A Data-Driven Methodology to Assess Raw Materials Impact on Manufacturing Systems Breakdowns

Maha Ben Ayed[1,2], Moncef Soualhi[1], Raouf Ketata[2], Nicolas Mairot[3], Sylvian Giampiccolo[3], and Noureddine Zerhouni[1]

[1] *Université de Franche-Comté, SUPMICROTECH, CNRS, institut FEMTO-ST, F-25000 Besançon, France*
maha.benayed@femto-st.fr
moncef.soualhi@femto-st.fr
noureddine.zerhouni@femto-st.fr

[2] *INSAT, Univ. Carthage, Boulevard de la terre, BP 676, Tunis, 1080, Tunisia*
raouf.ketata@insat.rnu.tn

[3] *SCODER, 1 rue de la Forêt Z.A. l'Orée du Bois, Pirey, 25480, France*
nicolas.mairot@scoder.fr
s.giampiccolo@scoder.fr

ABSTRACT

Data-driven Prognostics and Health Management (PHM) become a crucial layer in the realm of predictive maintenance (PM), particularly for metal-forming industries. In fact, non-compliant material characteristics affect badly the manufacturing tools leading to high machine breakdown frequency and poor quality products. To cope with this situation, a new methodology for breakdown prediction is proposed. In detail, the methodology starts by implementing an Extract, Transform, Load (ETL) process to create a new dataset from heterogeneous sources. Then, a feature selection method is used for dimensionality reduction and keeps only useful information. After that, a Machine Learning (ML) model predicts system breakdown occurrences using the selected features. Finally, thanks to these steps above, an auto-labeling algorithm to evaluate the severity impact of the material data is proposed and makes the originality of this paper. The developed methodology is applied to a real dataset of a French company, *SCODER*, that shows and points out promising perspectives in PM.

Keywords: Prognostics and Health Management, Manufacturing, Raw Material Data, Extract-Transform-Load, Features Selection, Machine Learning, Auto-labeling.

1. INTRODUCTION

Prognostics and health management (PHM) in industries made significant progress in the management of their data pipelines. It contributes to the development of advanced monitoring algorithms and optimizing production scheduling in their systems. In fact, a monitoring system that enables early detection of potential breakdowns plays a pivotal role in supporting decision-making, resulting in minimized breakdowns, reduced repair costs, and cost savings for businesses. However, the majority of prediction systems excessively focus on sensor data while ignoring the significance of raw materials. Indeed, the presence of non-compliant material characteristics can contribute to reduced production efficiency, especially when the manufacturing process heavily relies on specific material properties. Moreover, slight deviations from these properties can lead to operational issues, an increase in maintenance needs, and unplanned breakdowns. Additionally, when monitoring the condition of their systems, manufacturers often use quantitative methods to predict purely numerical results such as downtime, Remaining Useful Life (RUL), downtime frequency, and so on (Benaggoune, Mouss, Abdessemed, & Bensakhria, 2020). They rarely focus on qualifying the severity of a failure to facilitate decision-making. In fact, actual monitoring algorithms in industries focus on sensor data for fault detection, diagnostics and prognostics (FDDP) (Abd Al Rahman & Mousavi, 2020), due to their availability and development in this last decade. Also, industrial operators are increasingly aware that sensor data provides valuable information on the status of industrial systems (Raouf, Khan, et al., 2022). For instance, the analysis of vibration data has proven to be an efficient parameter in detecting anomalies and anticipating system failures (Farahat, Gupta, et al., 2020) In power transmission systems, vibration analysis can help monitor the condition of gears and belts,

https://doi.org/10.36001/IJPHM.2024.v15i1.3818

and identify mechanical issues (C. Li et al., 2016). Similarly, for rotating machinery such as electric motors vibrations can reveal imbalances and bearing faults (Popescu & Aiordachioaie, 2019). Similarly, the analysis of current data showed the possibility of non-invasive monitoring of the health and performance of electrical systems (Soualhi, Nguyen, Soualhi, Medjaher, & Hemsas, 2019), especially for detecting mechanical and electrical faults simultaneously. In fact, sensor measurement can reveal recurring anomalies, identify common causes, and highlight potential areas for improvement. Natural Language Processing has also been integrated into PHM applications. For example, (Ayed et al., 2023) used textual data from industrial machine history and NLP techniques to predict the origin of the next breakdown. In the literature, only a few works have explored the role of raw materials in conducting a comprehensive analysis and gaining an efficient understanding of breakdown mechanisms (Ramprasad, Batra, Pilania, Mannodi-Kanakkithodi, & Kim, 2017). Understanding the specific characteristics of raw materials becomes crucial, as they can introduce vulnerabilities and increase the likelihood of breakdowns throughout the production cycle (Rizzo et al., 2020). Hence, it becomes imperative to incorporate material characteristics into PHM practices to ensure a comprehensive assessment of breakdown risks. Recognizing this, we believe that material nature plays a crucial role in machines breakdowns (Arshadi, Gref, Geladi, Dahlqvist, & Lestander, 2008). It is essential to highlight a key assumption underlying our analysis. In fact, the quality of raw materials is the main factor influencing the frequency and duration of system breakdowns in our study. Despite the various factors that can affect an industrial system, such as environmental impacts, human factors and machine state, the quality of raw materials plays an essential role. This assumption is based on in-depth expertise and empirical observations. It underlies our work and guides our approach to research and data interpretation, aimed at highlighting how variations in raw material quality significantly predict breakdowns. In this paper, we introduce an innovative methodology that combines sensor data analysis with an assessment of raw material properties. By leveraging the benefits of both elements, including insights into the chemical composition and physical attributes found in raw materials, alongside the strengths of sensor data analysis, such as real-time monitoring, companies can quantify the impact of their raw materials on the process's health. This, allows them to enhance raw material management, improve decision-making processes, optimize maintenance schedules (Achouch et al., 2022), and achieve higher levels of performance and productivity in their operations (Soualhi et al., 2023). Navigating this intricate data landscape, we employ the ETL (Extract, Transform, Load) process, a choice motivated by the diverse origins of our datasets (Zhang et al., 2022). Building upon this foundation, we integrate feature selection strategies with state-of-the-art machine learning algorithms to ensure precise classification. During the ETL process stage, we extract data from two primary sources: machines and materials. Then, these data undergo transformation to ensure consistency and exploitability. Once refined, they are stored in a dedicated database, making it ready for analysis. This latter ETL process ensures that data from various sources are harmonized and reliable, thus making it an insightful analysis. The next step concerns feature selection where the objective is to reduce the dimensionality of the feature space and select the most influential features. Subsequently, machine learning algorithms are used to classify materials based on the selected features. The integration of feature selection and machine learning facilitates accurate and efficient material labeling, allowing operators to effectively categorize their material inventories (Rahman et al., 2020). The main contributions are summarized as follows:

- Create and deploy a methodology to predict machine breakdown occurrences based on raw material data
- Auto-label material coils according to their impact on the process to help industries schedule their material consumption.
- Improve industrial performance and anticipate frequent machine-tool breakdowns.

The is organized as follows: section 2 presents the related works with specific positioning. Section3 aims to present the proposed methodology. In section 4 we present the case study and results with a discussion on methodology performances. Finally, a conclusion and future works are presented in section 5.

2. RELATED WORKS

As outlined in the introduction, the methodology begins with an ETL process, followed by feature selection techniques, and raw material-based machine breakdown prediction. In the literature, numerous techniques exist for ETL and feature selection, but there is less exploration of works extending to raw material impact, covering manufacturing processes and supply chains (Y. Li et al., 2022). The ETL process has become a cornerstone of data management. Indeed, using a rigorous and systematic approach to ETL ensures reproducibility, accuracy and efficiency. Nevertheless, the modern data landscape, characterized by digital proliferation and exponential data growth, has spurred the evolution of data management systems. Beyond traditional ETL capabilities, today's systems incorporate advanced functionality such as data integration, quality improvement and governance. Recent work in state of the art describes ETL processes involving feature selection, particularly when dealing with a profusion of variables (Abiodun et al., 2021). The benefits of effective feature selection include improved model accuracy, reduced computational load and improved interpretability. The landscape of feature selection techniques is varied, encompassing filtering methods, wrapping methods and integrated methods (Mera-

Gaona, López, Vargas-Canas, & Neumann, 2021). Feature selection techniques can also rely on other strategies such as dimensionality reduction through Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) to offer ways to reduce the feature space (Winursito, Hidayat, Bejo, & Utomo, 2018). With regard to raw materials, (Stauffer et al., 2019) investigated the management of variability in critical attributes impacting process stability under appropriate granulation conditions in the pharmaceutical field. The process involved the analysis of various factors such as the availability and characteristics of raw materials that have an impact on the production process. In the industrial context, (Borràs-Ferrís, Palací-López, Duchesne, & Ferrer, 2022) proposed a methodology based on the Partial Least Squares (PLS) model to assess the ability of batches of raw materials to produce compliant products. This enables decisions to be made regarding the acceptance or rejection of batches of raw materials from new suppliers. In (Ahmad et al., 2021) authors used material characteristics to predict the shear strength of rockfill materials. Their findings show that the Support Vector Machine (SVM) outperforms other models in predicting the shear strength of rockfill materials, highlighting the crucial role of normal stress in affecting shear strength. In the same context, (Mosavi et al., 2020), groundwater hardness data from 135 wells were analyzed using Boosted Regression Trees and Random Forest (RF) models. The study identified the key influencing factors to be the distance from rivers, elevation, and groundwater depth. However, the research was limited by the quantity of data.

To contribute to the existing methods, our research aims to quantify and label the impact of raw materials on industrial performance. It focuses on data layer by adapting the existing methods from the state of the art to real industrial data. Our positioning about the state of the art can be presented as follows:

- We integrate raw material data with sensor data through a tailored ETL process to create a predictive model.
- We highlight the most relevant material characteristics for the manufacturing process.

Our objective, through this in-depth analysis, is to enhance our understanding of raw materials' effectiveness and their impact on optimizing manufacturing performance.

3. PROPOSED METHODOLOGY: FROM RAW DATA TO AUTO-LABELING

This section outlines the key steps of the proposed methodology for addressing challenges related to machine breakdown prediction and auto-labeling, while considering raw materials proprieties variation. This approach holds significant relevance for small and medium-sized enterprises (SMEs) specializing in metal parts manufacturing. These industries operate stamping lines where the properties of raw materials are rigorously identified and monitored to control their influence on process variations, particularly machine breakdowns. As mentioned in the introduction, this work is based on the assumption that the nature of the material is the main and only factor influencing machine breakdown, and that other process parameters are insignificant. The overall view of the proposed methodology is presented in Figure 1. The methodology, illustrated by Figure 1, is divided into offline and online phases. In the offline phase (subsection 3.1), the ETL process handles data on raw materials and machines, generally merged via intermediary data. This produces a dataset marked by features from C_1 to C_n and real occurrences O_r. A clustering model on occurrence O_r permit to identify severity thresholds and create labels. In this phase, we apply also the feature selection techniques to reduce ML parameters. This step is essential to perform predicting machine breakdowns per hour O_p. In the online phase (subsection 3.2), the characteristics of the materials are automatically integrated into the prediction model, estimating the occurrence O_r with each introduction of material. The predicted value, associated with the clustering labels, determines the final label of the material.

