Data-Driven Fault Detection in a Thermocouple Network Using Neighboring Redundancy, XGBoost Classifier, and Up–Down Counter

DOI:
10.1109/JSEN.2023.3342614

Citation for published version (APA):
Velandia cárdenas, D. A., Pulgarín, E. J. L., & Sofrony, J. I. (2024). Data-Driven Fault Detection in a Thermocouple Network Using Neighboring Redundancy, XGBoost Classifier, and Up–Down Counter. *IEEE Sensors Journal*, *24*(4), 5215-5223. https://doi.org/10.1109/JSEN.2023.3342614

Published in:
IEEE Sensors Journal

Data-driven fault detection in a thermocouple network using neighbouring redundancy, XGBoost classifier and up-down counter

Diego A. Velandia Cárdenas, Erwin Jose López Pulgarín, *Member, IEEE* and Jorge Iván Sofrony, *Member, IEEE*

***Abstract*—Fault Detection and Isolation (FDI) is of great interest for the control community since it can drive improved performance in a system by allowing predictive maintenance/repairing and catering for improved operational safety. Fault Detection and Isolation in large-scale smelting furnaces presents several challenges, as it requires the understanding of complex thermal and chemical reactions occurring inside the structure. Furthermore, the impossibility of having full operational information about the process makes the use of model-based methods very complex or unfeasible. This paper introduces a methodology to develop a Data-Driven FDI system for the detection of incipient and intermittent failures in a network made out of 322 thermocouples located on the shell of the furnace. Statistical metrics over Fault Counter Time Windows (FTCW) were used to identify different types of sensor failures, which led to establishing a baseline of known failure events and to create a dataset to train the Machine Learning (ML) classification models. A data-driven approach was proposed based on the sensors physical (neighbouring) redundancy, which led to some type of physical redundancy. A post-processing stage was used to stabilize the model's response in time, determining that the proposed FDI system successfully detects faults whilst reducing reported false negatives.**

***Index Terms*—FDI, machine learning, parameter variation, redundancy, sensors network, thermocouple, up-down counter, XGBoost.**

I. Introduction

Electric arc furnaces (EAC) are used for smelting ores into refined metals by driving electricity from the electrodes in its ceiling to the conducting bottom, through a bed made of pre-treated ore called calcine. The molten materials inside the furnace separate by density difference as they heat, allowing their extraction. Unlike blast furnaces, which are commonly powered by coke or coal, EACs power can be obtained from several sources according to availability and price.

Operating a smelting furnace involves constant control and monitoring of the process. It may be considered as a safety critical system, meaning that the lack of a strict safety monitoring scheme may entail catastrophic events [23]. Monitoring activities include tracking the temperature around the entire furnace's exterior wall, which results in a complex and costly operation due to the large number of required sensors embedded in the furnace's wall and the extreme environmental conditions.

The sensors operate in extreme conditions, increasing the risk of failure, which in turn can put at risk the health of the furnace. The temperature of the middle wall is a critical variable to monitor when assessing structural safety, as this is the zone where the molten metal separates from the slag [17]. Detecting sensor failures helps to increase the reliability of the health monitoring system.

FDI techniques have grown in complexity over the years as they accommodate more complex systems and processes. FDI systems should trigger actions that improve system performance, reduce maintenance times, and help to assess operational risk. Previous work [8] [20] defines four stages of FDI: detection, isolation, estimation, and adjustment. In the FDI context, a sensor failure is an undesired change in a measurement behaviour, leading to the appearance of precision, stability, and reliability issues. Sensor failures are due to a wide range of causes including regular wearing, misuse, environmental conditions, among others.

Failures can be classified as abrupt or incipient, (see Zhang et al. [25] and Samara et al. [18] for example). Abrupt failures are sudden alterations in the sensor's behaviour, turning it inoperable until corresponding adjustments are performed. This kind of failure can be identified in a timeline as, in most cases, a frozen signal or a noticeable (abrupt) measurement change. The accidental disconnection of a sensor is an example of abrupt failures requiring simple adjustments; a molten thermocouple, on the other hand, would require a total sensor replacement and can imply major efforts or even a total operation suspension.

Incipient failures are gradual or slight deviations of the sensor's response compared to its expected values, not disabling the sensor immediately but making its measurements less reliable compared to a sensor with no affectations. Incipient failures result harder to identify, as the change in the sensor's

D. Velandia and J. Sofrony were with the Department of Electrical and Electronic Engineering, Universidad Nacional de Colombia - Sede Bogotá, Colombia.

E. Lopez is with the Department of Electrical and Electronic Engineering (EEE), University of Manchester, Manchester, UK.

response is not easily detected by visual inspection or its rate of appearance is not predictable.

One form of FDI is physical redundancy, where additional sensors that measure the same variable are installed and used to determine whether there is a failure. This approach has considerable setbacks as installing more sensors implies larger investments (economical, logistic, space-wise). Given the sensor network's configuration depicted in Figures 2 and 3, an approximation to physical redundancy is proposed.

Analytical sensor redundancy, in contrast, relies on analytic knowledge of the process, together with sensor and process data, to establish if a sensor is experiencing a failure. Analytical redundancy can be performed following three approaches: model-based, knowledge-based and data-driven.

A model-based approach requires an analytical model that relates the system's inputs and outputs to compute the expected outputs and compare them with the measured values. The difference between measured and expected values are known as residuals. Obtaining a detailed model of an electric arc furnace is considerably complex as it requires highly detailed knowledge of the furnace's current state, making the model-based approach not suitable for the present study case [6].

