Data-Driven Quality Prognostics for Automated Riveting Processes

Sara Ferreira Pereira
sara.f.pereira@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Portugal
November 2017

Abstract— **Technologies based in robotics and automatics are reshaping the aerospace industry. Aircraft manufacturers and top-tier suppliers now rely on robotics to perform most of its operational tasks. Over the years, a succession of implemented mobile robots has been developed with the mission of automating important industrial processes such as welding, material handling or assembly procedures. However, despite the progress achieved, a major limitation is that the process still requires human supervision and an extensive quality control process. An approach to address this limitation is to integrate machine learning methods within the quality control process. The idea is to develop algorithms that can direct manufacturing experts towards critical areas requiring human supervision and quality control. In this paper we present an application of machine learning to a concrete industrial problem involving the quality control of a riveting machine. The proposal consists of an intelligent predictive model that can be integrated within the existing real time sensing and pre-processing sub-systems at the equipment level. The framework makes use of several data-driven techniques for pre-processing and feature engineering, combined with the most accurate algorithms, validated through k-folds cross validation technique which also estimates prediction errors. The model is able to classify the manufacturing process of the machine as nominal or anomalous according to a real-world data set of design requirements and operational data. Several machine learning algorithms are compared such as linear regression, nearest neighbor, support vector machines, decision trees, random forests and extreme gradient boost. Results obtained from the case study suggest that the proposed model produces accurate predictions which meet industrial standards.**

Keywords: Prognostics, Machine Learning, Data-driven, Manufacturing, Aeronautics

1. INTRODUCTION

"Modern industries have embraced the dawn of a databased epoch due to the extreme difficulty in obtaining the physical models for complicated processes."[1]

In aeronautics, there is growing pressure to automate and improve the existing manufacturing processes [2]. The increasing aircraft backlog, the strong competition and the new lines reinforce the need to automate. Setting the change are also the ergonomic issues and the high productivity and quality demands. The integration of autonomous robotic systems within the industrial environment has brought several benefits, including the improvement of employee safety, increased production rate and a better product quality [3], [4]. Often used to perform functions that were dangerous, unsuitable or too repetitive for human operators, automation systems came to bring increased payload and speed up the traditional processes of the industry.

Despite the technological evolution of riveting machines, the manufacturing process still requires a great deal of human supervision and automatic control. The integrity of the aircraft structural joints continues to rely, to a great extent, on the experience of trained technicians. Here it would be important to have aiding predictive systems. Predictive manufacturing is probably one of the most revolutionary technologies that the Industrial Internet of Things/Industry 4.0 brings. Being able to take the results of a prognostics system and drive operational activity can reduce the need for reactive maintenance, produce higher quality products, and enable increased availability and performance. Methods such as machine learning can play an important role leading to more effective operations that minimize cost and increase profit at the same time.

The author aim to show that prognostics (future prediction) of the riveting process can be developed based on machine log data. Nowadays there are already several applications of these techniques that improve society's quality of life, particularly in industrial plants, where the automation can relive the workload sustained by operators, encouraging a better allocation of human resources. On the other hand, deploying machine learning techniques for monitoring the quality of the process makes the assessment less prone to human error. This paper reports the usage of data-driven methods combined with machine learning algorithms and domain knowledge to expedite and improve the diagnostics of the quality of the aircraft skin assembly industrial process.

It is foreseen that human factors will continue to play a central role in aviation quality control and safety. Nevertheless, as the skill requirements for avionics technicians increase, the need for automated prognostics systems is increasing as well. This work intends to contribute to the development of predictive systems that can help controllers on their daily tasks by directing their attention to problematic manufacturing areas.

The paper is organized as follows. Section 2 briefly reviews related work. Section 3 discusses the proposed modeling approach. Section 4 discusses the methodology. Section 5 presents experiment results and demonstrates the proposed approach on real data. Section 6 discusses and concludes the paper. Future work is also discussed.

2. RELATED WORK

Data driven applications were first applied to areas such as health, earth science, business management and social theory. These areas, frequently defined by the lack of solid deductive theory, were must benefited by the knowledge extracted from data that enabled sound theory building [6] [7] [8] [9]. On industrial level, the use of these techniques, deriving from data science, became practical since the 90's when technology matured [10] [11], and have since then been established as a potential solution to help researchers make adequate

decisions.

Aerospace industry has also embraced the advances offered by data-driven predictive techniques, from structural prognostics [12] [13], to damage assessment [14], engine health monitoring [13] [15], seeded-fault predictions [15] [16], avionics [17] and even combination architectures of data-driven and model-based techniques for a prognosis fault-detection reasoner for the hole aircraft [18].

At manufacturing level, there are also numerous reports on quality prognostics for already highly automated assembling systems [19] [20] [21] [22] [23]. By predicting processes' outcome it is possible to anticipate and adapt to product variations, rendering the manufacturing system more flexible and efficient . To achieve these goals there are two possible approaches: one more remote, "off-line", consisting in static, and more often, long-term adaptations; and "on-line" dynamic adaptations consisting on more immediate changes to the process, autonomously determined by the system, often referred to as machine learning [20]. Figure 1 represents this general idea.