3.1. Offline Phase: Occurrence Predictor Construction

This section describes the key steps aiming at evaluating various feature selection techniques and choosing the most effective Machine Learning (ML) model for predicting breakdown occurrences.

3.1.1. Data Preprocessing & Structuring : ETL Process Application

The preprocessing step involves actions and techniques to clean and transform raw data into exploitable information for analytics and ML algorithms. A general overview of an ETL process is presented in figure 3.

- **Data Extraction:** The objective of this stage is to collect data from their source files. Indeed, data are extracted from two main sources. The first source consists of data generated by machine tools, typically captured through an installed monitoring system. This type of system allows real-time display of machine parameters and health indicators. These records, often referred to as *log* data, are automatically saved as time series in *.txt* files. The second type of data concerns material data, obtained either from suppliers or through laboratory analysis. Stored in an *SQL* database, these data provide essential information about the properties and characteristics of the raw materials used in the studied industrial processes. It is essential to recognize that raw data, regardless of their origin, may contain inaccuracies, inconsistencies, or gaps. Such issues can distort subsequent analyses or predictive models. This underscores the importance of the transformation phase, discussed below, to establish the appropriate preprocessing.

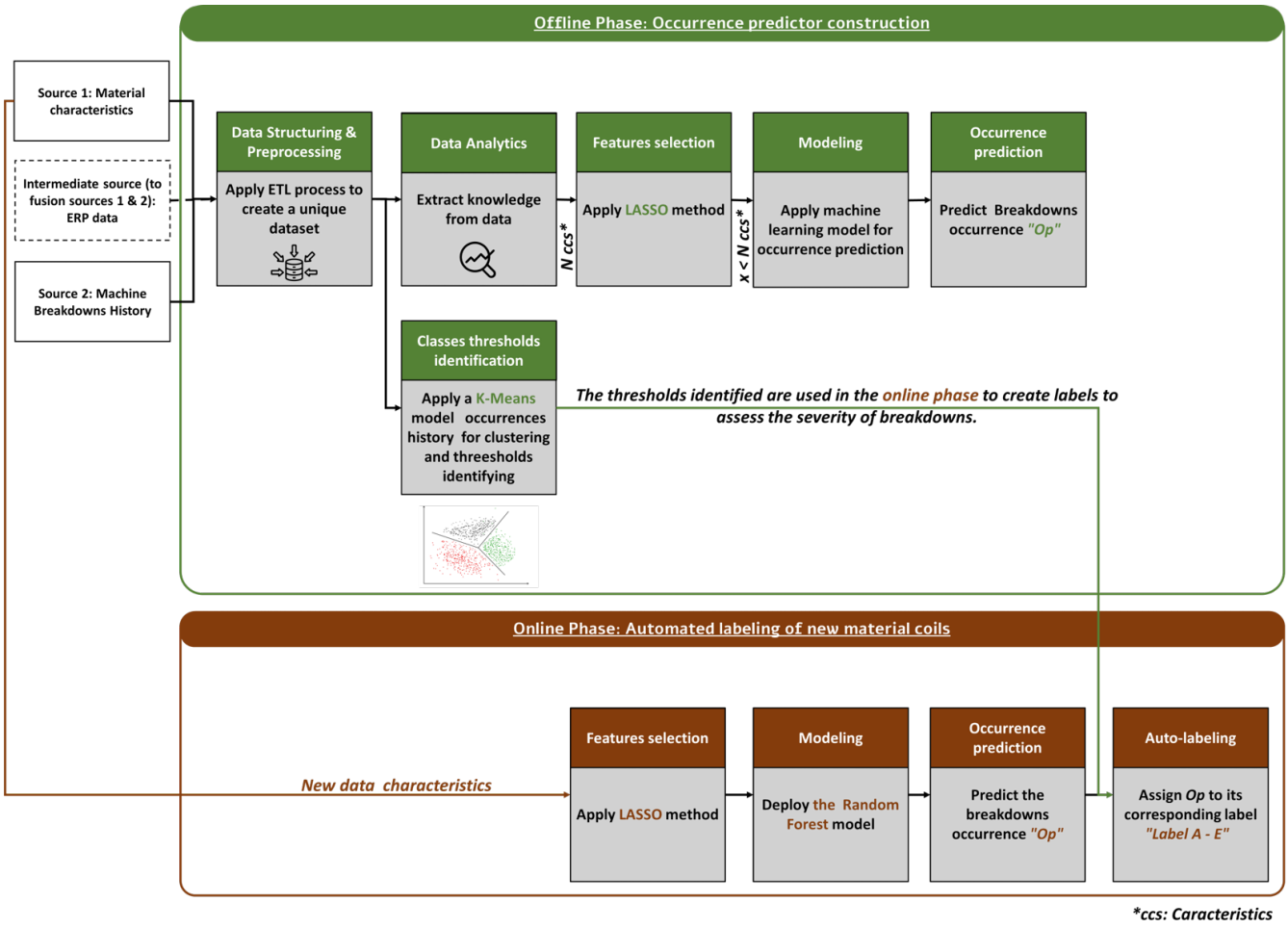

Figure 1. Flowchart of the proposed methodology for breakdowns occurrence prediction

- **Data Transformation:** This phase of the process focuses on transforming data into a suitable format for analysis and modeling. Indeed, data preprocessing is an essential step in this transformation process, underlining the importance of data quality and reliability (Omri, Al Masry, Mairot, Giampiccolo, & Zerhouni, 2021). It starts with a thorough inspection to identify errors and inaccuracies in the datasets, followed by corrective actions. It is important to note that the two data sources mentioned in the "Extraction" phase are not treated in the same way. In fact, machine data undergoes a specific cleaning process tailored to time-series data. The inspection phase helps extract the important columns to be used. Next, we proceed with data type conversion and removal of duplicate values stemming from acquisition system bugs to preserve data integrity. A critical step in data transformation is the occurrence calculation of machine breakdowns per hour based on its history reports. It is important to examine material data carefully and rectify any inconsistencies or outlier values in the measurements, whether they originate from suppliers or result from detailed laboratory analyses. For missing data points, we employ strategies such as imputation or deletion as needed. Furthermore, we apply statistical techniques to detect and address outliers. Depending on their nature and impact, outliers may be adjusted to match the dataset's characteristics or removed to avoid distorting the final results. A generic aspect of this phase involves scaling numerical data into a standardized range and applying mathematical transformations to avoid scale-related biases and enable those variables measured on different scales. The date was transformed into a common range scale $[0, 1]$. This will guarantee variables impact the subsequent analysis without scale bias, which consequently deals with fair and accurate assessments of the data on machine breakdowns. To provide a clear visual representation of these various data cleaning steps, figure 2 illustrates the steps involved in the data cleaning process. After preprocessing the data, we aggregate the transformed datasets (machine & material) using the 'date' variable into a single dataset.

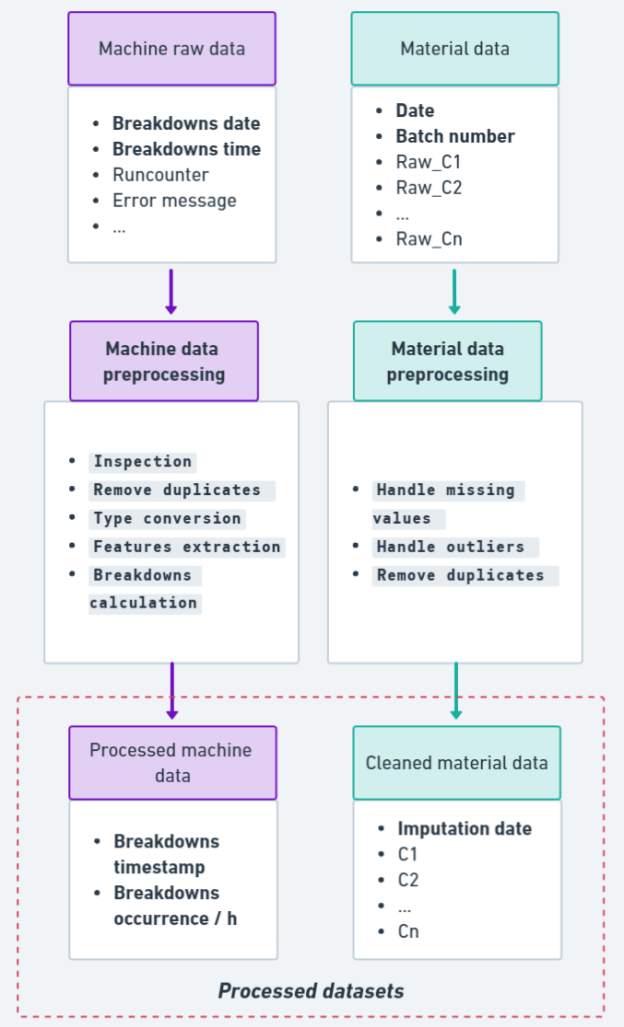

Figure 2. Data Preprocessing Steps

- **Data Loading:** Following the transformation and structuring of the data into a unified dataset, data are loaded into a single *Excel* file. Subsequently, the next stage encompasses data analytics, feature extraction, and the application of modeling algorithms to predict breakdowns.

3.1.2. Data Analytics for Machine Learning

In our methodology, data analytics holds a crucial significance when it evaluates the extracted features and selects the appropriate ones for the ML model. We employed techniques from Exploratory Data Analysis (EDA) to gain a deep understanding of data nature, distribution, correlations, as well as potential quality issues. These informations are essential for making the right decisions regarding feature selection and suitable ML selection. A conducted EDA also allows for the detection of possible redundancies or interactions between features, which can influence the final model selection. Four types of data analytics techniques were conducted: Statistical Analysis, Correlation Analysis, Data Distribution Visualization, and Linearity Analysis. Statistical Analysis involves the use of statistical techniques to understand and summarize data characteristics. This often includes calculating measures of central tendency (mean, median, mode), dispersion (standard deviation, variance), as well as summarizing distributions (quartiles, deciles). Then, a Correlation Analysis is held to examine the relationship between variables, particularly if they are linearly associated. The Pearson correlation coefficient is commonly used to quantify the strength and direction of this relationship, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. The Pearson correlation coefficient (r) is calculated using the following formula:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (1)$$

Another technique in EDA involves visualizing data distribution using graphs and diagrams, such as histograms, box plots, and density plots, to represent the distribution of values in a dataset visually. This helps in observing the shape of the distribution, detecting potential outliers, and identifying trends. Lastly, linearity analysis aims to evaluate whether the relationship between an independent variable and a dependent variable is linear. This evaluation can be performed by plotting a scatter plot of the two variables and visually examining the shape of the relationship. Statistical tests can also be used to confirm linearity. These analyses are of paramount importance for data understanding, selecting appropriate analysis methods, and, most importantly, narrowing down the choice of models to be tested for prediction (Ketata, Al Masry, Zerhouni, & Yacoub, 2023).