Knowledge-based approaches, on the other hand, do not require an analytic model of the system; instead they use the historical records of known failures and combine them through a diffuse inference engine with a rule base extracted from expert knowledge. These systems improve their performance as they operate, as the knowledge base is increased on each new successful event detection [3]. These approaches are recommended when a solid base of knowledge exists and there is a guarantee that an expert group will continuously review and adjust the system to achieve the best performance possible.

Data-driven approaches are particularly suitable when dealing with complex systems whose models are difficult to construct or expert knowledge is not continuously available for feedback. Data-driven models can be trained once and operate autonomously until the system changes (degrades) significantly.

Data-driven models can be classified as unsupervised or supervised according to the way they work. Unsupervised models use dimensionality reduction of high dimensional data to define metrics and group up the observations into different clusters [1] aiming for the maximum similarity among elements within a same cluster.

Supervised data-driven techniques assume a statistical model that modifies some of its internal parameters based on a sample of the data called *training set*. The input to supervised models can be the original data as it was produced, or a set of metrics derived from the original data, better known as features.

II. Proposed FDI methodology

This work proposes an FDI methodology to deal with both incipient and abrupt sensor failures in a smelting furnace's thermocouple sensor network. Sensor measurements were gathered from a process with little operational data and no baseline knowledge of existing faults, with a resulting dataset which was unbalanced due to the significantly fewer failure events compared to its normal operation.

The contributions of this paper are twofold, Figure 1 features different steps involved in the development of the contributions. Firstly, we introduced a methodology to create a knowledge baseline from unlabelled and chronologically ordered sensor data. Subsection IV-A describes how abrupt failures are identified and labelled on each sensor's signal by applying heuristic rules over rolling time windows; in parallel, detailed observation is carried out on rolling time windows of sensors neighbourhoods' data to identify incipient failures from sensor measurements alone, this process is explained in subsection IV-B. The combination of knowledge related to abrupt and incipient faults leads to the determination of a baseline, as introduced in subsection IV-C.

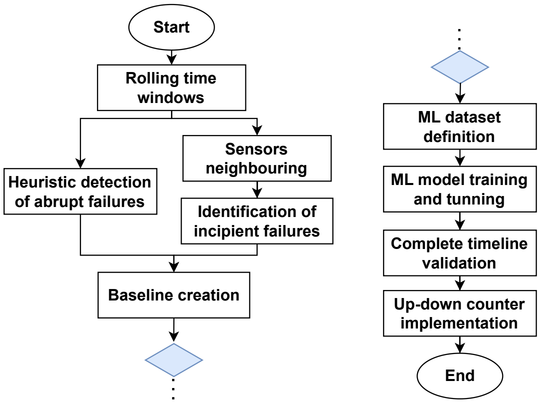
Start
Rolling time windows
Heuristic detection of abrupt failures
Sensors neighbouring
Identification of incipient failures
Baseline creation
ML dataset definition
ML model training and tunning
Complete timeline validation
Up-down counter implementation
End

Fig. 1. FDI model development process

Secondly, we introduce a data-driven FDI model that deals with highly unbalanced training data as it is expected under real operating conditions. In subsection IV-D training and validations datasets were derived from the baseline to train, tune and compare various machine learning models, revealing a shared setback due to the strong subsampling process forced by the data unbalance, an up-down counter like described in subsection IV-E was implemented as a post-processing stage to the ML model to perform a final filter on its computed results.

Section V describes the results gathered from implementing our methodologies and models for FDI of both abrupt and incipient failures. The final implemented model's architecture is described in section VI and section VII lists our conclusions and future work.

III. Background and related works

Cerro Matoso S.A. (CMSA) is Colombia's biggest lateritic nickel ore extraction, mining and smelting operation, producing over 35,000 tons of ferronickel (FeNi) per year [1]. The smelting stage of CMSA's operation takes place in an electric arc furnace measuring 21 meters in diameter and 7 meters in height, where calcine is fed into the furnace trough feeding tubes located at its ceiling. Ferronickel separation takes place

[1] Groups of data observations sharing similar characteristics and identified by a common label.

at the furnace's middle wall, and hence temperature must be constantly monitored, as the middle wall is the most thermally and chemically active region.

CMSA's personnel constantly modifies different operational parameters including electrode's electrical power and physical position, calcine input flow through the 27 different intake tubes, water flow for heat exchangers located at the medium wall; some important parameters, like calcine chemical composition, are difficult to track on-line. Progress has been made on using on-line operative measurements for predicting the calcine chemical composition [22], the expected temperature at the furnace's walls [11] and estimating the refractory lining thickness inside the furnace's walls [10]. All cases require having reliable temperature data to construct the models.

The furnace's circumference is divided into quadrants and each quadrant is divided into 18 alphabetically labelled sections, segmenting the furnace's middle wall in 72 sectors labelled numerically as shown in Fig. 2. Each of the 72 sectors represent the angular position of a heat exchanging plate, where the furnace's inner temperature is monitored using thermocouples embedded in its wall.

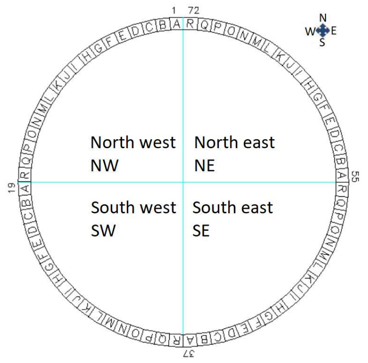

Fig. 2. Coolers distribution over the furnace's mid-wall (top view)

The furnace's wall is divided into three main sections along the vertical axis, upper, middle and lower wall. The furnace's middle wall is covered by 5 rows of heat exchanging plates, 4 rows of Plate Coolers (PCs) and one row of Waffle Coolers (WCs) as shown in Fig. 3. Plate Coolers are grouped by height and named alphabetically. The PC's temperature is monitored by one thermocouple per plate, and the WC's by two thermocouples per plate.