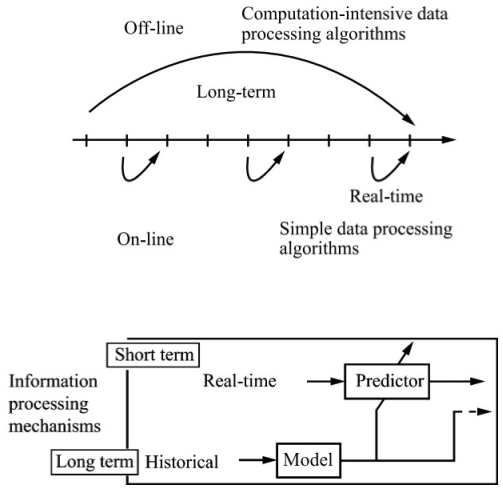


Figure 1: Off-line and on-line prognostics. General Framework. Adapted from [24]

The work carried out is focused on supervised leaning, for quality prognostics in the aerospace manufacturing industry. More specifically it is applied to the riveting process of aircraft skin panels.

Supervised learning techniques, in a broad sense, can be embedded under the same general data-driven machine learning framework. In fact, there were already developed some automated machine learning platforms [25]. The outlines from this implementation were adapted for building this prognostics proposal, given its successful applications throughout various popular machine learning competition platforms such as Kaggle, Codelab and DrivenData [2].

The success of data-driven techniques, even with all the resources of a great machine learning expert, is often more tied up to pipeline tweaking and feature engineering, than with great machine learning algorithms. Domain knowledge, and an engineering approach can in fact turn the most simple algorithms into efficient applications.

Therefore, the followed approach consists of the following steps:

1. design and implement evaluation metrics
2. make sure the framework (pipeline) is solid end to end
3. begin from a reasonable objective (baseline solution)
4. combine data-driven techniques with domain knowledge to attain an informative feature space
5. train and optimize different machine learning models
6. make sure the pipeline stays solid
7. evaluate model performances
8. select and test best performing models

The principle behind this "simple" generalization is that *"adding complexity slows future releases"* [26]. If a solid solution is found it can, then, be further optimized. Indeed, these "rules" unfold into several different particularizations, according to Martin Zinkevich [26] and Pedro Domingos [27] [28]. These were the guidelines for all the development.

3. BACKGROUND

The process of assembling aircraft skin panels starts of with a preliminary set up where all the panels are assorted, splice stringers are placed on the panel borders and pinned together manually with provisional rivets referred to as tacks. In most cases these are removed once the final rivets are installed, being its main purpose to hold the parts in the desired relative location when moving the pre-assembled structure from the fixtures to the automated riveter. After the tackled assembly has been transported to the drill-rivet machine, and positioned on the platform fixtures, the machine starts running a routine for each bay [3].

Simultaneously, a set of sensors retrieve information on the performed fastener head movements; another set of sensors signal any non-identified body entering the work space and stops the machine; counters provide data on how many fasteners/collars there are still on the cassettes and so on. All this information is collected and registered on the machine logs.

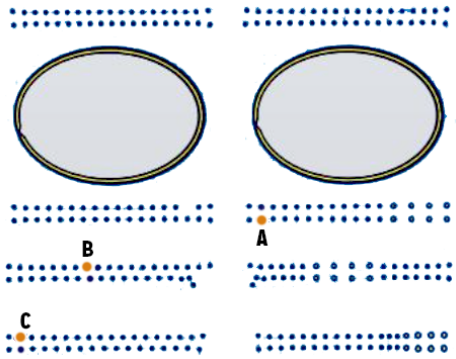
The quality control team is responsible for visually inspecting and measuring diameters/positioning of every rivet, looking for faulty installations. Drawing on experience of dealing with this procedure in this specific environment, one concludes that essentially five **types of non-conformities** can be identified:

I Holes not drilled
II Holes drilled with a deviation from the nominal location
III Holes drilled with a deviation from the nominal diameter
IV Fastener not installed (and consequently collar not installed)
V Collar not installed

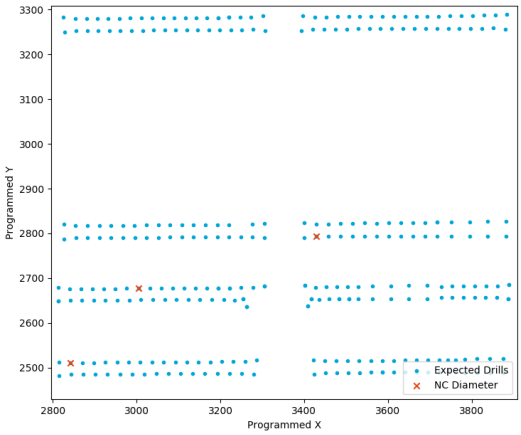
So as the process is finished, besides the assembled skin part result two data sources: the machine data logs and the quality control reports. To take advantage of the information contained in the data sources it is essential to consolidate these

[2]http://www.kaggle.com
http://www.codalab.com
http://www.drivendata.com

[3]The term bay is used to describe each section/block of rivets



(a) Adapted non-conformity report



(b) Induced location of anomalies on programmed coordinates plot (from non-conformity report to machine data log)

Figure 2: Anomaly mapping (type III: holes drilled with with a deviation from the nominal diameter) location. Values on the plot axis were removed due to confidentiality issues.

two representations of the industrial process (the machine logs and the non-conformity reports). The data on quality reports was leveraged as a binary classifier for the entries of the machine log, thus creating the label `"Anomaly"`, that was set to `1` if a non quality was found in a sample's intrinsic location, and `0` otherwise.

For some real world applications it is of key importance that the obtained models are particularly accurate at some sub-range of the domain of the target variable. Frequently, these specific sub-ranges of the target variable are poorly represented on the available training sample. In this case, those instances are the anomalies, with a representation of 0.16% relative to the whole rivets' instances.