3.1.3. Features Selection

In our methodology, feature selection represents a crucial step. Indeed, it is a common step among works interested in ML modeling, especially when dealing with multiple variables. For example. In (Raouf, Lee, & Kim, 2022), authors specifically showed that the chi-square test-based feature selection (Case VI) significantly enhanced classifier performance, achieving the highest accuracy, which underscores the critical role of feature selection in enhancing the effectiveness of mechanical fault diagnosis systems. Another application of feature selection techniques is presented by (Chen & Gao, 2020), who proposed an integrated group-based sensor selection algorithm for manufacturing systems. This approach reduces the number of sensors required for accurate Remaining Useful Life (RUL) estimation and demonstrated an average improvement of 86% in RUL calculation. To guarantee a feature selection process that yields meaningful results for breakdown prediction, we compared and contrasted three distinct methods: the Lasso (Least Absolute Shrinkage and Selection Operator), Recursive Feature Elimination (RFE), and ANOVA (Analysis of Variance). RFE is a tech-

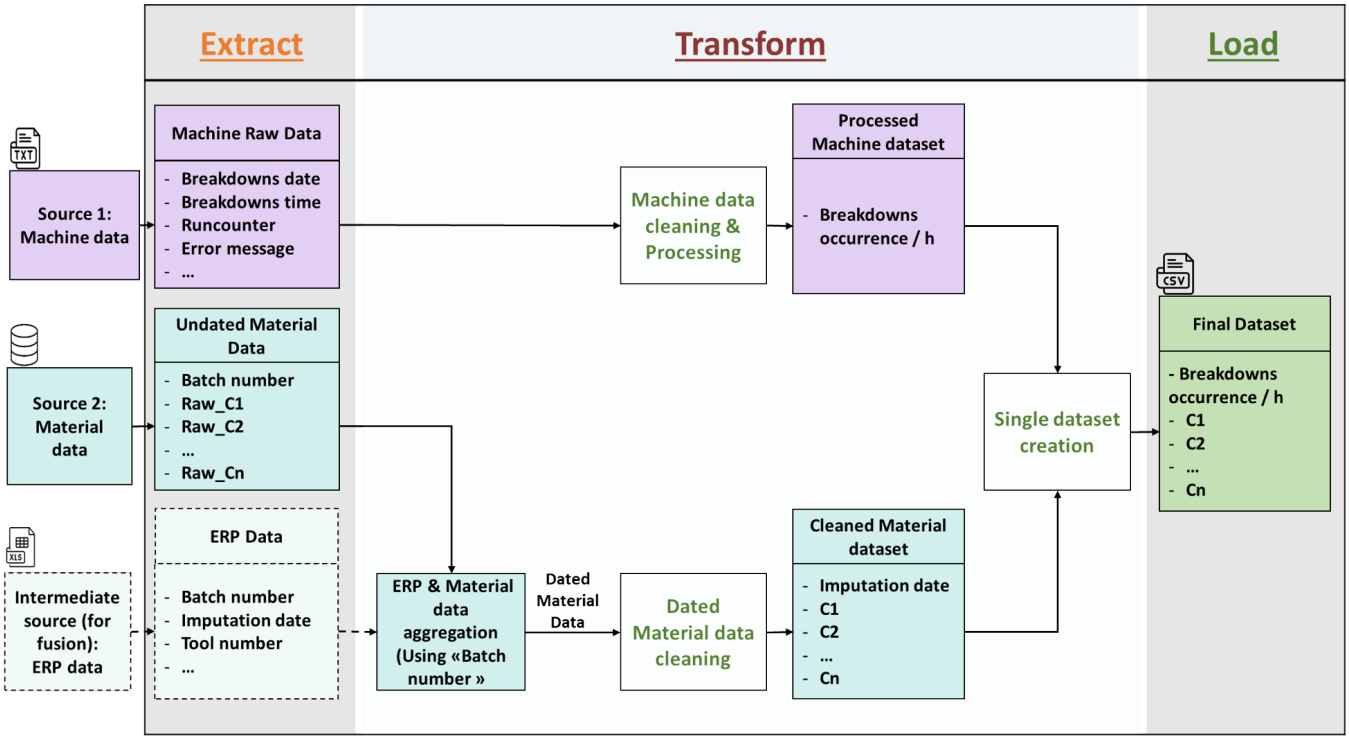

Figure 3. ETL Process for learning dataset construction

nique that functions by recursively removing attributes and constructing a model on the attributes that remain. It uses model accuracy to pinpoint which attributes (or combinations thereof) are most influential in predicting the target variable. Meanswhile, ANOVA is a statistical method employed to compare the means of three or more samples to discern if at least one of the sample means significantly deviates from the others. This method is frequently applied in feature selection to assess if numerical features exhibit significant differences across various categories. A detailed exposition of these two techniques and their operational mechanisms is provided in the appendix section. After evaluating the three mentioned methods for feature selection and model construction, Lasso was identified as the most pertinent method within this context. The Lasso method is a regularization technique that aims to shrink the coefficients of less important features to zero. This process is beneficial for reducing the dimensionality of the dataset and selecting the most relevant features for the predictive model. In the context of our study, where we aim to predict the O_r variable using the features C_1 through C_n, Lasso can be especially beneficial in reducing variance, preventing overfitting, and pinpointing the most relevant features by shrinking coefficients of some of them to zero. Mathematically, Lasso regression is defined by minimizing the sum of squared residuals (SSE) with a penalty on the absolute value of the coefficients:

$$\text{minimize} \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} |w_j| \right) \tag{2}$$

where:

- y_i are the observed values of O_r.
- $\hat{y}_i$ are the predicted values of O_r.
- w_j are the coefficients associated with features $\mathbf{C_1}$ to $\mathbf{C_n}$.
- p is the number of features.
- α is a regularization parameter. When $\alpha = 0$, Lasso reduces to plain linear regression. With a higher value of α, coefficients can be shrunk to zero, effectively eliminating those features from the model.

In the analysis of the dataset, the Lasso regularization technique was employed to discern the significance of each feature in predicting the O_r variable. The Lasso method is renowned for its ability to handle complex data by efficiently managing situations where the number of variables significantly exceeds the number of observations. This technique is particularly valuable in fields where large, multidimensional datasets are common as it offers an elegant solution to the overfitting issue that can arise from an overly complex model. By eliminating unnecessary variables, Lasso enables the construction of a simpler and more interpretable model while maintaining high predictive performance. In comparison to

ANOVA, which analyzes variances between different groups without necessarily reducing the dimensionality of the data, and RFE, which eliminates less significant variables without incorporating a direct regularization mechanism, Lasso proves to be a superior approach to feature selection. By integrating characteristic selection with regularization, Lasso achieves a balance between model simplicity and its ability to generalize to new material data inputs. To validate Lasso's effectiveness, we conducted a direct comparison with ANOVA and RFE, using model accuracy as the primary criterion. This approach demonstrated that Lasso not only reduces model complexity but also improves accuracy, confirming its superiority in our specific data context. With its ability to identify and retain only the most influential variables for prediction, Lasso is a promising method for researchers and practitioners seeking optimal efficiency and high precision in their predictive models. By comparing these three feature selection techniques, the we aim to identify the most informative features that exhibit a strong association with the target variable. These selected features will subsequently be used for building predictive models or conducting further analyses to gain insights into the underlying relationships between the features and the target variable.

3.1.4. Modeling

Predicting machine breakdowns using selected features extracted through advanced feature extraction techniques requires a performed model selection strategy (Minh, Wang, Li, & Nguyen, 2022). This selection is contingent upon several factors closely aligned with the problem's nature and the data's inherent characteristics. The primary step involves understanding the nature of the problem. In our context, the task necessitates a non-linear modeling approach, as evidenced by the data dynamics. The data volume significantly impacts the selection process. A large dataset may cause certain models to exhibit lethargic performance, leading to protracted training durations. Therefore, it is essential to select models that efficiently handle voluminous data without sacrificing predictive precision. Additionally, the model's complexity demands astute consideration. While intricate models have the prowess to encapsulate nonlinear relationships, they are also predisposed to overfitting, wherein the model becomes unduly specialized to the training dataset. Striking an optimal balance between model complexity and generalization capacity is of utmost importance (Montesinos López, Montesinos López, & Crossa, 2022). Model interpretability also plays a crucial role in the decision-making process. If elucidating and justifying the model's predictions is a priority, simpler models, like linear regression or decision trees, emerge as prime contenders due to their transparency in decision rationale. Our initial steps incorporated an exhaustive exploratory data analysis, investigating facets such as variable correlation, linearity, and distribution patterns. The insights derived indicated a non-linear association between our features and the target variable. Consequently, the necessity for non-linear models was accentuated. Given the inferred non-linearity, traditional linear regression might not be the ideal candidate for our case. We conducted a thorough analysis of three different ML models: Random Forest (RF), Artificial Neural Networks (ANN), and Support Vector Machine (SVM). The aim of this analysis was to identify the model that would perform the best in solving the problem at hand. To make an informed decision, we considered several factors, such as the models' architecture, implementation, and parameter settings. We provided a detailed description of each model's architecture in the appendix section, including the number of layers, neurons, and equations. To ensure the accuracy of our results, we carefully selected the optimal parameter settings for each model. These settings were determined through a series of rigorous experiments and iterations. Based on our findings, we concluded that the RF model is the most suitable for predicting machine breakdown occurrences. In fact, to evaluate and compare the performance of these ML models, we calculate their Root Mean Squared Error (RMSE) on a specifically designated test dataset. The RMSE is used to measure the square root of the average of squared differences between predicted and actual values, as presented in its equation 3:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

This process allowed us to identify the most proficient non-linear model that aligns seamlessly with our predictive objectives. RMSE serves as a reliable indicator of the models' predictive accuracy in this context.