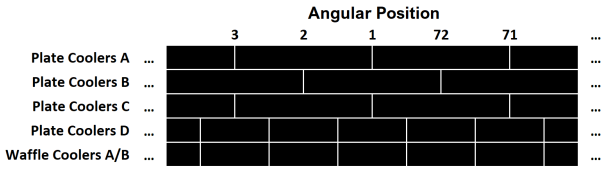

Fig. 3. Coolers distribution over the furnace's mid-wall (front view)

Accurately detecting incipient failures in the thermocouple network will allow CMSA to schedule and perform preventive actions on their sensing system rather than waiting for the need for immediate corrective actions, increasing the reliability of the measurements provided by the sensor network. This is of great interest since this is a keystone of the furnace's safety and operation control system. Non-reliable data can lead to several potential conditions ranging from low productivity due to misguided temperature-dependent control systems, up to events that can be extremely hazardous for the personnel, the plant, and the environment.

CMSA's on-line monitoring system currently has features that detect temperature anomalies and hence guide on-site inspections, but there are currently no available records of incipient failures in the maintenance log.

As incipient failures are not common during the sensor's lifespan, and it is desired to work exclusively on production line data with no artificial registers, appropriate data labelling rules are required to create a balanced baseline set from highly unbalanced data. The provided dataset includes temperature measurements from 322 thermocouples: 37 PC A, 34 PC B, 36 PC C, 72 PC D and 143 WC, whose temperature is logged periodically every 15 minutes starting from 30/09/2015 up to 30/09/2019, for a total $56,620,931$ individual data points.

Figure 4 depicts the data distribution of the filtered and normalized dataset [2] .

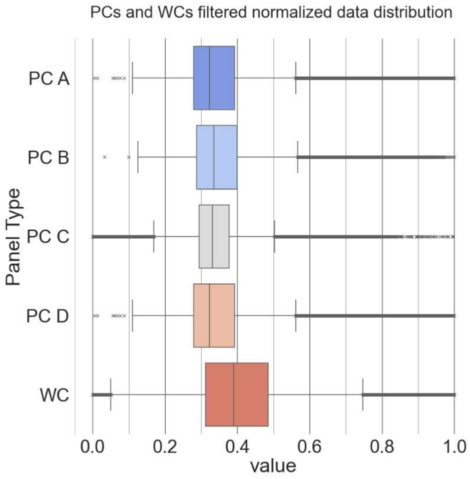

Fig. 4. Data distribution box plot

Figure 4 shows how the Waffle Coolers' temperature tends to be higher than the rest of the panels, as the normalization rule has the same boundaries for all panels. It is also possible to see that temperature values for the Waffle Coolers are more disperse in comparison to the values for the Plate Coolers; this in accordance to what is to be expected in the most active zone of the furnace.

The first setback to overcome is the absence of a baseline for incipient failures. This baseline is necessary for training and validation purposes of any proposed FDI model. Figure 5 depicts an example of an incipient failure, a short-time variation of the value from one of the PC B in the NE quadrant. These types of events can be rarely noticed by the plant operators as their values do not exceed operating conditions. Hence, no alarm is raised.

Supervised FDI models for a blast furnace have been documented in several cases. Shi et al. [21] proposed a neural networks-based system for detecting burden surfaces in a steel production furnace. The scheme was based on radar spectrum

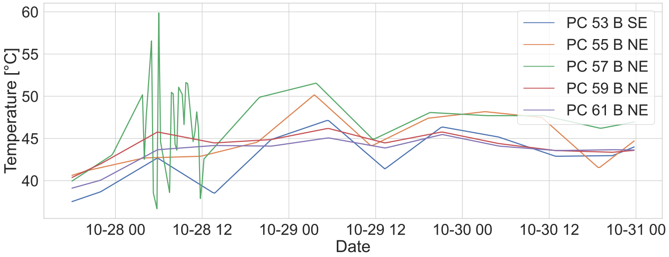

Fig. 5. Example of incipient failure in a thermocouple

generated images, where a neural network with six layers was trained to produce a set of features, and then a neural network with ten layers and a YOLOv3 [16] module was used to predict the existence of a burden surface. The proposed model achieved an average prediction over 99% for five prediction classes.

Leahy et al. [9] classified incipient faults in wind turbine components using a Support Vector Machine (SVM) and a manually-labelled dataset under 4 operational conditions: no fault, general fault, specific fault, fault prediction (about to happen); the authors achieved a recall of up to 97% in a multi-label classification system and encountered challenges keeping up with accuracy. The authors highlight their model's high accuracy and also note that one of their labels has over 50% of its samples wrongly classified, this event is not generating a noticeable impact in the accuracy score as the class with the classification issue has less than 100 observations, while the remaining classes include several thousands of accurately classified observations. For the present work, balanced training and validation datasets are employed so any possible classification issue has a significant impact on the performance metrics.

Mandal et al. [12] proposed a deep learning strategy to detect different failures in a nuclear reactor's thermocouples system. Their approach relies on a solidly established baseline that relates different sensors' known failures and their read signals. Mandal's work and results highlight the relevance of a reliable baseline, which is one of the work's key achievements.

IV. Methods

To exclude abrupt failures, it is necessary to identify them first. A combination of rolling time window analysis and heuristic rules is proposed to achieve this goal. Once abrupt failures have been identified, rolling time window analysis and sensors neighbouring is proposed as a tool to approximate physical redundancy. Data features are extracted from the different neighbour-window combinations, and an ML training and testing dataset is extracted from the features and the baseline data. Dataset subsampling methods are required, considering the expected high unbalanced nature between healthy and faulty sensor data.