The problem faced in these cases is commonly known as imbalanced data distributions, or imbalanced data sets. In other words, in these domains the cases that are more important for the user are rare and few exist on the available training set. The combination of the specific preferences of the user with the poor representation of these situations creates problems to modeling approaches at several levels. The main problem of imbalanced data sets lies on the fact that they are often associated with an user preference bias towards the performance on cases that are poorly represented in the available data sample.

Naturally experimenting with classification methods was the first resolution, until it started to become obvious that trying to predict such an imbalanced class, originate biased models with low accuracies. The results raised an unacceptable amount of mis-classifications, with low recall and even lower precision, despite all the efforts on reducing the effects of the imbalance (stratified split, under-sample of the majority class, over-sample of the minority class [31]).

A survey on the content of the different variables of the process was conducted alongside the manufacturer. With the knowledge gathered there were already some lines one could draw about what were important attributes, and what could be at least a partial relation between variables that would indicate the event of anomaly. The preliminary approach was constructed with the support of this technical knowledge and some elementary programming aid, without even going through the pre-processing techniques, resulting on what was designated as **baseline solution**. The objective here was essentially to get a visualization of the contrast between these pre-conceived notions on the target's behavior and the reality of which instances were in fact targeted as an anomaly by the quality control team. From the definition, technical baselines are references from which to measure progress of the system development. The reason most processes establish a baseline is so that progress can be monitored with respect to a common reference. In this case, the baseline solution is also a snap shot realization of what could be expected as a final result. Figure 3 is the graphical representation of the solution obtained.

In order to obtain this result some definitions had to be outlined, as to what constitutes each class:

- Expected Drills [●]: the locations of the expected drills are essentially the coordinates where the machine was programmed to execute drilling, accessed directly from the attributes.
- Type I (NC) [×] and Type III (NC) [×] : the locations of anomalies were mapped from the non-conformity reports.
- Type I (⊢) [●] and Type III (⊢) [●] : the deduction of anomalies' locations was made, from the information within the samples' attributes through logical inference, for each anomaly type at a time.

With a quick look at Figure 3 it can be seen that there are only two types of anomalies comprised in the specific part from the case study, Type I and Type III. In each case, an elementary deduction processes was carried out, in order to infer the locations.

Taking into account the results obtained with the baseline solution, it was deduced that it would be more effective to come up with predictors for each anomaly type individually. Considering this, and returning to the reasoning behind the creation of a new feature for label representation: the focus was primarily turned to the identification of Type III anomalies since, from the knowledge of the problem's domain, there is an irrevocable relation between its occurrence and diameter related features. Thus, creating a new predictor y_h for Type III anomalies would allow to reformulate the problem from a binary classification into regression. This kind of approach has been proven successful in other real world applications of machine learning techniques [33].

So now the question that remains is what is the essence of this new target y_h. Manually creating features requires a lot of time spent with the sample data, while evaluating the possibilities on how to expose them to the learning algorithms. This is the part of feature engineering that is often compared to an

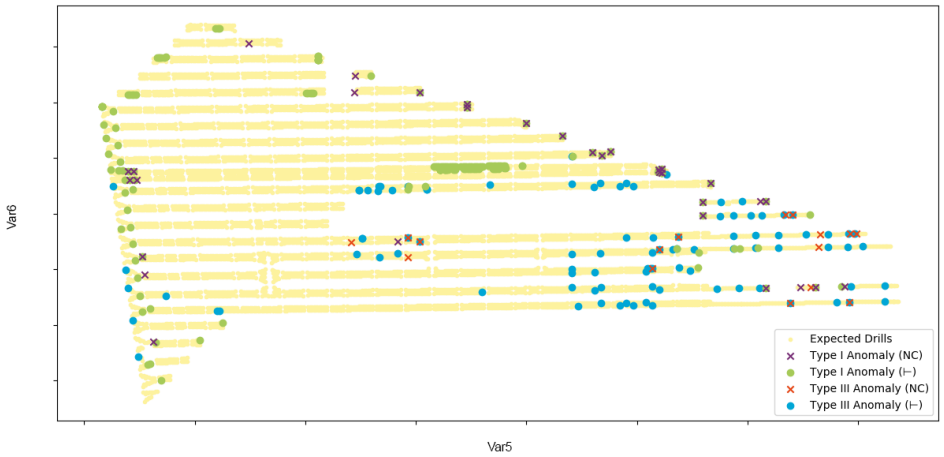


Figure 3: Baseline Solution. Plot of the expected drills' location superposed with Type I and Type III anomaly' locations deducted (⊢), and reported by quality control (NC). Values on the plot axis were removed due to confidentiality issues.

art form, and can really set the solution apart on competitive machine learning.

"At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used. (...) This is typically where most of the effort in a machine learning project goes. It is often also one of the most interesting parts, where intuition, creativity and "black art" are as important as the technical stuff.-- Pedro Domingos, *A few useful things to know about machine learning* [28].

Figure 4 depicts two views of a rivet installation. a longitudinal view, on the left, and a section view, on the right. During the drilling process, with the sample rate of approximately one in every ten drills, the resulting hole is probed in two different, depth relative (longitudinally), positions. In each one of these positions the diameter is measured in different directions, relative to the circumference.