3.1.5. Classes thresholds identification

In this study, the thresholding method involves comparing various clustering algorithms to determine suitable thresholds for labeling materials based on occurrences of machine breakdowns. The aim of this step is to create a set of classes that characterize and qualify the severity of predicted breakdowns by a model. This will help in preparing for the auto-labeling phase, which will be part of the online phase. The goal is to establish breakdowns intervals, with each interval characterizing a severity that will subsequently become a label. After structuring and preprocessing data, we obtained a dataset that contained historical breakdown data for each coil of material. We used clustering methods to predict clusters based solely on the occurrence values. We compared three methods: K-Means, DBSCAN, and Hierarchical. We chose to use K-Means as it performed the best in creating distinct and cohesive clusters that effectively categorized the severity levels of breakdown occurrences, and clearly showed the thresholds between clusters. This method was particularly effective for our dataset, striking a balance between simplicity

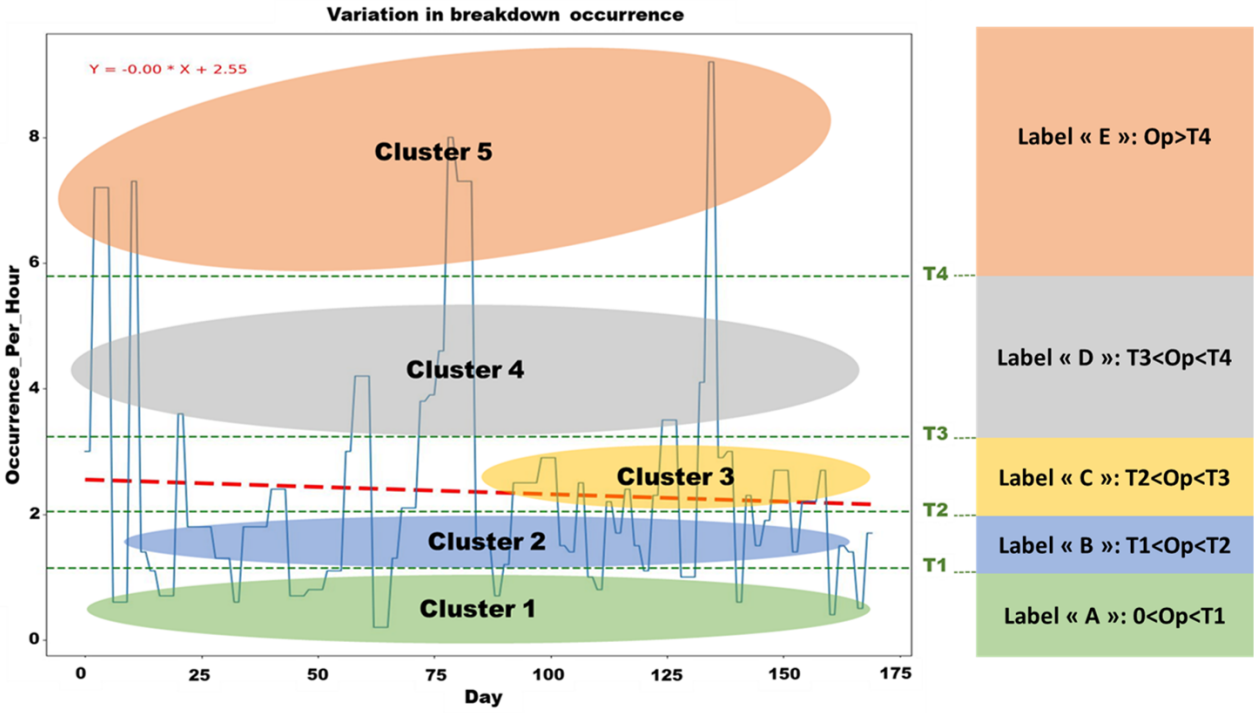

Figure 4. Classes Thresholds Identification for Labels Creation

and ability to handle the variability in the breakdown occurrence data. It provided a more straightforward basis for setting our labeling thresholds.The number of clusters, denoted as k, can be determined based on expert knowledge or other criteria, such as the Elbow method, silhouette analysis, or domain-specific insights, and has been set to 5 for our study. By applying the K-Means algorithm to the normalized data, breakdown occurrences are assigned to five clusters based on their similarity which enables us to determine the five labels denoted as **A, B, C, D, E**. When implementing K-Means clustering, we set thresholds to distinguish between clusters. We do this by calculating the centroids, which are the average values of the data points within each cluster. These thresholds can be placed between adjacent centroids along the relevant dimensions, creating clear boundaries between different levels of severity for breakdown occurrences. By analyzing the distribution of data points around these centroids and considering the variance within each cluster, we can determine the best thresholds to minimize misclassification. This allows us to accurately identify each cluster's severity level, which helps improve maintenance interventions and materials scheduling. Analyzing the resulting clusters provides insights into their characteristics, including statistical measures such as size, mean, and standard deviation. These insights facilitate the definition of thresholds, which are determined by considering factors such as the distance from the cluster mean. To validate and fine-tune the defined thresholds, recent data or pilot tests are used to ensure their accuracy in classifying new machine breakdowns. This rigorous approach ensures the effectiveness of the thresholding method in accurately classifying machine breakdowns for each coil of material in the study. Figure 4 illustrates, on one hand, the five identified clusters along with the thresholds between clusters. On the other hand, it also demonstrates how this step can be formalized into a standard for auto-labeling raw materials. This raw-data-based method for clustering non-time series data involves modifying the distance measurement method in the original clustering algorithm to suit the data, thereby preserving the most original characteristics of the data.

3.2. Online Phase: Automated Labeling of New Material Coils

This section explains the technical intricacies of our methodology's online phase, as illustrated in figure 1 and outlined by Algorithm 1. Our data acquisition process begins upon receiving new material coils, accompanied by a material certificate from the supplier, providing comprehensive physico-chemical characteristics. Emphasizing the reliability of supplier-provided data, free from errors and other issues, we directly apply Lasso method to select input features which will be an input to our predictive model. Our model is purposefully designed to provide precise predictions of breakdown occurrences per hour $\mathbf{O_p}$. Once we have this value, we assign this value to its corresponding label identified thanks

the the classes thresholds identification step described in section 3.1.5. The algorithm 1 describes the structure of the automatic labeling phase and also frames the integration of supplier data into the predictive model in order to assign them labels describing the severity of the predicted breakdown value.

Algorithm 1 New Metal Coil Auto-labeling

Input: Raw Material Data
Output: Coil Label

Initialization:
1. Get new input material (Coil): Variables from C_1 to C_{12}
2. Extract the most important features: C_i, C_j, C_k, C_l
3. Predict breakdown occurrence via deployed RF model
4. Auto-Labeling material coil:
if $Op \in [0, T1]$ **then**
    Coil is labeled A
else if $Op \in [T1, T2]$ **then**
    Coil is labeled B
else if $Op \in [T2, T3]$ **then**
    Coil is labeled C
else if $Op \in [T3, T4]$ **then**
    Coil is labeled D
else
    Coil is labeled E
end if

4. CASE STUDY AND RESULTS

This section presents a case study involving the auto-labeling of raw materials of a French company named *SCODER*, specialized in the manufacturing of metal parts in the automotive sector. It outlines the deployment of various steps to validate the efficiency and robustness of our prediction methodology. subsection 4.1 offers a detailed description of this case study, presenting the platform and the data acquisition process. And then, in subsection 4.2, we present the obtained results of material auto-labeling using the proposed methodology by comparing different ML models.

4.1. Case Study and Data Description

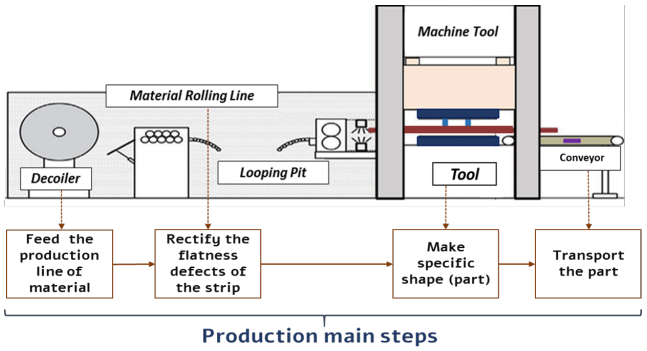

Figure 5. Case Study: Production Line Description

We consider the *SCODER* case study as a real-word application of our proposed methodology. The primary aim of this PHM project is to help industries enhance their production efficiency by minimizing machine breakdowns and increasing overall productivity. To illustrate the production process, figure 5 provides an overview of the metal parts manufacturing process. Initially, a decoiler supplies material to the metal rolling line, where flatness defects are subsequently corrected. A special tool shapes the material into the desired part, which is then transported by a conveyor. Production performance relies on several key parameters, including the material quality, machine condition, operator skill, and tool functionality. In our case study, based on expert knowledge, we attribute machine breakdowns exclusively to the characteristics of the metal coils used. Therefore, this PHM study encompasses two objectives:

- Predict the machine breakdown occurrence based on material proprieties.
- Perform an auto-labeling of raw material coils to quantify its impact on machine breakdowns.

The first step of our methodology, as described in section 3, was data gathering. Two main types of data were collected: materials data and machine history data, with an intermediate source which is ERP data. materials data contains the physical and chemical properties of metal coils. These properties are carefully recorded and stored in a dedicated SQL database. In fact, upon receiving each batch of material, we received a material certificate containing the characteristics of the material, batch number, reception date, etc. Table 1 describes the various raw material data we have. As we need to

Type	Raw Material Description	Constraints
Raw Material Characteristics	Mechanical and Chemical Characteristics	Collected $\forall$ Coil, $F_{mat} = 3 - 5$ records/day Size = 8.2 KB/record. Memory = 7MB/ Year
Features	`Batch Number`	Integer
	`Raw_`C_1	Float
	`Raw_`C_2	Float
	`Raw_`C_3	Float
	`Raw_`C_4	Float
	`Raw_`C_5	Float
	`Raw_`C_6	Float
	`Raw_`C_7	Float
	`Raw_`C_8	Float
	`Raw_`C_9	Float
	`Raw_`C_{10}	Float
	`Raw_`C_{11}	Float
	`Raw_`C_{12}	Float

Table 1. Raw Material Data Description

identify the imputation date for each metal coil, we rely on an extracted database from the ERP system, as described in table 2, where we record the consumption history for each material coil. To create a coherent dataset for materials, the ERP extracted data will be aggregated with corresponding material data during the data structuring phase, ensuring accurate and

complete integration. On the other hand, machine data were

Type	ERP data Description	Constraints
Raw ERP Data	Material consumption history	Collected $\forall$ Coil, $F_{erp} = 2 - 3$ records/day Size = 4.8 KB/record. Memory = 4MB/ Year
Features	`Batch Number`	Integer
	`Imputation Date`	Date
	`Tool Number`	Integer
	`Supplier`	Text
	`Quantity`	float

Table 2. ERP Extracted Data Description

obtained through a monitoring device equipped with a data acquisition system. This device was connected to various sensors that ensured machine protection by imposing maximum value limits and implementing process-oriented envelope curve monitoring. Whenever the envelope curve was reached, indicating a potential deviation or anomaly, the machine would automatically stop. The date and time of the breakdown, along with an error message, were recorded, as shown in table 3. The "date" variable played a crucial role in facilitating the subsequent integration of material and machine data, enabling comprehensive analysis. This rich set

Type	Raw Machine Description	Constraints
- *Raw Machine Data* (*Log Data*)	Machine functioning	F_m= 0.2 Hz Memory $\cong$ 1 GB/Year
Features	`Date`	Date
	`Time`	Time
	`Runcounter`	Integer
	`Error message`	Text
	`Speed`	Integer
	`Module`	Text
	`Position`	Integer
	`Brake angle`	Integer
	`Strokstart`	Integer

Table 3. Raw Machine Data Description.

of features provided a holistic view of the manufacturing process and formed the foundation for subsequent analysis and modeling. The integration of these datasets will be described further to provide a comprehensive understanding of the learning base that enables effective analysis and decision-making in an industrial context.