Different ML models are trained and evaluated, and a final post-processing stage is implemented to compensate for issues related to the subsampling performed during the ML training and testing dataset conformation.

A. Rolling window analysis and abrupt failures detection

The first stage is studying the abrupt failures within the data, which are simpler to detect and isolate. A combination of rolling window analysis with heuristic rules is applied to detect common abrupt failures in the provided dataset. Through rolling window analysis [13] large amounts of chronologically ordered data can be sampled and studied as individual blocks of standardized width. Data analysis criteria can be applied to each independent data window to determine whether a condition is met during that specific time-lapse.

Each data window is labelled after the initial timestamp, so a chronologically ordered dataset contains equally distanced observations. Given the periodicity of the data logging, rolling window widths are chosen as multiples of $1/4$ of an hour(i.e. a 1-hour window will contain 4 data points, a 1-day window will contain 96 data points and so on).

Outliers and frozen values are two common types of abrupt failures. Frozen values are characteristic of issues like sensor disconnection, short circuit, open circuit, among other disabling failures. For abrupt failure detection, sensors are studied individually, and their data is analysed as described below.

An initial time window is taken with the specified width. For outlier detection, all individual registers within the selected window are compared against the high and low-value thresholds; the selected time window is labelled as 'outlier' if the total count of registers outside the value thresholds surpasses a given maximum outliers count parameter. Frozen values are detected by computing the variance of the sensor's data within the time window; windows with zero variance means the value is not changing, and therefore it is labelled as 'frozen'. After both analyses are done, the time window is moved one timestamp forward and analyses are repeated until the last window of the given width has been studied.

An outlier count limit to decide whether the whole window is labelled as faulty by has to be established. Given the rolling window time width vs. registers count equivalence, a parameter sweep is carried out to study the number of windows that will be labelled as faulty for different outliers thresholds. Outlier threshold is defined then as a minimum count of values out of limits, defined as a percentage of the number of registers in the studied time window. Minimum value count will be rounded using the roof function and constrained to be not lesser than 1.

B. Sensors Neighbouring

Neighbourhood-based sensor configurations have been previously proposed by Fekete et al. [4], since nodes share similar properties due to their proximity. Given the current configuration of the presented furnace, where it is impossible to place more sensors for physical redundancy, and there is a lack of a baseline, set of rules, or model for analytical redundancy, a Neighbourhood-based sensor configuration is proposed to approach, as close as possible, physical redundancy based on data similarity between nearby nodes.

In the studied setting, sensor neighbouring focuses on a group of nearby sensors, or neighbourhood, around a central thermocouple. Neighbours are always considered within a same level, as thermal and chemical reactions inside the furnace change significantly in the vertical direction. Correlation matrices are proposed to verify the validity of the neighbouring approach. Correlation matrices have been used before to perform FDI in a Battery Management System based on the expected similarity between healthy batteries working in a single unit. [19]

For data displaying purposes, all thermocouples will be named as:

$$\text{[Panel type][angular position][level/letter][quadrant]}$$

Panel types can be either WC for Waffle Coolers or PC for Plate Coolers, and level or letter depends on the panel type. PC are organised by levels from A to D and WC contain two thermocouples per plate, named A and B. As an example, a sampled neighbourhood is defined around [**WC 33 A SW**]. The neighbourhood's timeline data is plotted in Fig. 6 and correlation matrices analysis is performed to verify similarity. Analysis using correlation matrices requires testing different time horizons to verify that sensors in a neighbourhood are related. Computing the correlation matrix for a full timeline tends to generate values close to zero. An effective strategy is to use rolling time windows and explore if the correlation changes. This strategy helps to avoid false negatives based on faulty data present in the dataset. The same behaviour was found for different neighbourhoods, which leads to the conclusion that sensor neighbouring may be considered as physical redundancy in the thermocouple network.

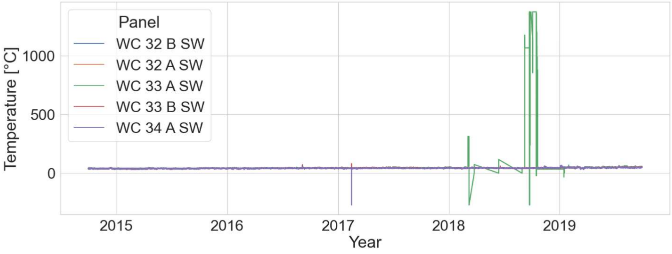

Fig. 6. Measured data for sample 1 neighbourhood

C. Identification of incipient failures and baseline creation

Heuristic rules allowed the identification of abrupt failures and the consequent reduction of the amount of data required to create a knowledge baseline for incipient failures. Sensor neighbouring, window analysis and correlation matrices were used to initially identify healthy sensors with correlations over 0.9 between all members of the neighbourhood. Visual inspection was used to search for incipient failures in the remaining data.

Sensor data was divided into blocks with a width of 2 weeks, and neighbourhoods were determined within a same plate type and level to allow visual identification of incipient failures. When an incipient failure was evidenced (an example is depicted in Fig. 7), the divergent thermocouple point-data was tagged as in *failure*. If a neighbourhood contains at least one sensor labelled as in failure, the whole neighbourhood will be labelled as in failure. This collection of labelled data is the knowledge baseline that allows, among other tasks, assessing the performance of an incipient failure FDI model.