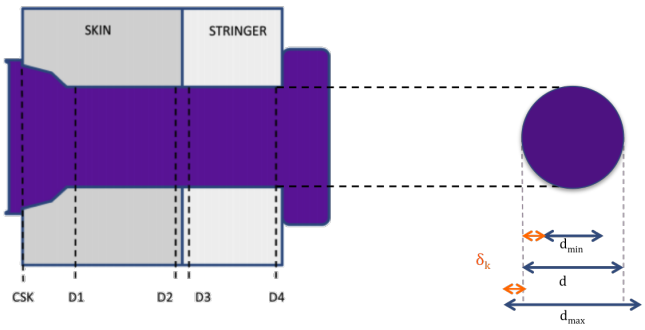


Figure 4: Schematics of a rivet hole. Representation of the variables used for label creation. Adapted from Hayes, C. [34]

The constructed feature, "Dev_max" $= \Delta\text{max}_\text{j}$, is a measure of the maximum deviation from the diameter tolerance interval found along the depth of the rivet hole from the sample X_j.

First it is performed an intermediate calculation of the k-th measurement maximum deviation from required tolerance, δ_k. Let k be the total number of measurements taken, and $t = [d_{\text{min}}, d_{\text{max}}]$ the tolerance interval. If d is the measured diameter at the instance k, then:

$$\delta_k = \text{max}_\text{r}(\text{r}_{\text{inf}}, \text{r}_{\text{sup}}) = \text{max}((\text{d}_{\text{min}} - \text{d}); (\text{d} - \text{d}_{\text{max}})), \quad (1)$$

where $r = \{r_{\text{inf}}, r_{\text{sup}}\}$ are the deviations from the inferior and superior tolerance limits, respectively. The result is $\Delta_j = \{\delta_1, \delta_2, ...\delta_k\}$, for each sample X_j, from which the maximum will once again be selected:

$$\Delta\text{max}_\text{j} = \text{max}_{\delta_\text{k}}(\Delta_\text{j}) \quad (2)$$

Therefore the new label is:

$$y_h = \Delta\text{max}_\text{j} \quad (3)$$

It is now possible to resort to regression techniques and aim at building a structure and distribution-independent model.

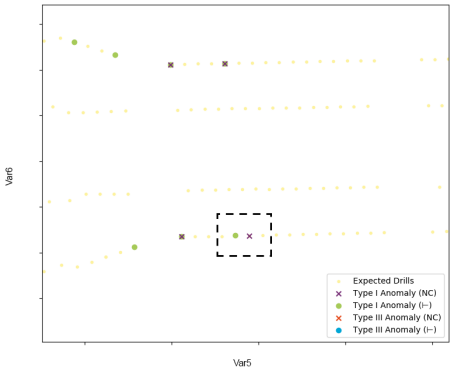
The logic reasoning over which y_h was constructed is:

a) if the value of the measured diameter, d, falls under or above the restrictions for the minimum and maximum diameter, respectively, it will indicate an anomaly.
b) it suffices that one of these situations occurs, for the hole X_j to be considered anomalous, thus δ_k is the maximum r obtained in the measurement k.
c) when $\{r_{\text{inf}}, r_{\text{sup}}\}$ are calculated it is insured that if the diameter complies the requirement, $r \leqslant 0$.
d) for each sample X_j there are k diameter measures taken therefore, since the same principle from item 3 applies, $\Delta\text{max}_\text{j}$ is the maximum deviation measured.
$\therefore$ if every measurement satisfies the tolerance requirements $\Delta\text{max}_\text{j} \leqslant 0$.

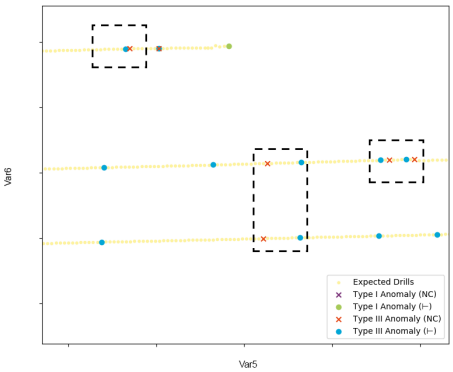
Accordingly,

$$\hat{y} = \begin{cases} 1, if \hat{y}_h = \hat{\Delta\text{max}}_j > 0 \\ 0, if \hat{y}_h = \hat{\Delta\text{max}}_j \leqslant 0 \end{cases} \quad (4)$$

So now that the problem is defined as a regression, and the focus is on type III anomalies only, some adjustments need to be done, especially with regards to the pre-existing anomaly



(a)



(b)

Figure 5: Zoomed in sections from baseline solution. Areas that derived Conjecture X are emphasized with a surrounding rectangle. Values on the plot axis were removed due to confidentiality issues.

labels, i.e., the samples with respect to type I anomalies that were previously classified as positive, are now set to zero. In fact, this conversion mirrors the behavior of a logistic classifier, by defining the threshold that dictates what is classified an anomalous sample.

Figure 3 displays an overall perspective on the placement of anomalies, a more detailed view - Figure 5 - helps to introduce the next argument. " *What is the level of trustworthiness of the data collected on-site?*". It is crucial to remind that all this work relies on human capabilities, thus being susceptible to human error. In this case, what is meant, and becomes quite noticeable when analyzing the results, is that probably a few mistakes regarding the location of the perceived anomalies were committed when doing the mapping to the technical drawing that is the essence of the quality control report.

Considering the proximity of locations (e.g. Figures 5a), or the similarity of the patterns (e.g. Figures 5b) this could possibly go unnoticed even by the people responsible for repairing, regarding the fact that once facing the approximate location they are able to evaluate and fix what indeed needs to be fixed, without reporting back on these peculiar details.

The hypothesis of human error being the reason for these coincidences is a hefty assumption, and wasn't confirmed by any means, nonetheless from now on it is gonna be referred to as: **Conjecture X**.

