. 081602.

[8] Bauer, M., Friedrichs, J., Wulff, D., Werner-Spatz, C., 2017, "Development and Validation of an On-Wing Engine Thrust Measurement System", ASME Paper GT2017-63277.

[9] Sridhar, V., Chana, K. S., Singh, D., 2017, "Computational and Experimental Study of a Platinum Thin-Film Based Oil Condition and Contamination Sensor", ASME Paper GT2017-63788.

[10] Schwartz, B., Fittro, R., Knospe, C., 2017, “Understanding the Effect of Systematic Errors on the Accuracy of Experimental Measurements of Fluid-Film Bearing Dynamic Coefficients”, ASME Paper GT2017-64665.

[11] Sridhar, V., Chana, K. S., 2017, “Development of a Combined Eddy Current and Pressure Sensor for Gas Turbine Blade Health Monitoring”, ASME Paper GT2017-63807.

[12] Sridhar, V., Chana, K. S., 2017, “Tip Clearance Measurements on an Engine High Pressure Turbine Using an Eddy Current Sensor”, ASME Paper GT2017-63803.

[13] Nakajima, T., Segawa, K., Kitahara, H., Seo, A., Yamashita, Y., Kudo, T., 2017, “Experimental Investigation of the Grouped Blade Vibration for Steam Turbine by Non-Contact Sensors”, ASME Paper GT2017-64047.

[14] Dewallef, P. and Borguet, S., 2013, "A Methodology to Improve the Robustness of Gas Turbine Engine Performance Monitoring Against Sensor Faults", *J. Eng. Gas Turbines Power*, **135**(5), p. 051601.

[15] Van Paridon, A., Bacic, M., Ireland, P. T., 2016, “Kalman Filter Development For Real Time Proper Orthogonal Decomposition Disc Temperature Model”, *ASME Paper GT2016-56330.*

[16] Hurst, A. M., Carter, S., Firth, D., Szary, A., Van De Weert, J., 2015, “Real-Time, Advanced Electrical Filtering for Pressure Transducer Frequency Response Correction”, *ASME Paper GT2015-42895.*

[17] Gutierrez, L. A., Pezzini, P., Tucker, D., Banta, L., 2014, “Smoothing Techniques For Real-Time Turbine Speed Sensors”, *ASME Paper GT2014-25407.*

[18] Yi, T.H., Huang, H.B., Li, H.N., “Development of sensor validation methodologies for structural health monitoring: A comprehensive review”, *Measurement* 109 (2017) 200-214.

[19] Prasad, A., Edward, J. B., Ravi, K., 2018. “A review on fault classification methodologies in power transmission systems: Part-I”, *Journal of Electrical Systems and Information Technology*, **5** (2018) 48-60.

[20] Prasad, A., Edward, J. B., Ravi, K., 2018. “A review on fault classification methodologies in power transmission systems: Part-II”, *Journal of Electrical Systems and Information Technology*, **5** (2018) 61-67.

[21] Hanachi, H., Liu, J., Mechefske, C., 2018, “Multi-mode diagnosis of a gas turbine engine using an adaptive neuro-fuzzy system”, *Chinese Journal of Aeronautics* 31(1): 1-9.

[22] Hanachi, H., Liu, J., Kim, I.Y., Mechefske, C.K, “Hybrid sequential fault estimation for multi-mode diagnosis of gas turbine engines”, *Mechanical Systems and Signal Processing* 115 (2019) 255-268.

[23] Tayarani-Bathaie, S.S,. Khorasani, K., “Fault detection and isolation of gas turbine engines using a bank of neural networks”, *Journal of Process Control* 36 (2015) 22-41.

[24] Navi, M., Meskin, N., Davoodi, M., “Sensor fault detection and isolation of an industrial gas turbine using partial adaptive KPCA”, *Journal of Process Control* 64 (2018) 37-48.

[25] Zhang, Y., Bingham, C., Yang, Z., Ling, B.W.K., Gallimore, M., “Machine fault detection by signal denoising-with application to industrial gas turbines”, *Measurement* 58 (2014) 230-240.