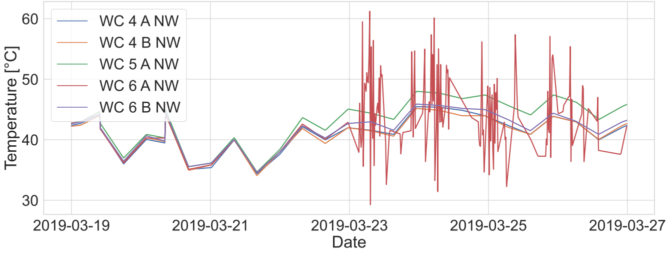

Fig. 7. Incipient failure evidenced by visual inspection

It must be highlighted that the baseline data comes from various thermocouple neighbourhoods and timestamps. Therefore, value discontinuities exist between the boundaries of every consecutive pair of neighbourhood data streams. Additionally, highly unbalanced data is expected between incipient failures and the healthy data.

D. ML dataset and model

Rolling time window analysis is performed on each baseline data stream to avoid boundary discontinuities. When the rolling window hits a stream's boundary, the window's starting date moves to the next stream's starting date. To define the ML labels dataset, the concept of population sample size from statistical theory is adapted to establish the minimum number of individual registers needed to be in failure within a time window to be considered entirely as in failure. Following the population sample equation $n = \dfrac{Z^2Npq}{NE^2 + Z^2pq}$ we can determine the minimum number n of faulty registers required for a window to be labelled faulty.

Having a total population N (i.e. the window width in registers), the confidence interval Z is set to 95%, the sampling error E is set to 10% to grant a slightly more flexible sample size compared to the frequently used value of 5%, and the p and q event probabilities are directly computed from baseline composition as these events are mutually exclusive.

To define the ML features dataset, for each neighbourhood time window, individual sensor median value, variance, and slope are computed and taken as representative features, as well as the correlation values between each sensor within the neighbourhood. The same rolling window boundary rule was applied to compute the labels that indicate whether a time window is considered in failure.

Once the features dataset has been established, the baseline is divided into training and validations datasets for the existence or absence of incipient failures. Both datasets are selected so they contain the same number of faulty and non-faulty data, avoiding sampling-induced bias in the model. Information loss in the ML datasets is expected due to the required subsampling of the non-faulty data. Complete timeline validation is proposed to assess ML models under expected industrial data conditions and evaluate the effects of the subsampling process.

As there is currently no previous information regarding the most suitable ML model and time window width for the case study, three ML models were initially considered: Random Forest (RF) [5], Support Vector Machine (SVM) [24] and Extreme Gradient Boosting (XGB or XGBoost) [22].

The experiments' performance can be measured based on sensitivity, recall, and accuracy [7]. Sensitivity denotes the model's capability to accurately detect faulty data. Recall relates to the model's capability for detecting data without faults, and accuracy is an overall measurement of the model's performance for accurately classifying data either as 0 or 1.

ML model validation is usually performed using a balanced data set under the assumption that the training data set reflects the highest possible number of scenarios present in the initial data. For the present case study, the training, and validation datasets represent a percentage of the total data below 1%, as stated in subsection V-B.

E. Complete timeline validation and up-down counter

Complete timeline validation uses the whole baseline to evaluate the performance of a trained ML model, considering that highly unbalanced data such as the one used here can produce skewed results.

As in k-folds or hold out validation, performance is measured using the general metrics, with a special clarification related to accuracy calculation, which cannot be considered a reliable metric since the data is highly unbalanced.

Up-down counters (UPC) [14] act as non-memoryless filters and help the discrete time decision-making process based on computed residuals. UPC requires 6 parameters: low, high and detection threshold values, and the internal counter's initial value, increase and decrease rates. When a residual is given to the UPC, it increases its internal counter value if the residual value is greater or equal to its detection threshold, or decreased otherwise. The UPC output is set to 1 if the internal counter reaches the high threshold, and set to zero if the internal counter reaches the low threshold.

Up-down counters can be symmetrical or asymmetrical depending on whether its internal counter increases and decreases its value at the same rate, not limited to linear-based functions. Up-down counter filtering is proposed as a post-processing stage to reduce the influence of intermittent changes in the output of the ML model; the UPC detection threshold is set to 1.

For the discrete time fault detection algorithm proposed in subsection IV-D, a considerable percentage of false positives and false negatives is expected, considering that a significant amount of non-faulty data was left out of the baseline due to the size constraint imposed by the data balancing requirement.

V. Results

Previous sections explained different techniques, which combined established a Machine Learning training and validation process through methods that were required due to the constraints of the case study addressed. This section is dedicated to the main results of the FDI system designed and its validation.

A. Rolling time window analysis and abrupt failures detection

Outlier detection by applying heuristic rules allowed the identification of several windows to avoid during baseline creation. Table I summarises the outlier detection results for the parameter sweep of window width and outliers threshold.

	Outliers threshold [%]						
Window	**10**	**30**	**50**	**70**	**90**	**100**	
1 hour	0.50	0.50	0.50	0.47	0.47	0.47	Windows with outliers [% of total]
6 hours	0.60	0.54	0.49	0.44	0.39	0.37	
12 hours	0.68	0.56	0.48	0.40	0.33	0.38	
1 day	0.79	0.60	0.45	0.34	0.26	0.23	

TABLE I
PERCENTAGE OF WINDOWS IDENTIFIED AS 'WITH OUTLIERS' FOR DIFFERENT WINDOW WIDTHS AND OUTLIERS THRESHOLD

Frozen windows were tagged using a window width parameter sweep and determining the percentage of frozen windows from the existing total. Using a window width of 1 hour, 2% of the total data was identified as frozen. For a 6-hour window width, 1.07% of the total data was identified as frozen. For a 12-hour window width, 0.84% of the total data was identified as frozen. And for a 1-day window width, 0.79% of the total data was identified as frozen.