4.2. Application of the proposed methodology

In this subsection, we undertake a comprehensive exploration of how each step in the proposed methodology can impact the collected data. Additionally, we examine the different interdependencies between these stages to gain a deeper understanding of their effect on the three datasets gathered from the production line.

4.2.1. Offline phase

In subsection 3.1.1, the ETL process was proposed as a method of structuring and preprocessing data. Data relating to machines and raw materials are extracted over time, with each change of raw material coil. In this context, a coil of material is assimilated to a production cycle, during which the characteristics of the consumed material generate an occurrence of machine stoppage, as illustrated in figure 5. In detail, the characteristics of the material are extracted by merging the consumption history of the material with the supplier's material certificate database. Furthermore, the operating history of the machine provides the temporal component necessary to identify occurrences of machine stoppages, hour by hour. It is evident that this varies according to the material and over time. After extracting, transforming, and loading a unique set of learning data, we obtained the dataset described in table 4. This dataset consists of 12 inputs ranging from

Features Type	Features Description	Constraints
Inputs	C_1	Float
	C_2	Float
	C_3	Float
	C_4	Float
	C_5	Float
	C_6	Float
	C_7	Float
	C_8	Float
	C_9	Float
	C_{10}	Float
	C_{11}	Float
	C_{12}	Float
Output	O_r	Float

Table 4. Learning dataset Description

C_1 to C_{12} and a single output representing the corresponding breakdown occurrence; we have collected the history of 96 coils of material, resulting in 96 observations. In summary, we explore a 96×13 matrix presented as follows:

$$\begin{bmatrix} C_1 & C_2 & C_3 & \dots & C_{12} & O_r \\ x_{11} & x_{12} & x_{13} & \dots & x_{1,12} & y_1 \\ x_{21} & x_{22} & x_{23} & \dots & x_{2,12} & y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{96,1} & x_{96,2} & x_{96,3} & \dots & x_{96,12} & y_{96} \end{bmatrix}$$

The next step in our process involves performing an exploratory data analysis (EDA) to better understand the patterns, relationships, and discerns ML models to test.
In our dataset, the output feature O_r has a mean value of $\mu = 2.58$ and a standard deviation $\sigma = 3.822$, ranging from a minimum of 0.2 to a maximum of 35.9 with a median of 1.7. The columns C_1 to C_{12} represent various measures. Specifically, C_1 has an average of approximately 24.57 with a standard deviation of 1.025, and its values span from 21.1 to 26.2. The other columns also display diverse distributions, as detailed in their descriptive statistics. The variable O_r, repre-

senting the number of breakdowns per hour indicates variability in the production process. The statistical description of the dataset, including measures such as mean, minimum, maximum, and standard deviation for each variable, is presented in table 5. This table provides a comprehensive overview of the central tendency, dispersion, and distribution shape of the dataset's features. To quantify the statistical relationships

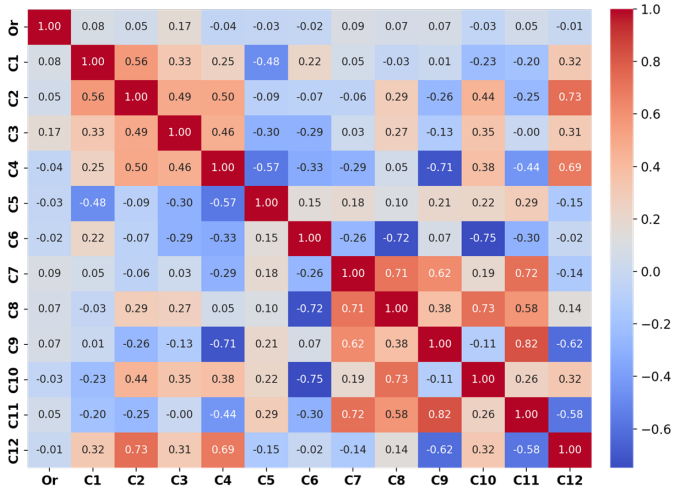

Figure 6. Correlation Matrix

between the various material features and the target variable O_r, we elaborate a correlation analysis. This helps to identify which features have a strong influence on O_r, providing valuable insights for predictive modeling and feature selection. The figure 6 displays the correlation matrix. It indicates that the variable O_r exhibits varying correlations with the other columns. This heatmap demonstrates a positive correlation between O_r, C_3 and C_7. Conversely, it shows a negative correlation with C_4, C_5, and C_{10}. others variables like C_1, C_2, and C_6 show relatively weak or negligible correlations with O_r. The presence of these varying correlations suggests that while some material features might play a substantial role in influencing O_r, others might have a minimal or no impact. This correlation matrix, though insightful, underscores the significance of other exploratory analyses, such as linearity analysis. In fact, the purpose of linearity analysis is to examine the relationship between each feature and the output variable O_r. The figure 7 displays a series of scatter plots for some material characteristics. We chose to show the relationship between some chemical characteristics (C_1, C_2, C_3, C_4) and mechanical characteristics $(C_9, C_{10}, C_{11}, C_{12})$ with the target variable O_r. For most features, the data points appear dispersed, suggesting no strong linear relationship with O_r. The color gradient, ranging from purple (low $O_r$values) to yellow (high $O_r$values), offers insights into regions with similar $O_r$values. While some areas exhibit color density, indicating consistent $O_r$values for certain feature ranges, the overall patterns hint at potentially complex, non-linear relationships. For some features, like C_2 and C_3, there is a distinct pattern where the points form a linear relationship with the output variable O_r. This indicates that these features might be good predictors for the target variable. For other features, such as C_1 and C_{12}, the relationship is not as clear, and the points are more scattered. This indicates that these features might not be as strong predictors for the target variable. These observations can be valuable when selecting features for modeling, as they can help to identify which features are most relevant to the target variable. To evaluate feature importance more precisely, we tested three different methods: Lasso, RFE, and ANOVA, which provided distinct results. It's crucial to note that the global performance of feature selection methods cannot be directly compared without applying a machine learning algorithm. In this step, we can only interpret the results of each feature selection technique with its corresponding metric, as illustrated in figure 8. In the Lasso method, features such as C_3, C_4, C_7, and C_{10} emerged as pivotal, with non-zero coefficients affirming their significance, while other features were relegated to zero, hinting at their lesser influence. The direction of the coefficients, whether positive or negative, indicates their relationship with the target variable O_r, and the magnitude delineates the feature's impact when other variables are held static. In the RFE framework, C_1 was identified as the most important feature, followed by C_5, C_8 and then other features. It is crucial to note that this interpretation is grounded in a linear perspective and might differ in a non-linear setting. Lastly, ANOVA was employed to discern the differences in means across multiple groups for every feature. However, no features were identified as significant based on the ANOVA F-test, as all p-values surpassed the usual 0.05 threshold. This urges a cautious approach in their immediate validation. To compare the results of the three feature extraction methods (ANOVA, Lasso, and RFE), we proceed directly to the modeling step. Indeed, to accurately predict the breakdowns occurrence, we test the three models mentioned in the methodology: SVM, RF, and ANN. Table 6 summarizes the nine combinations created between the ML models and the feature selection methods, also detailing the inputs chosen by each technique, the key parameters used for each method, the main conclusions, and additional remarks that provide more context regarding the nature and specifics of each approach. Testing each combination allowed us to calculate the Root Mean Squared Error (RMSE). According to table 6, it is evident that each selection method offers a distinct perspective on the importance and relevance of the features with respect to the target variable Gold. The performance metrics in ML can show which one is the most important, taking into account the context, assumptions, and specifics of each approach. In fact, results shows that machine learning models, feature selection methods, and the number of features play a crucial role in model performance. RF and SVM tend to perform similarly and generally better than ANN, maybe due to their ability to cap-

	O_r	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
count	96	96	96	96	96	96	96	96	96	96	96	96	96
mean	2.580	24.570	537.631	625.816	0.076	1.372	0.002	0.008	0.038	0.043	0.051	0.001	0.014
std	3.822	1.025	14.911	10.993	0.002	0.026	0.001	0.001	0.014	0.004	0.003	0.000	0.002
min	0.2	21.1	517	596	0.071	1.36	0.001	0.006	0.021	0.039	0.047	0.001	0.012
25%	1.1	23.9	525	620	0.074	1.36	0.001	0.007	0.024	0.041	0.048	0.001	0.012
50%	1.7	24.7	530	628	0.076	1.37	0.002	0.008	0.041	0.043	0.051	0.001	0.012
75%	2.8	25.4	552	633	0.077	1.37	0.002	0.009	0.046	0.046	0.053	0.002	0.015
max	35.9	26.2	569	638	0.08	1.51	0.003	0.01	0.058	0.049	0.056	0.002	0.017