In order to further test the conjecture with more sophisticated algorithms, a new binary target variable (label) is created: y_{CONJX} = "CONJX". *Testing this assumption also becomes a goal of this research.* It is important to clarify that the number of anomalies existing in the data-set is different for each label. This is attributed to the fact that the data-set has to be that consisting on the samples with probing measurements, so that the feature $= \Delta\max_j$ could be generated. Therefore, while in the case of label y_{orig} there are eight anomaly instances falling into the data-set, in the case of label y_{CONJX} there are 15.

4. FRAMEWORK

The very first step is to consolidate and clean the data-set to get the data in tabular form, suited for data driven techniques. Established this, the following procedure is to split the dataframe into training and validation sets: $\{\mathbf{X}_{\text{train}}, \mathbf{X}_{\text{CV}}\} \longrightarrow \{\mathbf{y}_{\text{train}}, \mathbf{y}_{\text{CV}}\}$. In order to optimize the validation process and performance evaluation, it is employed the K-folds Cross Validation split. K-folds is a technique that deals with the variance problem: evaluation scores vary when testing over different validation sets, so if the data-set is only split once, the evaluation metrics obtained are not rigorous. When the data-set is split into k folds, the algorithm runs on iterative mode along k instances, where the model is trained on nine-folds and then tested on the remaining fold, and since there are k folds there are k possible combinations of $k-1$ training folds plus one validation fold. K was set equal to ten, meaning the model is trained and tested in ten different combinations of sets, and in the end it can be evaluated by averaging the performances obtained [35][36]. The cross validation set is kept separate from the training set all along the process, and is by no means accounted for when fitting objects. Instead, the parameters fitted to the training set, resulting from every operation applied, are flushed aside so they can be applied in the end to $\mathbf{X}_{\text{CV}}$ for evaluation purposes. This separation is crucial to prevent the system from overfitting, and to guarantee that the model created generalizes well to unseen data.

After the splitting is done, the training set is sent to different pipelines for standardization and feature engineering steps.

Data Cleaning

Thoroughly, each variable was examined, and some were immediately discarded for adding absolutely no value, based on being duplicate information, referring to some other process, presenting no variation, and some other were just missing data. Variables regarding identification details, were also discarded once were not considered to be useful to diagnose anomalies.

There is another set of variables, critical to be discarded, which is that of attributes that depended on the rivet hole probing measurements. On one hand, if those attributes were included, it would result in data leakage at feature level inducing biased models, on the other, it would be impossible to extend the results to unseen data if it contained samples where the measures had not been taken.

Besides the obvious ones to be removed, some assumptions were made about what data should not be needed, being all of it recorded to test later, if necessary.

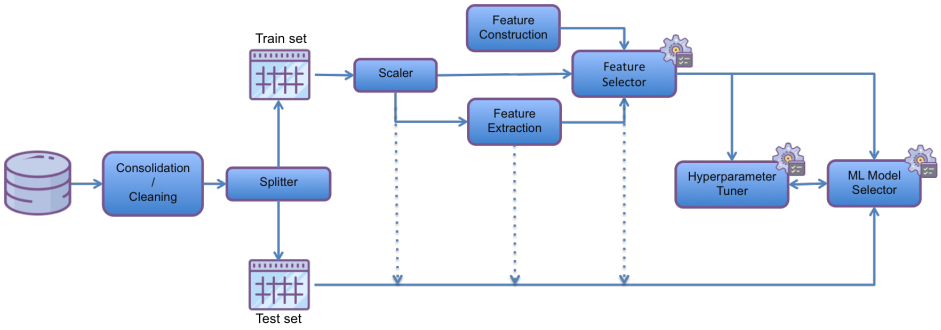


Figure 6: Adopted Framework. Each block stands for a step in the iteration process. Gear symbols evidence the steps of greater focus.

Data Scaling

Features with very different scales and with large variances usually lead to difficulties visualizing data, but more importantly, they may have a very negative impact on the performance of many machine learning algorithms. Unscaled data can also have effects on the convergence speed rate, or even prevent it from happening, with many gradient-based estimators. To avoid these occurrences, the variables were standardized.

Extended Statistic Analysis

If one pictures a scenario where there are a whole lot of independent features where each correlates well with the target, one can imagine how easy learning from these features would be. It is indeed important to visualize the variables and how they correlate with each other, and particularly with the target. When two features are highly correlated between themselves, it means they explain the same variance, thus being redundant.

However, if the target is given by a very complex function of features it may present itself very hard to learn. This is referred to in most machine learning literature as the *curse of dimensionality* [37]. In real world case studies, raw datasets are hardly ever in a ”friendly”learning form, but one can construct features from it that are, and remove those ones with insignificant relevance. That’s what is gonna be explored next.

In order to reduce feature space while maintaining the most possible information, one can extract the feature space principal components and see how much information is retained if the feature space is reduced to a certain number of principal components. This technique is called principal component analysis (PCA) The first five principal components are then extracted and are subsequently added as input variables for feature selection methods.

Feature Selection

Some complex predictive modeling algorithms that perform feature importance evaluation and selection internally while elaborating the model where compared. The algorithms chosen were the ones that could compute the mentioned feature importance during the model preparation process, without actually performing feature selection. The scores obtained are displayed on figure 7. The method for ordering features based on feature importance given by Mean Decrease Impurity is iterative, through different random states of the random forest.

The selected features are sent over to the model selector and to the hyper-parameter tuner. Every routine can immediately be tested by the evaluator module, since it is a flexible framework. For that purpose all the transformations applied to the training set until the moment of evaluation are also applied to the validation set. When another iteration starts, all the modifications to the data-set are reset.