B. Identification of incipient failures, baseline creation and ML dataset definition

As expected, comparatively few individual observations of incipient failure were obtained compared to the available data set. Tagging existing data with known incipient failures is the cornerstone for creating a baseline. As the baseline needs to be as balanced as possible, samples of healthy sensor data were selected to create a balanced baseline. Table II summarises the created data set based on a rigorous data inspection. The baseline combines both data in failure and healthy data, and individual registers are labelled to be used either for training or validation purposes.

Panel type	**Registers count**	**Registers in fail**	**Faulty data [%]**
Plate Cooler A	21,500	11,732	54.56
Plate Cooler B	648	264	40.74
Plate Cooler C	10,512	5,424	51.59
Plate Cooler D	5,136	2,592	50.46
Waffle Cooler	7,056	4,128	58.50
Total	**44,852**	**24,140**	**53.82**

TABLE II
BASELINE REGISTERS COUNT, REGISTERS IN FAIL COUNT AND PERCENTAGE OF REGISTERS LABELLED AS IN FAILURE

The baseline data set size is constrained by the number of identified incipient failures. The baseline data set constructed contains a total of 44,852 individual registers, which represents the 0.08% of the available data.

By applying the algorithm described in subsection IV-D different data sets were obtained, one per window width. Each data set register includes 5 individual medians, 5 individual variances, 5 individual slopes and 10 correlations.

The population sample equation from subsection IV-D and data from table II were used to determine the minimum number of registers required to label a whole window as in failure. For PC A to C and WC the same minimum failure registers count is required to label the window as in failure. For a window width of 1 hour, 4 faulty registers are the minimum count to label the window as in failure; 20 registers for a window width of 6 hours; 32 registers for a window width of 12 hours, and 78 registers for a window width of 1 day. For the specific case of the PC D there's only one different minimum count, 49 faulty registers are the requirement for a width window width of 1 day.

With the given rules, any time window having a faulty register count above the fore-mentioned thresholds is labelled as in failure. Table III summarises the resulting faulty window count for each time window width and panel type.

Window width	Window count	Windows in fail	Percentage faulty windows [%]
1 hour	44,848	24,079	53.69
6 hours	44,828	23,713	52.89
12 hours	44,804	23,701	52.89
1 day	44,756	24,133	53.92

TABLE III
LABELS DATASETS FOR WINDOW WIDTH PARAMETER SWEEP

C. ML model training and tuning

Table IV summarises the results obtained from training different ML models and using different rolling time window widths.

ML model	Window width	Sensitivity	Recall	Accuracy
Random Forest	**1 h**	0.9764	0.9809	0.9785
	6 h	0.9690	0.9821	0.9753
	12 h	0.9897	0.9945	0.9920
	1 d	0.9973	0.9974	0.9978
Support Vector Machine	**1 h**	0.9693	0.0364	0.5326
	6 h	0.2803	0.9790	0.6148
	12 h	0.3576	0.9591	0.6431
	1 d	0.4574	0.9257	0.6762
XGBoost	**1 h**	0.9723	0.9798	0.9758
	6 h	0.9506	0.9725	0.9611
	12 h	0.9734	0.9884	0.9805
	1 d	0.9973	0.9990	0.9980

TABLE IV
CENTRAL-COMPOSITE EXPERIMENT RESULTS

It is possible to observe from table IV that XGBoost for a window width of 1 day shows the highest sensitivity, recall, and accuracy among all performed experiments. Hyperparameter optimization using the early stop technique [15] is used to improve the model's performance. Table V summarises the performance of XGBoost models after hyperparameter optimization.

The reader may observe that the four time window widths have been kept during the hyperparameter optimization process. This axis of the experimental result will be kept up throughout the ML model development process to observe its influence in the final results.

Window width	Sensitivity	Recall	Accuracy	Final estimators
1 h	0.9856	0.9876	0.9876	481
6 h	0.9887	0.9887	0.9887	568
12 h	0.9972	0.9973	0.9972	754
1 d	0.9975	0.9994	0.9984	355

TABLE V
XGBOOST HYPERPARAMETER OPTIMIZATION RESULTS

D. Complete timeline validation and up-down counter

Full timeline data sets used for complete timeline validation are defined around the Waffle Cooler [**WC 42 B SE**] and the Plate Cooler [**PC 28 A SW**]. [**WC 42 B SE**] is the sensor with the highest presence of incipient failures, having 28.78% of its data with presence of such failures in the measured values; sensor [**PC 28 A SW**] has 0.5% of its data with presence of incipient failures.

The algorithm depicted in subsection IV-D is applied for different window widths, and the corresponding trained XGBoost model is used to detect incipient failures. Table VI compiles the performance metrics results for the complete timeline validation using the neighbourhood around the panel [**WC 42 B SE**] as an example.

Window width	Sensitivity	Recall	Precision
1 h	0.4627	0.9459	0.8061
6 h	0.4794	0.9535	0.8163
12 h	0.5168	0.9243	0.8064
1 d	0.4980	0.7906	0.7059

TABLE VI
FULL TIMELINE VALIDATION RESULTS

A significant sensitivity drop is evidenced, suggesting a difference between the traditional hold-out validation carried out during the ML model training process and how the model would perform with online discrete time data input (full-time validation). Visual inspection of predicted labels reveals intermittence in the model's output, as shown in the example of Fig. 8 for predictions in a two-day lapse during the year 2017.