[26] Bakdi, A., Kouadri, A., Bensmail, A., 2017, “Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme”, *Control Engineering Practice* **66** (2017) 64-75.

[27] Portnoy, I., Melendez, K., Pinzon, H., Sanjuan, M., 2016, “An improved weighted recursive PCA algorithm for adaptive fault detection”, *Control Engineering Practice* **50** (2016) 69-83.

[28] Stork, C.L., Kowalski, B.R., Distinguishing between process upsets and sensor malfunctions using sensor redundancy, Chemometrics and Intelligent Laboratory Systems 46 1999 117–131

[29] Verma, N., Singh, D., Data Redundancy Implications in Wireless Sensor Networks, Procedia Computer Science 132 (2018) 1210–1217, International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

[30] Zhao, Z., Liu, Z., Sun, Y., Liu, J., WOS-ELM-Based Double Redundancy Fault Diagnosis and Reconstruction for Aeroengine Sensor, Journal of Control Science and Engineering Volume 2017, Article ID 1982879, 14 pages https://doi.org/10.1155/2017/1982879

[31] Mahapatro, A., Panda, A.K., Choice of Detection Parameters on Fault Detection in Wireless Sensor Networks: A Multiobjective Optimization Approach, Wireless Pers Commun (2014) 78:649–669 DOI 10.1007/s11277-014-1776-1

[32] Mattera, C.G., Quevedo, J., Escobet, T., Shaker, H.R., Jradi, M., A Method for Fault Detection and Diagnostics in Ventilation Units Using Virtual Sensors, Sensors 2018, 18, 3931; doi:10.3390/s18113931

[33] Stetter, R., Witczak, M., Pazera, M., Virtual Diagnostic Sensors Design for an Automated Guided Vehicle, Applied Sciences 2018, 8, 702; doi:10.3390/app8050702

[34] Capriglione, D., Pietrosanto, A., Online Fault Detection of Rear Stroke Suspension Sensor in Motorcycle, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, Vol. 68, no. 5, May 2019

[35] Losi, E., Venturini, M., Manservigi, L., Ceschini, G.F., Bechini, G., 2019, "Anomaly Detection in Gas Turbine Time Series by means of Bayesian Hierarchical Models", J. Eng. Gas Turbines Power, Vol. **141**, 111019 (9 pages), doi: 10.1115/1.4044781.

[36] Stoumpos, S., Theotokatos, G., A novel methodology for marine dual fuel engines sensors diagnostics and health management, International J of Engine Research 1-21, DOI: 10.1177/1468087421998635

[37] Darvishi, H., Ciuonzo, D., Eide, E.R., Rossi, P.S., Sensor-Fault Detection, Isolation and Accommodation for Digital Twins via Modular Data-Driven Architecture, IEEE Sensors Journal 2021, Vol. 21, Issue 4, 9216114, Pages 4827-4838, DOI: 10.1109/JSEN.2020.3029459

[38] Pourbabaee, B., Meskin, N., Khorasani, K., "Robust sensor fault detection and isolation of gas turbine engines subjected to time-varying parameter uncertainties", *Mechanical Systems and Signal Processing* 76-77 (2016) 136-156.

[39] Rahme, S., Meskin, N., "Adaptive sliding mode observer for sensor fault diagnosis of an industrial gas turbine", *Control Engineering Practice* 38 (2015) 57-74.

[40] Naderi, E., Khorasani, K., "Data-driven fault detection, isolation and estimation of aircraft gas turbine engine actuator and sensors", *Mechanical Systems and Signal Processing* 100 (2018) 415-438.

[41] Yang, Z., Ling, B.W.K., Bingham, C., "Fault detection and signal reconstruction for increasing operational availability of industrial gas turbines", *Measurement* 46 (2013) 1938-1946.

[42] Jombo, G., Zhang, Y., Griffiths, J.D., 2018, "Automated gas turbine sensor fault diagnostics", *Proceeding of ASME Turbo Expo 2018*, June 11-15, Oslo, Norway.