Table 5. Data Statistical Analysis Description

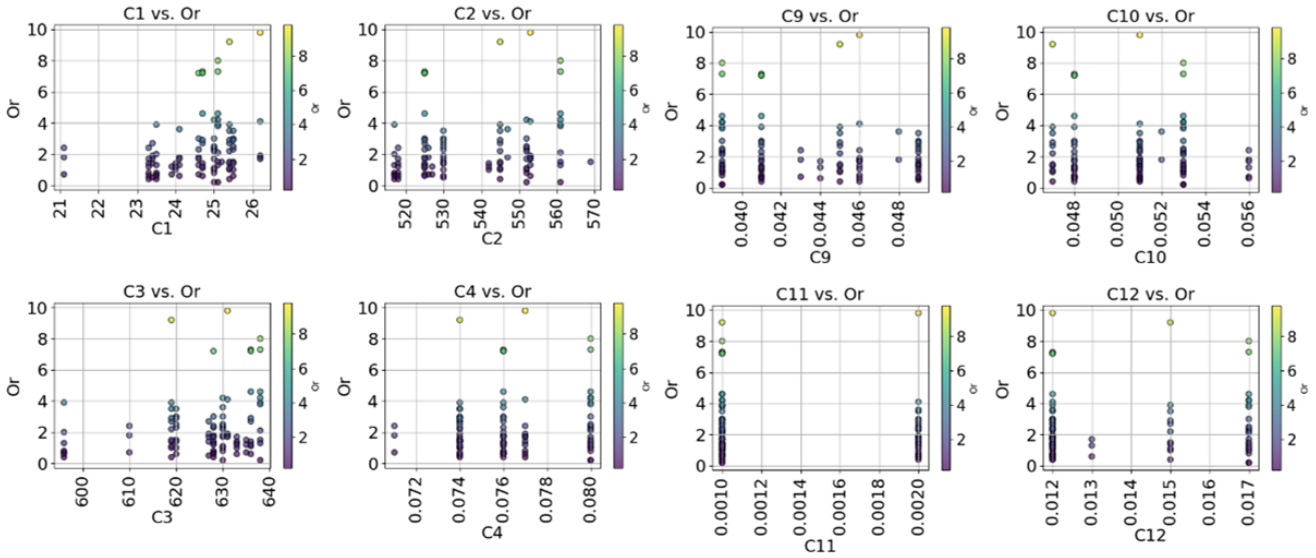

Figure 7. Linearity Analysis

ture complex non-linear relationships in data. In terms of feature selection methods, ANOVA and Lasso provide similar performance, while RFE results in slightly worse outcomes. This discrepancy might be due to the different ways these methods select important features, with RFE potentially choosing less informative features. The number of features does not show a clear trend in impact on performance; in some cases, using selected features (C_1 to C_{12}) yields better results, while in other cases, a subset of features performs better. This suggests that some features are more informative than others, and including non-informative features can degrade model performance. The parameters of the models also significantly impact performance; for instance, the maximum number of iterations for ANN affects model convergence, and the number of estimators and random state for RF can also influence outcomes. These results highlight the importance of testing different combinations of machine learning models, feature selection methods, and the number of features to find the optimal configuration for a given dataset. The best combination found in this analysis was the RF model with Lasso as the feature selection method with 4 features (C_3, C_4, C_7, C_{10}). This combination yielded the lowest RMSE of 1.7, indicating superior predictive performance compared to other combinations. The RF model is known for its ability to handle complex data structures and relationships, and the Lasso method efficiently identified the most relevant features, contributing to the model's high accuracy. These results emphasize the potential of combining machine learning models with appropriate feature selection methods to optimize occurrence prediction. The presented development allows us to predict on occurrence by material coil; a final step remains before proceeding to the deployment of the method, which is to create the labels. Indeed, the labels are created by comparing the three clustering methods: DBSCAN, Hierarchical, and K-means. This clustering technique was applied to our dataset spanning 93 days to segment data based on O_r variable. The results, presented in figure 9, revealed distinct cluster structures with specific thresholds between each cluster. For the DBSCAN method, it was not possible to specify the number of clusters in advance, as this method works by identifying high-density areas in the feature space, which are separated by low-density areas. In our case, using the default parameters of DBSCAN, we obtained a single cluster, indicating that data are very grouped in the feature space. We proceeded

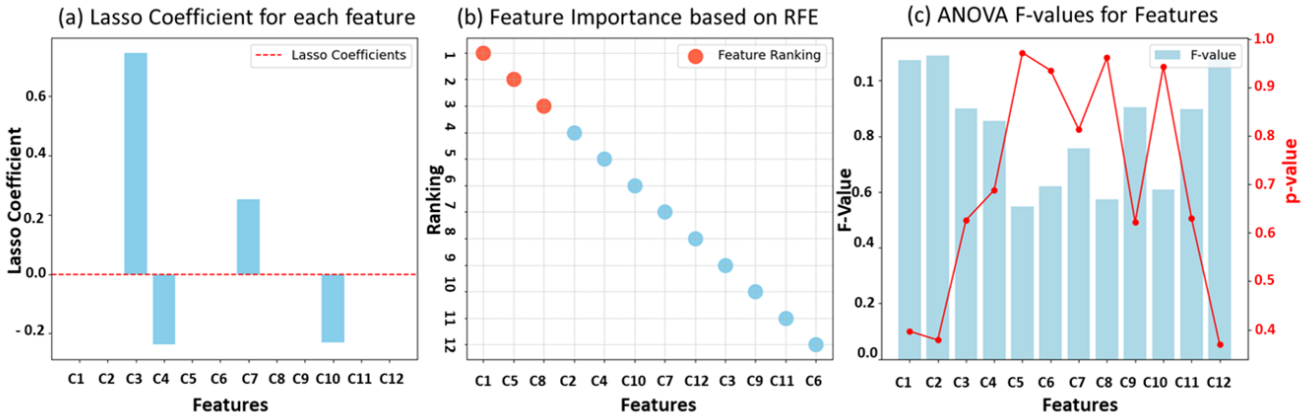

Figure 8. Feature Selection Methods Results

Combination	Feature Selection Method	ML Model	Inputs	Output	Parameters	RMSE
1	Lasso	SVM	C_3, C_4, C_7, C_{10}	O_r	-	1.308
2	Lasso	ANN	C_3, C_4, C_7, C_{10}	O_r	max_iter=500	16.56
3	Lasso	RF	C_3, C_4, C_7, C_{10}	O_r	n_estimators=100, random_state=0	1.70
4	RFE	SVM	C_1, C_5, C_8	O_r	-	3.08
5	RFE	ANN	C_1, C_5, C_8	O_r	max_iter=500	13.28
6	RFE	RF	C_1, C_5, C_8	O_r	n_estimators=100, random_state=0	3.414
7	ANOVA	SVM	C_1 to C_{12}	O_r	-	3.308
8	ANOVA	ANN	C_1 to C_{12}	O_r	max_iter=500	13.16
9	ANOVA	RF	C_1 to C_{12}	O_r	n_estimators=100, random_state=0	1.968

Table 6. RMSE Comparison of ML Models

to adjust the parameters of DBSCAN in an attempt to obtain a specific number of clusters, but this gave us 12 clusters, which is not a satisfactory result in our context. This situation illustrates the limitations of DBSCAN when it comes to controlling the number of clusters, especially in datasets where the structure of clusters is not clearly defined. K-Means identified five clusters with thresholds at 0.9, 1.7, 2.15, 2.85, and 3.83, and the silhouette score was 0.61. Hierarchical clustering produced different results to K-Means with thresholds at 1.2, 1.3, 2.4 and 3.3 and a silhouette score of 0.58 . These thresholds represent transition points between different data groups and can be used to understand the relationships between different observations in the dataset. The silhouette scores suggest that K-Means clustering produced the most cohesive and separated clusters, followed closely by hierarchical clustering. Given the nature of our problem, which is auto-labeling, we have chosen K-means as it has the highest silhouette score. Indeed, in our case, clustering is not performed for prediction, so a simple preferment clustering method was sufficient. We are mainly interested in the precision of the thresholds between clusters to ensure accurate auto-labeling. The selection of K-means as the best clustering method is based on our specific objective and preferences. The fact that K-means provided a better choice confirms that sometimes simple ML models can deliver good results. The key is that they match our dataset and meet our needs without being overqualified.

4.2.2. Online phase

In the previous subsection 4.2.1, we applied the offline phase of our methodology to the *SCODER* case study. This allowed us to establish the different techniques to be deployed in the online phase. After conducting our experiments and developing an appropriate RF model for the, we can now discuss the findings related to predicting a global occurrence in the manufacturing industry. We randomly selected 20 coils of material from the historical material records. For each coil, we extracted the four relevant features, namely C_3, C_4,C_7, and C_{10} identified by the Lasso model and used these features as inputs to test our model in order to predict the target variable's value O_r. Figure 10 presents a comparison between the real O_r value and the values predicted by our model O_p. The real values are represented in blue, and the predicted ones are represented in red. This visualization allows us to assess the

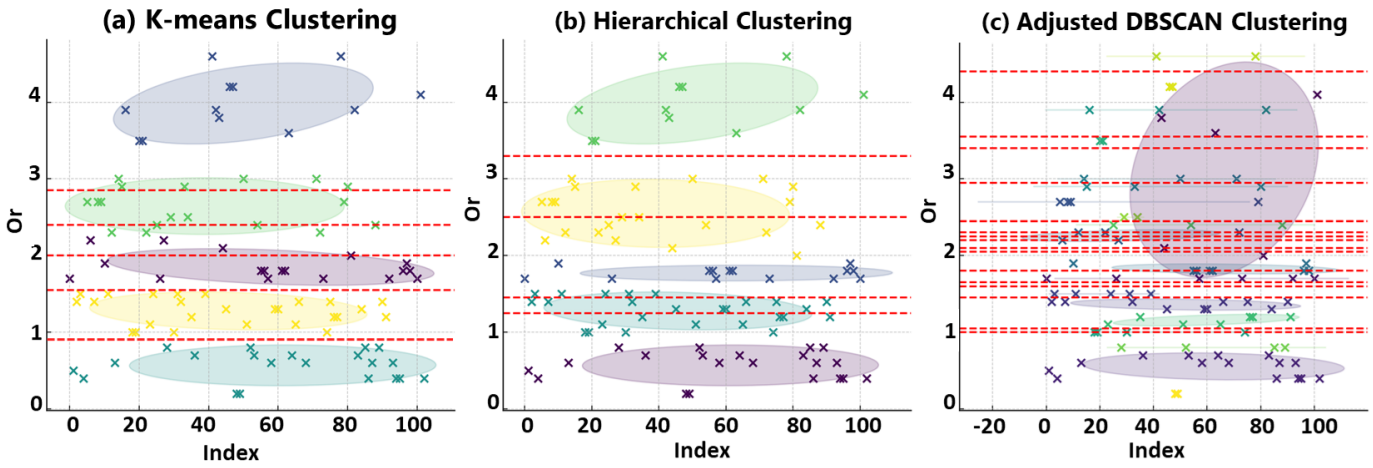

Figure 9. Thresholds Identifications

accuracy of our model by comparing the proximity of the two curves. The accuracy of the model was also measured using the following formula:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\text{Real Value}_i - \text{Predicted Value}_i}{\text{Real Value}_i}\right| \quad (4)$$