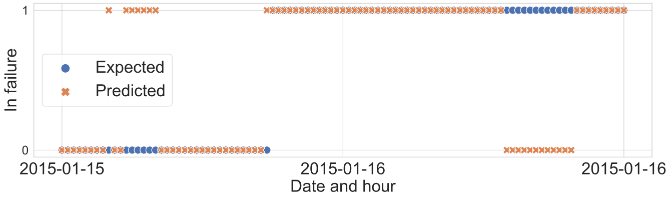

Fig. 8. Comparison of predicted vs. expected labels by ML model

Symmetrical and asymmetrical up-down counters were tested as a post-processing stage to reduce the prediction intermittence. Unit linear increase and decrease rates are selected for the UPC internal counter, with an initial value of 0. The low threshold value is set to 0 and the high threshold value is chosen by a parameter sweep. An asymmetrical filter instantly increases or decreases the internal counter to its limit value, while the opposite internal counter change rate is kept linear with unit value.

Figures 9 and 10 illustrate the change in sensitivity and recall, respectively, for the fault detection process by including a symmetrical up-down counter after the ML model prediction. A High threshold value parameter sweep is performed and results are grouped by rolling window width for performance comparison purposes.

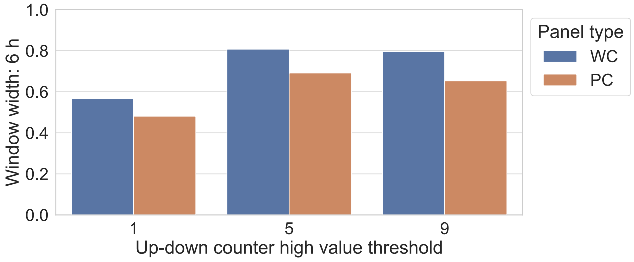

Fig. 9. Fault detection sensitivity using symmetric up-down counter

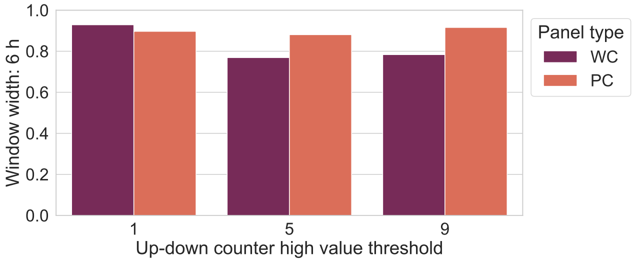

Fig. 10. Fault detection recall using symmetric up-down counter

Overall, detection sensitivity improves considerably with the addition of the up-down counter, almost doubling their original values. As for detection recall, a slight performance reduction occurs when the up-down counter is used. Similar improvement was noted for the different time window widths, being a 6-hour window width the case with the highest overall sensitivity and recall among all scenarios. The most suitable model is chosen by selecting an appropriate high threshold value with the most balanced trade-off between detection sensitivity and recall scores.

Numerical data of the plotted results suggest that the best performance is obtained for a high threshold value of 5 and a window width of 6 hours. Sensitivity was rated at 69.17% for the complete timeline validation of the Plate Coolers and 80.77% for Waffle Coolers; sensitivity results for Plate Coolers and Waffle Coolers are 88.07% and 76.92% respectively.

VI. Final model generation process

The final model has a 5-stage architecture, which starts with combining data from neighbouring thermocouples into one time window at a time (from now on called segment) from the initial 322 columns data set. Second, the existence of abrupt failures is evaluated in the segment via the heuristic rules; the whole segment is labelled as 'abrupt failure' if abrupt failures are detected and the process starts over with the next segment.

Segments free of abrupt failures are considered in the third stage of the process (feature extraction), which computes the features as discussed in subsection V-C and injects the computed values into the trained XGBoost model (i.e. the fourth stage). In the fifth and last stage, the ML model outputs are fed to the symmetrical UPC, finally detecting incipient failures.

For each individual segment, the sequential five-stage model computation process, which includes the ML model for incipient failure detection and the UPC, can be considered as $O(1)$ time complexity. The segment's shape is fixed, and the UPC's internal counter's value changes only based on the most recent input value.

One limitation of our approach is its reliance on expert knowledge to both generate a baseline and to verify its performance. Similar to other data-driven approaches, data labelling and data engineering is required; for this industrial process, additional process verification is required from planned maintenance and repairs to the oven. Another limitation arises from edge cases related to potentially missing sensors or faulty sensors reporting abrupt failures, reducing the number of available sensors in a neighbourhood, potentially missing incipient failures.

VII. Conclusions

A data-driven fault detection and isolation (FDI) model for incipient failures in an industrial thermocouple network was achieved, with a detection accuracy of up to 80% of faulty time windows. The model includes a detection stage for abrupt failures using rolling windows, a features computing stage, a detection stage for incipient failures using an ML model, and a post-processing stage using a symmetrical up-down counter.

A methodology to create a balanced baseline of abrupt and incipient failures based on heuristic rules and expert knowledge was introduced. This methodology allowed us to process sensor measurements from an industrial furnace with a complex thermocouple network and a lack of operational logs. The methodology used the concept of physical redundancy approximation by using measurements from neighbouring sensors.

Different models were tested with a range of time window widths to get the highest sensitivity and recall metrics. A time window of 6 hours produced the highest sensitivity and recall over the entire dataset. Results were improved further by implementing a symmetric up-down counter to post-process the outputs of the model and improve predictions.

Future work will explore different data-driven models with alternative approaches for FDI to look for improved accuracy and better capability for handling highly unbalanced data.

Acknowledgement

This work has been funded by the Colombian Ministry of Science, Technology and Innovation through the grant number 786, "Convocatoria para el registro de proyectos que aspiran a obtener beneficios tributarios por inversión en CTeI".