Model Selector & Hyperparameter Tuner

In the model selector module the proposed machine learning algorithms are fitted to the training set, taking into account the features chosen. The hyper-parameter tuner selects the optimal hyper-parameters for the fitted ML algorithm. This steps are often executed in parallel, and always taking advantage of the tenfold cross validation method.

The models created are then validated on the cross validation set, by applying the same transformations, previously modeled to the training set. The evaluator module generates evaluation metrics for the results obtained.

Extended Testing

Given the case where good performances are obtained in the validation process, it is desired to extend the predictions to unseen data, as previously mentioned. This is the final goal of all the techniques and models developed.

For this goal to be met, there are essentially two key factors involved. The first is regarding the features considered for this project’s approach, more specifically why all of those attributes that depended on the rivet hole probing measurements were discarded. On one hand, if those attributes were included, it would result in data leakage at feature level inducing biased models. On the other one, it would be impossible to extend the results to unseen data, if that data was referent to samples where the measures had not been taken.

Which is intrinsically connected to the second key factor: keeping only the samples where data relative to the that same variables dependent of the probing procedure was not missing, allowed the construction of the label $\hat{y}_h$ to validate the models in terms of regression. It is important to emphasize the fact that the features were only used for label creation, and were not introduced in the feature space, as mentioned on the first key factor.

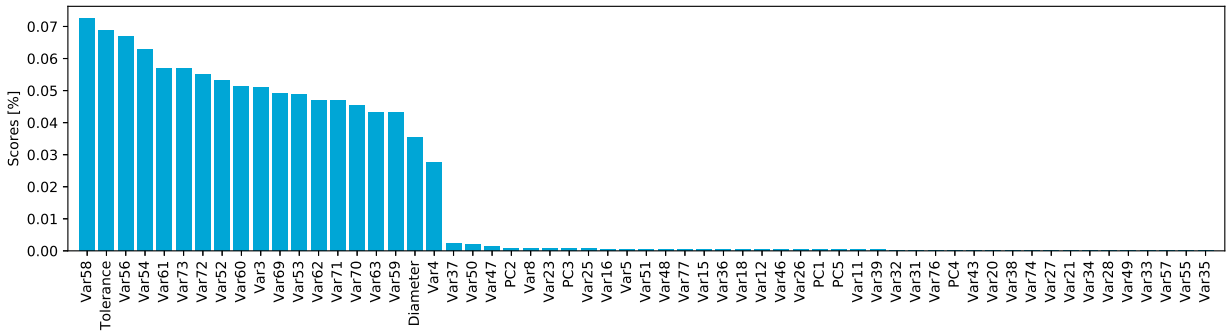


Figure 7: Feature importance computed resorting to Random Forest scores. Features names were omitted due to confidentiality issues.

So, why are these considered key factors? Well, discarding all the samples with missing data regarding the probing procedure left aside a large number of observations that were unseen through the hole cross-validation process. In fact they could not even be used for that purpose for lacking the information to construct the regression label.

Furthermore, making the overall process independent of the probing variables, also means that now the validated models can be applied to those samples that were discarded on a first instance. Although there are no $y_n =$"Dev_Max" labels regarding these samples, there is still the classification information from the quality control reports.

Thus, the fundamentals behind the idea of extending the results is that, if the models achieved good performances on the validation set, it should be possible to apply those models to unseen data. Where unseen data is now those samples previously discarded. Then, making use of the principle established for conversion of the regression results into classification labels, the results obtained by applying the regressor models can be converted and compared to the labels from quality control, y_{orig}, and from conjecture X, y_{CONJX}.

As for the nature of this process of extending the results, it falls somewhere in-between a semi-supervised approach, for the lack of labels for the regressor's outcome, and a logistic regression, for the conversion of the regression results into classification labels. In fact, this conversion itself mirrors the behavior of a logistic classifier, by defining the threshold that dictates what is classified an anomalous sample.

Hopefully, the models that achieve best performances in the validation set (the data-set used until now for validating and tuning the models), will now be able to generalize well to the test set, $\mathbf{X}_{\text{TEST}}$. The new testing set has now approximately 9000 observations, where according to the label y_{orig} there are eight anomaly instances, and according to y_{CONJX} there are only two.

5. Results

Diagnostics Results

The results regarding the quality diagnostics perspective were already delivered in table 1. This framework already grants an optimized method for quality assessment, either from the perspective of the original label y_{orig} as from the y_{CONJX}. Undoubtedly the evaluation scores improve under the umbrella of Conjecture X, after all the conjecture itself was inferred from the results. In the case of type I anomaly recall raises from 75% to 85% while in type III anomalies this raise is even more drastic, being almost a difference of 40%. Recall is a critical evaluation metric for the case study since false negatives need to be penalized the most.

Besides being the symptomatic results of exploratory data analysis, these were also the results that established the baseline for expanding the research. Analyzing the spacial distribution of diagnosed anomalies vs. the anomalies reported by the quality control team, lead to the suspicion of miscataloged anomaly locations (figure 5). Based on this conjecture, a new binary label was created, shifting the positive classification to the samples judged to be the correct. This label was from that moment referred to as $y_{\text{CONJX=}}$ "Conjecture X".