where n is the number of samples, and Real Value$_i$ and Predicted Value$_i$ are respectively the real and predicted values for the i-th sample. This formula provides a measure of the relative error between the predicted and real values, and higher accuracy indicates better performance of the model. When the expert receives the label made by a the prediction methodology, he plays a crucial role in using this information to efficiently plan raw material consumption. The model's decision serves as a valuable guide, but human expertise is still essential to interpret these data in the specific context of the ongoing operation. The expert must consider not only the model's results, but also other factors such as the machine's operating conditions, variations in the tool shape and any production constraints. By integrating these elements with the model's predictions, the expert can make informed decisions that maximize production efficiency while minimizing the risks of machine breakdowns. This holistic approach, which combines the power of data analysis with human expertise, is essential to ensure that raw material planning is optimized and machine stoppages are minimized, thereby achieving production goals while preserving the quality of the production process and final product. The developed methodology achieved a Technology Readiness Level (TRL) of 7. This is a significant milestone for companies that are looking to optimize their usage of raw materials. This helps production to adapt to the dynamic and changing demands of the industrial environment, resulting in minimized interruptions and breakdown risks while maximizing resource use. In practical terms, this advancement results in a precisely labeled raw material inventory at *Scoder*, facilitating a more targeted and effective consumption of coils. In fact, labeling materials into categories holds strategic importance for our production chain, guiding scheduling and risk management. Materials categorized as **(A)** and **(B)**, expected to have fewer breakdowns, are scheduled for nighttime production due to limited technical support. Conversely, materials in **(D)** and **(E)**, known for complexity or higher fault tendency, are assigned to skilled operators during the day, reducing risks. If incoming materials fail to meet the required standards, they can be quickly returned to the supplier, avoiding waste and reducing costs. By replacing the old model of random selection or FIFO (First In First Out) selection with a strategic material allocation, this methodology represents a significant step

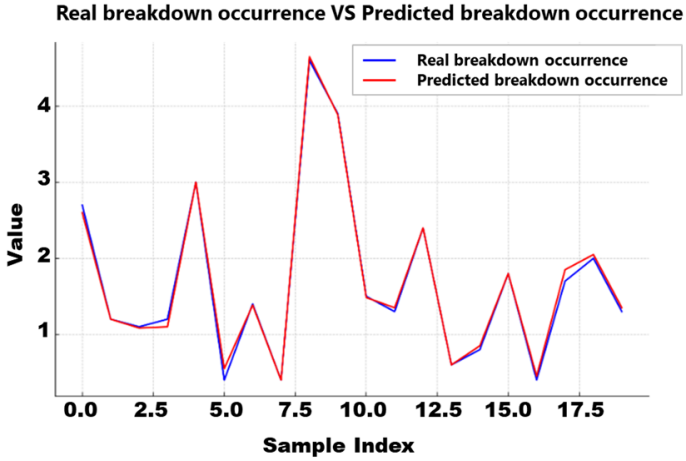

Figure 10. Predicted Breakdown Occurrence VS Real Breakdown Occurrence

forward in industrial planning. It aligns production processes with both sustainability and efficiency goals.

5. CONCLUSION

This study presents a comprehensive methodology aimed at addressing the crucial challenge of integrating raw material characteristics into the prognostics and health management (PHM) framework for predictive maintenance in the manufacturing industry. By acknowledging the pivotal role of raw materials in influencing machine performance and reliability, our methodology goes beyond conventional sensor-based diagnostics and prognostics, thus offering a holistic solution. The implementation of the ETL process has successfully resulted in the development of a real-time data fusion and preprocessing framework. This framework enables the seamless integration of raw material data, allowing for accurate material labeling and subsequent predictive analysis. Through this, our methodology facilitates intelligent raw material management, minimizing machine breakdowns, and enhancing production reliability. Our comparative analysis of feature selection methods revealed that the Lasso technique When combined with a RF model demonstrated superior performance in accurately predicting material impact on machine breakdowns. This underscores the importance of feature selection in enhancing the predictive power of machine learning algorithms. Applying our methodology to a real-world metal material dataset from *SCODER*, we successfully labeled metal coils into distinct categories based on their characteristics. This labeling not only aids in intelligent inventory management but also contributes to the proactive anticipation of machine breakdowns, thus minimizing production disruptions. The combination of ETL, Lasso, and RF offers a powerful toolset for manufacturers to optimize raw material consumption, reduce breakdowns, and boost production efficiency. As industries continue to navigate the challenges of modern manufacturing, this methodology holds the potential to redefine the landscape of predictive maintenance and contribute to sustainable and resilient manufacturing processes. A significant challenge of our work is to fully integrate raw material characteristics with predictive analytics to enhance the PHM systems. This complexity stems from the varying quality of raw materials and its nuanced impact on machine performance, which is difficult to quantify and incorporate into existing models. For future work, we aim to focus on predicting the impact of raw material on part quality. This involves developing more sophisticated models that can accurately capture the relationship between raw material characteristics and the global quality of manufactured parts. Such models will not only advance our understanding of material science but also improve manufacturing processes by enabling more precise control over the accuracy of the output, reducing waste, and increasing efficiency.

ACKNOWLEDGMENT

The work was carried out with the support of *SCODER*, a small to medium-sized enterprise located in Pirey, France, as well as SupmicroTech-ENSMM in Besançon, France. The project has been co-financed by the National Research Agency (ANR) as part of the "France Relance" plan, a program aimed at preserving the human research and development (R&D) capacities of companies, while making young graduates and doctors available to businesses. The authors would like to express their sincere gratitude to all parties involved for their invaluable contribution to this research work. Their involvement, whether financial, scientific, or technical, greatly enhanced the quality and efficiency of the research.

REFERENCES

Abd Al Rahman, M., & Mousavi, A. (2020). A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry. *Ieee Access*, *8*, 183192–183271.

Abiodun, E. O., Alabdulatif, A., Abiodun, O. I., Alawida, M., Alabdulatif, A., & Alkhawaldeh, R. S. (2021). A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. *Neural Computing and Applications*, *33*(22), 15091–15118.

Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhouib, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in industry 4.0: Overview, models, and challenges. *Applied Sciences*, *12*(16), 8081.

Ahmad, M., Kamiński, P., Olczak, P., Alam, M., Iqbal, M. J., Ahmad, F., ... Khan, B. J. (2021). Development of prediction models for shear strength of rockfill material using machine learning techniques. *Applied Sciences*, *11*(13), 6167.

Arshadi, M., Gref, R., Geladi, P., Dahlqvist, S.-A., & Lestander, T. (2008). The influence of raw material characteristics on the industrial pelletizing process and pellet quality. *Fuel processing technology*, *89*(12), 1442–1447.

Ayed, M. B., Soualhi, M., Mairot, N., Giampiccolo, S., Ketata, R., & Zerhouni, N. (2023). Explainable prediction of machine-tool breakdowns based on combination of natural language processing and classifiers. In *Intelligent systems conference* (pp. 105–121).

Benaggoune, K., Mouss, L. H., Abdessemed, A., & Bensakhria, M. (2020). Holonic agent-based approach for system-level remaining useful life estimation with stochastic dependence. *International Journal of Computer Integrated Manufacturing*, *33*(10-11), 1089–1104.

Borràs-Ferrís, J., Palací-López, D., Duchesne, C., & Ferrer,

A. (2022). Defining multivariate raw material specifications in industry 4.0. *Chemometrics and Intelligent Laboratory Systems*, *225*, 104563.

Chen, K.-C., & Gao, Z.-J. (2020). Integrated group-based valuable sensor selection approach for remaining machinery life estimation in the future industry 4.0 era. In *2020 international symposium on vlsi design, automation and test (vlsi-dat)* (pp. 1–4).

Farahat, A., Gupta, C., et al. (2020). Similarity-based feature extraction from vibration data for prognostics. In *Annual conference of the phm society* (Vol. 12, pp. 10–10).

Gelzinis, A., Verikas, A., Vaiciukynas, E., Bacauskiene, M., Minelga, J., Hållander, M., ... Padervinskis, E. (2014). Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders. In *2014 ieee symposium on computational intelligence in healthcare and e-health (cicare)* (pp. 125–132).

Ketata, F., Al Masry, Z., Zerhouni, N., & Yacoub, S. (2023). Explainable machine learning approach with augmentation for mortality prediction. In *2023 ieee international conference on advanced systems and emergent technologies (ic_aset)* (pp. 01–06).

Li, C., Sanchez, R.-V., Zurita, G., Cerrada, M., Cabrera, D., & Vásquez, R. E. (2016). Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals. *Mechanical Systems and Signal Processing*, *76*, 283–293.

Li, Y., Zhang, Q., Zhu, Y., Yang, A., Liu, W., Zhao, X., ... others (2022). A model study on raw material chemical composition to predict sinter quality based on ga-rnn. *Computational Intelligence and Neuroscience*, *2022*.

Mera-Gaona, M., López, D. M., Vargas-Canas, R., & Neumann, U. (2021). Framework for the ensemble of feature selection methods. *Applied Sciences*, *11*(17), 8122.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 1–66.

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109–139). Springer.

Mosavi, A., Hosseini, F. S., Choubin, B., Abdolshahnejad, M., Gharechaee, H., Lahijanzadeh, A., & Dineva, A. A. (2020). Susceptibility prediction of groundwater hardness using ensemble machine learning models. *Water*, *12*(10), 2770.

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., & Zerhouni, N. (2021). Towards an adapted phm approach: Data quality requirements methodology for fault detection applications. *Computers in industry*, *127*, 103414.

Popescu, T. D., & Aiordachioaie, D. (2019). Fault detection of rolling element bearings using optimal segmentation of vibrating signals. *Mechanical Systems and Signal Processing*, *116*, 370–391.

Rahman, M. M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S., & Razak, R. A. (2020). A review of machine learning methods of feature selection and classification for autism spectrum disorder. *Brain sciences*, *10*(12), 949.

Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, *3*(1), 54.

Raouf, I., Khan, A., Khalid, S., Sohail, M., Azad, M. M., & Kim, H. S. (2022). Sensor-based prognostic health management of advanced driver assistance system for autonomous vehicles: A recent survey. *Mathematics*, *10*(18), 3233.

Raouf, I., Lee, H., & Kim, H. S. (2022). Mechanical fault detection based on machine learning for robotic rv reducer using electrical current signature analysis: A data-driven approach. *Journal of Computational Design and Engineering*, *9*(2), 417–433.

Rizzo, A., Goel, S., Luisa Grilli, M., Iglesias, R., Jaworska, L., Lapkovskis, V., ... Valerini, D. (2020). The critical raw materials in cutting tools for machining applications: A review. *Materials*, *13*(6), 1377.