References

[1] Luis Carlos Bonilla Flórez and Rubén Rangel de Hoyos. “Modelo de optimización del consumo unitario de carbón del proceso RKEF en Cerro Matoso SA”. MA thesis. Universidad del Norte, 2016.

[2] Jaiber Camacho-Olarte et al. “A Data Cleaning Approach for a Structural Health Monitoring System in a 75 MW Electric Arc Ferronickel Furnace”. In: *Engineering Proceedings* 2.1 (2020). ISSN: 2673-4591.

[3] Zhenzhen Dong et al. “Research on Agricultural Machinery Fault Diagnosis System Based on Expert System”. In: *Proceedings of 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC 2018* Imcec (2018), pp. 2057–2060.

[4] S. P. Fekete et al. “Neighborhood-Based Topology Recognition in Sensor Networks”. In: *Algorithmic Aspects of Wireless Sensor Networks*. Ed. by Sotiris E. Nikoletseas and José D. P. Rolim. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 123–136. ISBN: 978-3-540-27820-7.

[5] Radhia Fezai et al. “Effective Random Forest-Based Fault Detection and Diagnosis for Wind Energy Conversion Systems”. In: *IEEE Sensors Journal* 21.5 (2021), pp. 6914–6921.

[6] Mohamed Hafid and Marcel Lacroix. “Inverse method for simultaneously estimating multi-parameters of heat flux and of temperature-dependent thermal conductivities inside melting furnaces”. In: *Applied Thermal Engineering* 141 (December 2017 2018), pp. 981–989. ISSN: 13594311.

[7] Guy S. Handelman et al. “Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods”. In: *American Journal of Roentgenology* 212.1 (2019), pp. 38–43.

[8] Patton. R. J. “Using Analytical Redundancy”. In: *Computing & Control Engineering* 8.May (1991), pp. 127–136.

[9] Kevin Leahy et al. “Diagnosing wind turbine faults using machine learning techniques applied to operational data”. In: *2016 IEEE International Conference on Prognostics and Health Management (ICPHM)*. 2016, pp. 1–8.

[10] Jersson X. Leon-Medina et al. “Monitoring of the refractory lining in a shielded electric arc furnace: An online multitarget regression trees approach”. In: *Structural Control and Health Monitoring* 29.3 (Nov. 2021).

[11] Jersson X. Leon-Medina et al. “Temperature Prediction Using Multivariate Time Series Deep Learning in the Lining of an Electric Arc Furnace for Ferronickel Production”. In: *Sensors* 21.20 (2021). ISSN: 1424-8220.

[12] Shyamapada Mandal et al. “Nuclear Power Plant Thermocouple Sensor-Fault Detection and Classification Using Deep Learning and Generalized Likelihood Ratio Test”. In: *IEEE Transactions on Nuclear Science* 64.6 (2017), pp. 1526–1534.

[13] Daiki Matsunaga, Toyotaro Suzumura, and Toshihiro Takahashi. *Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis*. 2019. arXiv: `1909.10660 [cs.LG]`.

[14] Ahmet Arda Ozdemir, Peter Seiler, and Gary J. Balas. “Wind turbine fault detection using counter-based residual thresholding”. In: *IFAC Proceedings Volumes (IFAC-PapersOnline)* 44 (1 PART 1 2011), pp. 8289–8294. ISSN: 14746670.

[15] Lutz Prechelt. “Early Stopping — But When?” In: *Neural Networks: Tricks of the Trade: Second Edition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 53–67. ISBN: 978-3-642-35289-8.

[16] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*. 2018. arXiv: `1804.02767 [cs.CV]`.

[17] Aroba Saleem et al. “Electromagnetic Measurement of Molten Metal Level in Pyrometallurgical Furnaces”. In: *IEEE Transactions on Instrumentation and Measurement* 69.6 (2020), pp. 3118–3125.

[18] Paraskevi A. Samara et al. “A statistical method for the detection of sensor abrupt faults in aircraft control systems”. In: *IEEE Transactions on Control Systems Technology* 16.4 (2008), pp. 789–798. ISSN: 10636536.

[19] Michael Schmid, Hans-Georg Kneidinger, and Christian Endisch. “Data-Driven Fault Diagnosis in Battery Systems Through Cross-Cell Monitoring”. In: *IEEE Sensors Journal* 21.2 (2021), pp. 1829–1837.

[20] Reza Shahnazi and Qing Zhao. “Adaptive Fuzzy Descriptor Sliding Mode Observer-based Sensor Fault Estimation for Uncertain Nonlinear Systems”. In: *Asian Journal of Control* 18.4 (2016), pp. 1478–1488. ISSN: 19346093.

[21] Qiudong Shi et al. “A Blast Furnace Burden Surface Deep learning Detection System Based on Radar Spectrum Restructured by Entropy Weight”. In: *IEEE Sensors Journal* 21.6 (2021), pp. 7928–7939.

[22] Diego A. Velandia Cardenas et al. “Data-driven classification of the chemical composition of calcine in a ferronickel furnace oven using machine learning techniques”. In: *Results in Engineering* (2023), p. 101028. ISSN: 2590-1230.

[23] Peng Wang, Xiaoyan Chen, and Lei Yu. “Application of Functional Safety Theories in Furnace Safety Supervisory System”. In: *2017 9th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. 2017, pp. 164–167.

[24] Shuang Yu et al. “The OCS-SVM: An Objective-Cost-Sensitive SVM With Sample-Based Misclassification Cost Invariance”. In: *IEEE Access* 7 (2019), pp. 118931–118942.

[25] Tongshuai Zhang et al. “Fault diagnosis for blast furnace ironmaking process based on two-stage principal component analysis”. In: *ISIJ International* 54.10 (2014), pp. 2334–2341. ISSN: 09151559.