[43] Li, Y.G., 2009, "Gas Turbine Performance and Health Status Estimation Using Adaptive Gas Path Analysis", *Journal of Engineering for Gas Turbines and Power* 132(4) 041701.

[44] Ying, Y., Cao, Y., Li, S., Li, J, Guo, J., 2016, "Study on gas turbine engine fault diagnostic approach with a hybrid of gray relation theory and gas-path analysis", *Advances in Mechanical Engineering* 8(1) 1-14, doi: 10.1177/1687814015627769

[45] da Silva, P.R.N, Gabbar, H.A., Vieira Junior, P., da Costa Junior, C.T., 2018, "A new methodology for multiple incipient fault diagnosis in transmission lines using QTA and Naïve Bayes classifier", Electrical Power and Energy Systems 103 (2018) 326-346.

[46] Sanchez, J.S., Pla, F., Ferri, F.J., 1997, "On the use of neighbourhood-based non-parametric classifiers", Pattern Recognition Letters 18 1997 1179–1186.

[47] Kullaa, J., "Detection, identification, and quantification of sensor fault in a sensor network", *Mechanical Systems and Signal Processing* 40 (2013) 208-221.

[48] Ceschini, G. F., Gatta, N., Venturini, M., Hubauer, T., Murarasu, A., 2018, "Optimization of Statistical Methodologies for Anomaly Detection in Gas Turbine Dynamic Time Series", *J. Eng. Gas Turbines Power* 140(3), 032401 (10 pages), doi: 10.1115/1.4037963.

[49] Ceschini, G. F., Gatta, N., Venturini, M., Hubauer, T., Murarasu, A., 2018, "Resistant Statistical Methodologies for Anomaly Detection in Gas Turbine Dynamic Time Series: Development and Field Validation", *J. Eng. Gas Turbines Power* 140(5), 052401 (11 pages), doi: 10.1115/1.4038155.

[50] Tang., L., Volponi, A.J., "Intelligent Reasoning for Gas Turbine Fault Isolation and Ambiguity Resolution", *Proceedings of ASME Turbo Expo 2018, GT2018-75540, June 11-15, 2018, Oslo, Norway*.

[51] Plotts, K., Diatzikis, E., "A Survey of New Technologies Used by Siemens Energy for the Monitoring and Diagnosis of a Global Fleet of Power Generation Systems", *Proceedings of ASME Turbo Expo 2009, GT2009-59967, June 8-12, 2009, Orlando, Florida, USA*.

[52] Chen, F., Roşu, G., Venkatesan, R.P., "Rule-based Analysis of Dimensional Safety", *RTA'03. LNCS 2706*, pp.197 – 207, 2003.

[53] www.cbc.ca

[54] d'Amorim, M., Hills, M., Chen, F., Roşu, G., "Automatic and Precise Dimensional Analysis", *Technical Report* UIUCDCS-R-2005-2668, December 2005.

[55] Chizeck, H.J., Butterworth, E., Bassingthwaighte, J.B., "Error Detection and Unit Conversion: Automated Unit Balancing in Model Interface Systems", *IEEE Eng Med Biol Mag. 2009*; 28(3): 50-58.

[56] Wu, H., Chen, J., Zhang, H., "Study on the Smart Diagnosis of Gas Turbine with Distributed Computing Technique", *Proceedings of ASME Turbo Expo 2019, GT2019-91950, June 17-21, 2019, Phoenix, Arizona, US*.

[57] Jaw, L.C., Lee, Y.J., "Engine Diagnostics in the Eyes of Machine Learning", *Proceedings of ASME Turbo Expo 2014, GT2014-27088, June 16-20, 2014, Düsseldorf, Germany*.

[58] Luo, Y.F., Zeng, J., 2004, "Gas Turbine Fault Diagnosis Based on a Support Vector Machine", *Journal of Engineering for Thermal Energy and Power,* 19(4), pp. 354-357.

[59] Elangovan, M., Sugumaran, V., Ramachandran, K.I., Ravikumar, S., "Effect of SVM kernel functions on classification of vibration signals of a single point cutting tool", *Expert Systems with Applications* 38 (2011) 15202-15207.