Soualhi, M., Nguyen, K. T., Soualhi, A., Medjaher, K., & Hemsas, K. E. (2019). Health monitoring of bearing and gear faults by using a new health indicator extracted from current signals. *Measurement*, *141*, 37–51.

Soualhi, M., Soualhi, A., Nguyen, K. T., Medjaher, K., Clerc, G., & Razik, H. (2023). Open heterogeneous data for condition monitoring of multi faults in rotating machines used in different operating conditions. *International Journal of Prognostics and Health Management*, *14*(2).

Stauffer, F., Vanhoorne, V., Pilcer, G., Chavez, P.-F., Vervaet, C., & De Beer, T. (2019). Managing api raw material variability in a continuous manufacturing line–prediction of process robustness. *International Journal of Pharmaceutics*, *569*, 118525.

Winursito, A., Hidayat, R., Bejo, A., & Utomo, M. N. Y. (2018). Feature data reduction of mfcc using pca and svd in speech recognition system. In *2018 international conference on smart computing and electronic enterprise (icscee)* (pp. 1–6).

Zhang, Y., Sheng, M., Liu, X., Wang, R., Lin, W., Ren, P., ... Song, W. (2022). A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. *Health Information Science and Systems*, *10*(1), 22.

APPENDIX

ANNEX 1: USED MACHINE LEARNING MODELS

1. **Random Forest (RF):** The RF algorithm combines numerous decision trees to render predictions (Gelzinis et al., 2014), as shown in figure 11. It excels at understanding complex nonlinear connections and handling datasets with multiple dimensions. In RF, $B = 100$ trees are

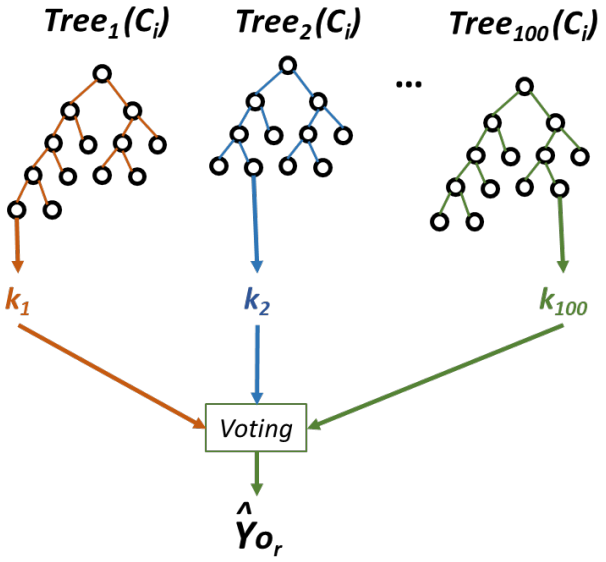

Figure 11. RF Architecture

constructed using bootstrap samples from the dataset, and for each tree, predictions are either averaged for regression or a majority vote is taken for classification. The prediction of the b^{th} tree for a given set of inputs $C_1, \ldots, C_{12}$ is denoted as $Y_b(C_1, \ldots, C_{12})$. The O_r prediction of the RF model, denoted as $\hat{Y}_{Or}$, is calculated as follows:

$$\hat{Y}_{Or} = \frac{1}{100} \sum_{b=1}^{100} Y_b(C_1, \ldots, C_{12}) \tag{5}$$

In this study, the RF algorithm operates with 12 input features derived from observations of material properties, labeled C_1 to C_{12}. These features are selected during the feature selection stage and are integrated into a training database that reflects the actual occurrences or outcomes for each set of material features. Each entry in this training dataset corresponds to a decision tree, and the RF uses multiple such decision trees to make predictions. The process involves using samples and combining the outputs of these decision trees to arrive at a final prediction. The RF model can be configured to use different features, and the number of decision trees in the forest is set to 100 to optimize both performance and computational efficiency.

2. **Artificial Neural Networks (ANN):** ANN model consists of interconnected nodes or neurons organized into layers, where each neuron processes information and contributes to the network's overall computation. ANNs are renowned for their ability to model complex, nonlinear relationships in data, although they often require substantial data and computational resources to achieve optimal performance. A neuron's output in our specific setup is typically represented by:

$$\hat{Y}_{Or} = f\left(\sum_{i=1}^{12} w_i C_i + b\right) \tag{6}$$

where w_i are weights, C_i are the input features derived from observations of material properties, b is the bias, and f is an activation function. In our specific case study, we adapt the ANN architecture, presented in figure 12, to work with these input features. These inputs, labeled C_1 to C_{12}, are selected during the feature selection stage to ensure they are relevant to our problem. We can configure the ANN to accommodate different sets of features, allowing us to explore various combinations and their impact on the model's performance.

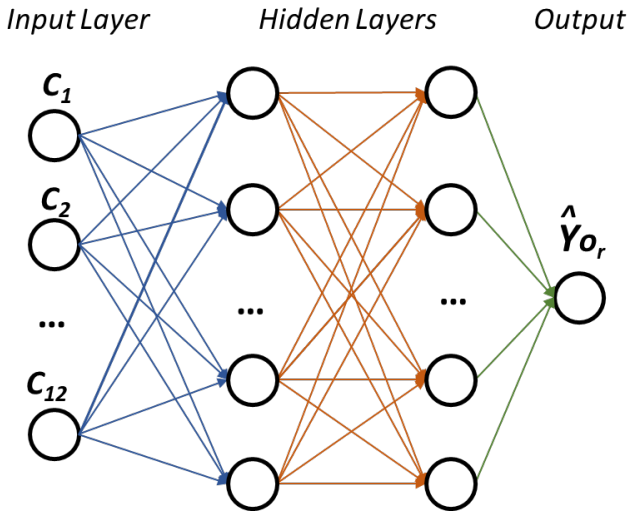

Figure 12. ANN Architecture

Additionally, the number of neurons and layers in the network can be adjusted based on the number and complexity of the selected features. By fine-tuning these architectural elements and training the network on our material property data, we can harness the power of ANNs to uncover intricate relationships and make accurate breakdown occurrence predictions.

3. **Support Vector Machine (SVM):** The SVM is a supervised learning algorithm adept at classification and re-

gression tasks. The SVM architecture, presented in figure 13, is designed to find an optimal hyperplane that minimizes the errors in predicting continuous outcomes. The decision function for SVM regression, in our case study, can be expressed as:

$$\hat{Y}_{Or} = \langle w, \mathbf{C} \rangle + b \tag{7}$$

where $\langle w, \mathbf{C} \rangle$ is the dot product of the weight vector w and the input vector $\mathbf{C} = (C_1, C_2, \ldots, C_{12})$, and b is the bias. By employing a suitable kernel function, such as the linear, polynomial, or radial basis function (RBF) kernel, and fine-tuning the model's hyperparameters, we can train the SVM to perform regression on specific material properties effectively. This approach aims to predict a continuous output based on the linear or non-linear relationships learned from the input features.

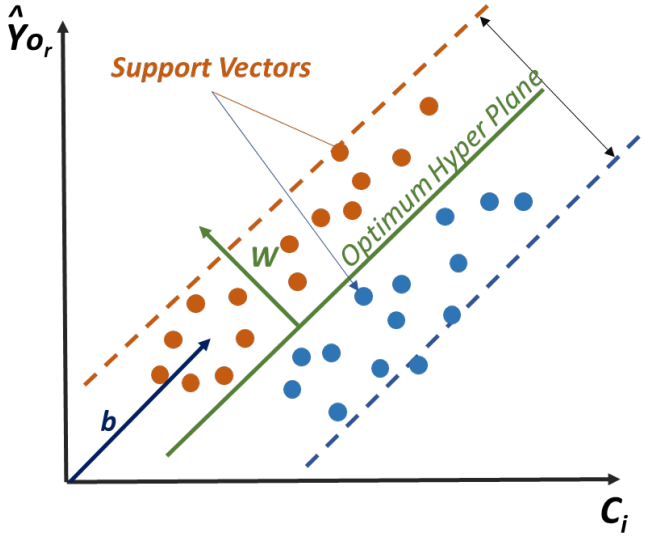

Figure 13. SVM Architecture

4. **DBSCAN Clustering**: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering method that defines clusters as continuous regions of high density. It does not require the number of clusters to be specified in advance. Instead, it operates based on two parameters: a distance value ('epsilon') and the minimum number of points required to form a dense region ('minPts'). For a given point p:
   - If there are at least 'minPts' points within an 'epsilon' distance of p, then p is a core point.
   - If p is within an 'epsilon' distance of another core point, it's a border point.
   - Otherwise, p is a noise point.

   The equation that characterizes density in the DBSCAN context is given by:

$$D(p, \epsilon) = \{q \in D | \text{dist}(p, q) \leq \epsilon\} \tag{8}$$

   where $D(p, \epsilon)$ denotes the number of points within an 'ϵ' distance of p and 'dist(p, q)' is a distance function (typically Euclidean) between points p and q.

5. **Hierarchical Clustering**: Hierarchical clustering builds a tree of clusters. The approach can be either agglomerative (bottom-up) or divisive (top-down). In the agglomerative approach, every data point starts as its own cluster, and pairs of clusters are merged based on their similarity until only one large cluster remains. The distance between two clusters A and B in the average linkage method is given by:

$$d(A, B) = \frac{1}{|A| \times |B|} \sum_{a \in A} \sum_{bin B} \text{dist}(a, b) \tag{9}$$

   where $|A|$ is the size of cluster A, $|B|$ is the size of cluster B, and 'dist(a, b)' is a distance function (typically Euclidean) between points a and b.

6. **Recursive Feature Elimination (RFE)**: Recursive Feature Elimination is a feature selection method employed in the study. It iteratively eliminates less important features based on the coefficients of a chosen machine learning algorithm. Starting with the full feature set, the model ranks the features based on their importance and eliminates the least significant ones. The RFE process can be mathematically expressed as:

$$F_{\text{selected}} = \underset{F \subseteq F_{\text{all}}}{\text{argmax}}\ J(F) \tag{10}$$

   where F_{selected} are the selected features, F_{all} is the full feature set, and $J(F)$ is the performance metric of the model trained using features F.

7. **ANOVA (Analysis of Variance)**: ANOVA, a statistical method, was also used for feature selection. It evaluates the relationship between each feature and the target variable by calculating an F-statistic:

$$F = \frac{\text{explained variance between groups}}{\text{unexplained variance within groups}} \tag{11}$$

   A p-value is then determined to assess the likelihood that the observed differences in means are due to chance. Features with lower p-values are considered more significant and are selected for further analysis.