Anomaly/ Metrics	Type III		Type I	
	Original	Conj. X	Original	Conj. X
True Positives	8	14	24	29
True Negatives	$\simeq$ 10 000	$\simeq$ 10 000	$\simeq$ 10 000	$\simeq$ 10 000
False Positives	88	82	146	143
False Negatives	8	2	8	5
Precision	$\simeq$ 8%	$\simeq$ 15%	$\simeq$ 14%	$\simeq$ 17%
Recall	$\simeq$ 50%	$\simeq$ 88%	$\simeq$ 75%	$\simeq$ 85%

Table 1: Evaluation Metrics for classification executed with data analytics

Prognostics cross-validation results

Respecting the prognostics approach the validation results were obtained applying a tenfold cross validation technique, meaning the evaluator module considered the accuracy estimates obtained through ten iterations. On each iteration, the data-set was split differently into training and testing set, where the test set was hold out until the actual evaluation process. A ten-fold cross validation confers a good trade-off between having only one validation set (which would minimize the variance on the testing set, but induce high bias on the model), and the leave-one-out approach (that make bias negligible, however, since there is only one sample on the testing set, the variance in the estimates of the model's error would be very high). Besides ensuring a good bias-variance trade-off in the testing process and on the model itself, unfolding the samples also creates a greater room for improvement, since more experiences can be conducted.

The models' performance was evaluated in regards to the mean absolute error (MAE) and the root mean squared error (RMSE) for the regression models' predictions on the testing set. Root mean squared error will always be larger or equal to mean absolute error. Together this evaluation metrics can

be used to diagnose the variation in the prediction errors: the greater the difference between then, the greater the variance in the individual errors in the test set.

Furthermore, recalling that the regression approach was generalized from the classification problem, the results were filtered resorting to the binary classification labels y_{orig} and y_{CONJX}. MAE and RMSE were calculated for each corresponding set of y_h. This gives a better understanding of how the models perform when predicting the maximum deviation from the diameter tolerance interval found along the depth of the rivet hole ($y_h = \Delta\text{max}_\text{j}$), in the set of cases where it was considered anomalous, and the set where it wasn't. This evaluation was performed regarding the original labels, y_{orig}, from the quality control reports, and the labels from "1", y_{CONJX}, formulated accordingly to the results obtained on the baseline solution.

The six columns to the right on table 2, are in fact evaluation metrics for the results of this conversion of the predicted maximum deviation from the diameter tolerance interval found along the depth of the rivet hole (equation 4).

It is important to recall that when the label differentiation is created, the number of anomalies existing in the data-set $\mathbf{X}$ (cross validation data-set) is different for each label. This is attributed to the fact that the cross validation data-set has to be that consisting on the samples with probing measurements, so that the regression label y_n can be generated. Therefore, while in the case of label y_{orig} there eight anomaly instances falling into the data-set ($\mathbf{X}$) and 8 on the test set ($\mathbf{X}_{\text{TEST}}$), in the case of label y_{CONJX} there are 15 in the cross validation data-set and only two in the test set.

The unbalanced distribution of anomalies across the cross validation and testing set, makes it so that the number of positives and negatives is different when considering each label.

The metrics adopted were precision, recall and F2-Score. These rates are calculated on top of a confusion matrix. When evaluating supervised classification problems it is common to refer to it and, for a binary class, it consists of four measures: True Positives, True Negatives, False Positives and False Negatives.

Precision can then be defined as a measure of how often the positive class was predicted and it corresponded in fact to a positive sample. Recall represents the rate of predicted positives over the actual positives, it is the conjugate of the false negative rate(FNR), or miss rate.

From the manufacturer's point of view, and for that matter, for the quality control procedure, it is critical to reduce the number of false negatives to zero. A false negative would mean that the prognostic indicates that there is no anomaly instance at a determined location when in fact there is. Therefore, the intent is to keep the recall rate the highest possible, since for the machine learning model to be adequate for the automation of the quality control process, none anomaly can go unnoticed. On the other hand, it is desirable to have a good precision. Precision is the measure of improvement introduced by adopting the model developed: the greater the precision, the smallest the area flagged as critical and in need of assessment by the quality control team.

Considering this, another rate is taken into account for the purpose of model evaluation, the $\mathscr{F}_\beta$ Score. This measure was derived so that *"it measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision"* [38]. In this situation, given the preponderant importance of attaining good recall rates over precision, $\beta = 2$, placing more emphasis on false negatives.

The lower accuracy delivered by the linear models can be plausibly sustained by the fact that the distribution of observations is in fact non-linear. Having said that, and regarding the great performances obtained in the cross validation of the remaining models, this are considered not adequate and will not be further considered.

The best performance overall is that of the random forest, with the smallest regression errors for every case, as well as one of the best classification scores achieved pos-conversion (equation 4). Follows the decision tree performance regarding MAE and RMSE, however when the conversion to anomaly labels is performed it fails to classify $1/8$ anomalies in the y_{orig} cross validation set, and $2/15$ in the y_{CONJX} set.

In terms of classification scores, k-nearest neighbors model comes in first place, identifying the total number or anomalies in both sets, same as the random forest, but with a greater precision, although the regression errors are somewhat bigger than the top mentioned models.

Regarding the conjecture constructed in the context of the baseline solution ("Conjecture X"), one can say that this results are supportive of it. For the most algorithms implemented, the results obtained (regardless of classification labels) through regression, consistently support the truth values assigned in y_{CONJX}. This can be observed by performing the conversion of $\hat{y}_h$ regression predictions, into $\hat{y}$ classification predictions through equation 4. To truly confirm the conjecture created would require more information on the overhaul process. However, regardless of that fact, one can conclude that this results support it.

Another interesting fact about this predictions is that they outperform the baseline solution in terms of recall, even though they are otherwise built uninformed of the quality-measurements *in loco* regarding the rivet-hole diameters. This is indeed the actual definition of a predictive model, otherwise it would be considered a simple diagnosis.