[60] Shi, M., Zhao, R., Wu, Y., He, T., 2021, "Fault diagnosis of rotor based on Local-Global Balanced Orthogonal Discriminant Projection", Measurement 168 (2021) 108320.

[61] Niazi, K.A.K., Akhtar, W., Khan, H.A., Yang, Y., Athar, S., 2019, "Hotspot diagnosis for solar photovoltaic modules using a Naive Bayes classifier", Solar Energy 190 (2019) 34-43.

[62] Zhang, Z., Han, H., Cui, X., Fan, Y., 2020, "Novel application of multi-model ensemble learning for fault diagnosis in refrigeration systems", Applied Thermal Engineering164 (2020) 114516.

[63] Brooks, K.S., Bauer, M., "Sensor validation and reconstruction: Experiences with commercial technology", *Control Engineering Practice* 77 (2018) 28-40.

[64] Chandola, V., Banerejee, A., Kumar, V., "Anomaly Detection: A Survey"

[65] Hyder, A.K., Shahbazian, E., Waltz, E., "Multisensor Fusion", NATO Science Series, II. Mathematics, Physics and Chemistry – Vol. 70, DOI 10.1007/978-94-010-0556-2.

[66] Muhammed, T., Shaikh, R. A., "An analysis of fault detection strategies in wireless sensor networks", Journal of Network and Computer Applications, 78 (2017) 267-287.

[67] Zhong, G. X., Yang, G. H., "Fault detection for discrete-time switched systems with sensor stuck faults and servo inputs", ISA Transactions, 58 (2015) 196-205.

[68] Iannelli, L., Johansson, K. H., Jönsson, U. T., Vasca, F., 2006, "Averaging of nonsmooth systems using dither", Automatica, 42 (2006) 669-676.

[69] Lo, C., Lynch, J.P., Liu, M., "Distributed model-based nonlinear sensor fault diagnosis in wireless sensor networks", Mechanical Systems and Signal Processing 66-67 (2017) 470-484.

[70] Pytharouli, S., Chaikalis, S., Stiros, S. C., "Uncertainty and bias in electronic tide-gauge records: Evidence from collocated sensors", Measurement, 125 (2018) 496-508.

[71] Manservigi, L., Venturini, M., Ceschini, G. F., Bechini, G., Losi, E., 2019, "A General Diagnostic Methodology for Sensor Fault Detection, Classification and Overall Health State Assessment", ASME paper GT2019-90055.

[72] Cejnek, M., Bukovsky, I., "Concept drift robust adaptive novelty detection for data streams", Neurocomputing, 309 (2018) 46-53.

[73] Lanzante, J., 1996, "Resistant, Robust and Non-Parametric Techniques for the Analysis of Climate Data: Theory and Examples, Including Applications to Historical Radiosonde Station Data", International Journal of Climatology, 16(11), pp. 1197-1226.

[74] Kafadar, K., 1983, "The Efficiency of the Biweight as a Robust Estimator of Location", J. Res. Natl. Bur. Stan., 88(2), p. 105.

[75] Rousseeuw, P. and Croux, C., 1993, "Alternatives to the Median Absolute Deviation", Journal of the American Statistical Association, 88(424), p. 1273.

[76] Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L., 2013, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median", Journal of Experimental Social Psychology, 49(4), pp. 764-766.

[77] Yahaya, S., Othman, A. and Keselman, H., 2004, “Testing the Equality of Location Parameters for Skewed Distributions Using S1 with High Breakdown Robust Scale Estirnators”, Statistics for Industry and Technology, pp.319-328.

[78] Hoaglin, D., Mosteller, F., and Tukey, J., 1983, Understanding robust and exploratory data analysis, Wiley, New York.

[79] Pinelli, M., Venturini, M., and Burgio, M., 2003, "Statistical Methodologies for Reliability Assessment of Gas Turbine Measurements ", Proceedings of the Turbine Technical Conference and Exposition, (ASME Paper GT2003-38407), pp. 787-793.

[80] Gomez, J., 2011, Kalman filtering, Nova Science Publishers, Hauppauge, N.Y.

[81] Losi, E., Venturini, M., Manservigi, L., Ceschini, G.F., Bechini, G., Cota, G., Riguzzi, F., “Data selection and trasformation to predict gas turbine trip by means of machine learning models”, Proceedings of ASME Turbo Expo 2021, GT2021 June 7-11, 2021, Virtual, Online, GT2021-58914 (submitted)

[82] Zaidan, M.A., Relan, R., Mills, A.R., Harrison, R.F., 2015, “Prognostics of gas turbine: an integrated approach”, Expert Systems with Applications 42 (2015) 8472-8483.

[83] Losi, E., Venturini, M., Manservigi, L., 2019, “Gas Turbine Health State Prognostics by Means of Bayesian Hierarchical Models”, ASME J. Eng. Gas Turbines Power, Vol. 141, 111018 (17 pages), doi: 10.1115/1.4044689.

[84] Kumar, Y., Sahoo, G., “Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA”, I.J. Information Technology and Computer Science, 2012, 7, 43-49, DOI: 10.5815/ijitcs.2012.07.06

[85] Durga Bhavani, D. Vasavi, A., Keshava, P.T., 2016, “Machine Learning: A Critical Review of Classification Techniques”, International Journal of Advanced Research in Computer and Communication Engineering 2319-5940, DOI 10.17148/IJARCCE.

[86] Zhao, Y.P., Wang, J.J., Li, X.Y., Peng, G.J., Yang, Z., “Extended least squares support vector

machine with applications to fault diagnosis of aircraft engine", ISA Transactions (2019), doi:10.1016/j.isatra.2019.08.036.

[87] Leong, W.C., Kelani, R.O., Ahmad, Z., "Prediction of air pollution index (API) using support vector machine (SVM)", Journal of Environmental Chemical Engineering (2019), doi: 10.1016/j.jece.2019.103208

[88] Subasi, O., Di, S., Bautista-Gomez, L., Balaprakash, P., Unsal, O., Labarta, J., Cristal, A., Krishnamoorthy, S., Cappello, F., "Exploring the capabilities of support vector machines in detecting silent data corruptions", Sustainable Computing: Informatics and Systems 19 (2018) 277-290.

[89] Mountrakis, G., Im, J., Ogole, C., "Support vector machines in remote sensing: A review", ISPRS Journal of Photogrammetry and Remote Sensing 66 (2011) 247-259.

[90] Zhou, D., Zhang, H., Weng, S., 2015, "A new gas path fault diagnostic method of gas turbine based on support vector machine", Journal of Engineering for Gas Turbines and Power, 137(10), p. 102605.

[91] Krawczyk, B., Galar, M., Woźniak, M., Bustince, H., Herrera, F., "Dynamic ensemble selection for multi-class classification with one-class classifiers", Pattern Recognition 83 (2018) 34-51.

[92] Milgram, J., Cheriet, M., Sabourin, R., "“One Against One” or “One Against All”: Which One is Better for Handwriting Recognition with SVMs?", Tenth International Workshop on Frontiers in Handwriting Recognition, Université de Rennes 1, Oct 2006, La Baule (France). inria00103955.

[93] Zhou, L., Wang, Q., Fujita, H., "One versus one multi-class classification fusion using optimizing decision directed acyclic graph for predicting listing status of companies", Information Fusion, 36 (2017) 80-89.

[94] Liu, Y., Bi, J.W., Fan, Z.P., "A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm", Information Sciences 394-395 (2017) 38-52.

[95] López, J., Maldonado, S., "Multi-class second-order cone programming support vector machines", Information Sciences 330 (2016) 328-341.

[96] Zhang, Z., Krawczyk, B., Garcìa, S., Rosales-Pérez, A., "Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data", Knowledge-Based Systems 106 (2016) 251-263.

[97] Petschke, D., Staab, T.E.M., 2019, “A supervised machine learning approach using naive Gaussian Bayes classification for shape-sensitive detector pulse discrimination in positron annihilation lifetime spectroscopy (PALS)”, Nuclear Inst. and Methods in Physics Research, A 947 (2019) 162742.

[98] Aralikatti, S.S, Ravikumar, K.N., Kumar, H., Nayaka, H.S., Sugumaran, V., 2020, “Comparative Study on Tool Fault Diagnosis Methods Using Vibration Signals and Cutting Force Signals by Machine Learning Technique”, Structural Durability & Health Monitoring 14(2), pp. 127 – 145.

[99] Aker, E., Othman, M.L., Veerasamy, V., Aris, I.b., Wahab, N.I.A., Hizam, H., 2020, “Fault Detection and Classification of Shunt Compensated Transmission Line Using Discrete Wavelet Transform and Naive Bayes Classifier”, Energies 2020, 13, 243; doi:10.3390/en13010243.

[100] Cheng, D., Zhang, S., Deng, Z., Zhu, Y., Zong, M., “kNN Algorithm with Data-Driven k Value”, X. Luo, J.X. Yu, and Z. Li (Eds.): ADMA 2014, LNAI 8933, pp. 499–512, 2014.© Springer International Publishing Switzerland 2014

[101] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844

[102] Tran, T.M., Thi Le, X.M., Nguyen, H.T., Huynh, V.N., 2019, “A novel non-parametric method for time series classification based on *k*-Nearest Neighbors and Dynamic Time Warping Barycenter Averaging”, Engineering Applications of Artificial Intelligence 78 (2019) 173–185.

[103] Chaudhuri, B.B., 1996, “A new definition of neighborhood of a pointin multi-dimensional space”, Pattern Recognition Letters 17 (1996) 11-17.

[104] Aslinezhad, M., Hejazi, M.A., 2020, “Turbine blade tip clearance determination using microwave measurement and k-nearest neighbour classifier”, Measurement 151 (2020) 107142.

[105] Moshrefzadeh, A., 2021, “Condition monitoring and intelligent diagnosis of rolling element bearings under constant/variable load and speed conditions”, Mechanical Systems and Signal Processing 149 (2021) 107153.

[106] Robert P.W. Duin, David M.J. Tax, “Classifier Conditional Posterior Probabilities”, Lecture Notes in Computer Science 1451(1451):611-619, DOI: 10.1007/BFb0033285.

[107] Manservigi, L., Murray, D., Artal de la Iglesia, J., Ceschini, G.F., Bechini, G., Losi, E., Venturini, M., “Detection of Unit of Measure Inconsistency by means of a Machine Learning Model”, Proceedings of ASME Turbo Expo 2020, June 22-26, London, England. ASME paper GT2020-16094.

APPENDIX

Table A1 - Confusion matrix for twelve UOMs, training with 10 % of filtered data

	Site #1 through #10 (true UOM = *kPa absolute*)				Site #11 (true UOM = *bar absolute*)			
	SVM RO	NB	K-NN	INN	SVM RO	NB	K-NN	INN
mmH$_2$O absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mmH$_2$O gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
inH$_2$O absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar gauge	0.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
kPa absolute	95.5 %	98.6 %	77.3 %	91.3 %	0.0 %	0.0 %	0.0 %	0.0 %
inH$_2$O gauge	4.0 %	3.2 %	22.7 %	8.7 %	0.0 %	0.0 %	0.0 %	0.0 %
kPa gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.3 %	0.0 %	0.3 %	0.0 %
bar absolute	0.0 %	0.0 %	0.0 %	0.0 %	99.7 %	100.0 %	99.7 %	100.0 %
bar gauge	0.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

Table A2 - Confusion matrix for twelve UOMs, training with 10 % of non-filtered data

	Site #1 through #10 (true UOM = *kPa absolute*)				Site #11 (true UOM = *bar absolute*)			
	SVM RO	NB	K-NN	INN	SVM RO	NB	K-NN	INN
mmH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mmH_2O *gauge*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar gauge	0.0 %	0.0 %	0.0 %	0.9 %	0.3 %	0.0 %	0.0 %	0.0 %
kPa absolute	97.4 %	94.1 %	76.2 %	88.8 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *gauge*	2.1 %	5.9 %	23.8 %	10.2 %	0.0 %	0.0 %	0.0 %	0.0 %
kPa gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi absolute	0.1 %	0.0 %	0.0 %	0.0 %	0.3 %	0.6 %	0.3 %	0.3 %
psi gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.3 %	0.0 %	0.3 %	0.0 %
bar absolute	0.0 %	0.0 %	0.0 %	0.0 %	99.1 %	99.4 %	99.7 %	99.7 %
bar gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

Table A3 - Confusion matrix for twelve UOMs, training with 10 % of non-filtered data

	Site #1 through #10 (true UOM = *kPa absolute*)				Site #11 (true UOM = *bar absolute*)			
	SVM RO	NB	K-NN	INN	SVM RO	NB	K-NN	INN
mmH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mmH_2O *gauge*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar gauge	0.0 %	0.0 %	0.0 %	0.9 %	0.3 %	0.0 %	0.0 %	0.0 %
kPa absolute	97.4 %	94.1 %	76.2 %	88.8 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *gauge*	2.1 %	5.9 %	23.8 %	10.2 %	0.0 %	0.0 %	0.0 %	0.0 %
kPa gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi absolute	0.1 %	0.0 %	0.0 %	0.0 %	0.3 %	0.6 %	0.3 %	0.3 %
psi gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.3 %	0.0 %	0.3 %	0.0 %
bar absolute	0.0 %	0.0 %	0.0 %	0.0 %	99.1 %	99.4 %	99.7 %	99.7 %
bar gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

Table A4 - Confusion matrix for twelve UOMs: site cross-validation by means of filtered data

	Site #1 through #10 (true UOM = *kPa absolute*)				Site #11 (true UOM = *bar absolute*)			
	SVM RO	NB	K-NN	INN	SVM RO	NB	K-NN	INN
mmH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mmH_2O *gauge*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar gauge	0.0 %	0.0 %	0.0 %	0.2 %	0. %	0.0 %	0.0 %	0.0 %
kPa absolute	94.2 %	94.4 %	89.5 %	88.5 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *gauge*	5.8 %	5.5 %	10.5 %	11.3 %	0.0 %	0.0 %	0.0 %	0.0 %
kPa gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.1 %	0.0 %	0.0 %
bar absolute	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	99.9 %	100.0 %	100.0 %
bar gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

Table A5 - Confusion matrix for twelve UOMs: site cross-validation by means of non-filtered data

	Site #1 through #10 (true UOM = *kPa absolute*)				Site #11 (true UOM = *bar absolute*)			
	SVM RO	NB	K-NN	INN	SVM RO	NB	K-NN	INN
mmH_2O *absolute*	0.0 %	0.0 %	2.6 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mmH_2O *gauge*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar gauge	0.0 %	0.0 %	0.0 %	1.4 %	0. %	0.0 %	0.0 %	0.0 %
kPa absolute	92.8 %	92.2 %	86.4 %	88.8 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *gauge*	7.2 %	7.8 %	11.0 %	9.8 %	0.0 %	0.0 %	0.0 %	0.0 %
kPa gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
bar absolute	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	100.0 %	100.0 %	100.0 %
bar gauge	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

Table A6 – Confusion matrix for six UOMs, training with 10 % of non-filtered data

	Site #1 through #10 (true UOM = *kPa absolute*)				Site #11 (true UOM = *bar absolute*)			
	SVM RO	NB	K-NN	INN	SVM RO	NB	K-NN	INN
mmH_2O *absolute*	0.0 %	0.0 %	2.6 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mmH_2O *gauge*	N/A				N/A			
mbar absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *absolute*	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
mbar gauge	N/A				N/A			
kPa absolute	92.8 %	92.2 %	86.4 %	88.8 %	0.0 %	0.0 %	0.0 %	0.0 %
inH_2O *gauge*	N/A				N/A			
kPa gauge	N/A				N/A			
psi absolute	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
psi gauge	N/A				N/A			
bar absolute	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %	100.0 %	100.0 %	100.0 %
bar gauge	N/A				N/A			