Prognostics extended results

Since there are no y_h labels for this test set, the evaluation metrics can only focus on the predictions obtained by converting the $\hat{y}_h$, accordingly to equation 4 into classification predictions. This process was designated as extended testing and it's main purpose is to evaluate how well the models generalize. The results achieved through this generalization are summarized in table 3.

Support vector regression, that previously on the cross validation set showed signs of instability, achieves the worst performance from the five models, confirming the suspicion.

The nearest neighbors model doesn't generalize well either, under none of the classification labels.

A sense of duality is installed when evaluating the efficiency of the generalization of the remaining three tree based algorithms: it is divided by the constraint of under which label they are evaluated. It is interesting to observe that for the original anomaly label all three performances are mediocre; while for the conjectured label all are perfect recall-wise.

Model	MAE (y_h) [mm]				RMSE (y_h) [mm]				Precision [%]		Recall [%]		F2Score [%]	
	$[y_{orig}=1]$	$[y_{CONJX}=1]$	$[y_{orig}=0]$	$[y_{CONJX}=0]$	$[y_{orig}=1]$	$[y_{CONJX}=1]$	$[y_{orig}=0]$	$[y_{CONJX}=0]$	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}
Suport Vector Machine	0.009114	0.008203	0.066245	0.066553	0.012796	0.010533	0.606014	0.607670	6.4	10.5	88.9	93.3	24.8	35.3
Decision Tree	0.008099	0.0075933	0.009577	0.0084829	0.0106916	0.009577	0.077216	0.077410	8.3	13.5	88.9	86.7	30.3	41.7
Random Forest	0.008096	0.006805	0.008181	0.008179	0.010964	0.009126	0.063272	0.063438	8.5	14.3	100.0	100.0	31.7	45.5
Extreme Gradient Boost	0.012660	0.009542	0.008841	0.008805	0.015340	0.012432	0.080614	0.080799	7.8	12.7	88.9	93.3	29.0	40.1
K-Nearest Neighbor	0.009554	0.008113	0.010577	0.010592	0.012915	0.011094	0.085517	0.085747	9.1	15.2	100.0	100.0	33.3	47.2
Bayesian Ridge	0.015730	0.019747	0.024905	0.024892	0.021719	0.025095	0.098812	0.099054	1.5	1.0	55.6	26.7	6.6	4.3
Lasso	0.028269	0.027499	0.012786	0.012739	0.030885	0.094793	0.029343	0.095036	0.3	0.5	11.1	13.3	1.1	2.2
Linear Regression	0.044734	0.039188	0.029415	0.029376	0.045316	0.043226	0.097615	0.097819	0.5	1.0	22.2	26.7	2.3	4.3

Table 2: Comparison between algorithms performances. Regression errors are filtered according to the binary classification labels. Classification metrics are obtained with the conversion of regression predictions (equation 4)

Model	True Positives		True Negatives		False Positives		False negatives		Precision [%]		Recall [%]		F2 Score [%]	
	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}	y_{orig}	y_{CONJX}
Support Vector Machine	0	0	8913	8918	1	1	8	2	0	0	0	0	NaN	NaN
Random Forest	3	2	7782	7787	1131	1132	8	0	0.3	0.2	27.3	100.0	1.3	0.9
K-Nearest Neighbors	1	1	7843	7849	1070	1070	7	1	0.1	0.1	12.5	50.0	0.5	0.5
Decision Tree	2	2	8006	8012	907	907	6	0	0.2	0.2	25.0	100.0	1.1	1.1
Extreme Gradient Boost	3	2	7825	7830	1088	1089	5	0	0.3	0.2	37.5	100.0	1.3	1.0

Table 3: Extended Results. Acknowledging the fact there are no labels for the results of regression, the metrics presented are only concerning classification.

6. CONCLUSIONS

With the proposed approach, data-driven quality diagnostics and prognostics were developed to tackle a challenge presented by an aircraft manufacturer. Here, regressive supervised machine learning methods were used to deduce quality classification predictions with the aim of performing anomaly detection. The models resulting from supervised learning were also applied to data where there were no regression labels.

This solution also draws knowledge from expert information about the domain, following the idea of "machine learning with world knowledge". This insight was particularly important in the development of the diagnostics models, to help create new features (feature engineering), and to detect incorrect labels. The results demonstrated the effectiveness of the data-driven approaches in the process industry domain; they revealed the non-linearity profile of the automated riveting process; and provided insight on how to combine information from different sources to enhance quality prognostics.

The research goals were two-fold:

I to optimize the diagnostics on quality control of the automated riveting process through data analytics;
II to create a predictive prognostics framework on quality control of the automated riveting process through data-driven machine learning techniques.

The first stated goal was attained using exploratory data analysis and simple statistic classification tools. It was possible to attain considerably high recall in the identification of the anomalies, previously labeled by the quality control team. Recall here can be considered the most important evaluation metric for the classification results, considering the industry goal. While the metric of precision measures the degree of optimization relative to the current industrial quality control protocol; recall assures the same standards are met.

Furthermore, a tool for visualization of the results, exposing the diagnosed critical areas was developed. This tool can be adopted at industry level to direct the technicians on the overhaul procedure, and was considered advantageous by the manufacturer. This was the first approach to optimize the quality diagnostics.

The visualization tool, when applied to the specific data of this case study, raised questions regarding the precision of the anomaly labeling by the quality control team. Several anomalies appeared to be incorrectly labeled by the expert team. A conjecture based on the problematic instances was elaborated: *the goal of analyzing if this conjecture is sustained by further results was declared.* To do so, a new binary label was created, based on the conjecture. The diagnostics optimization result was established as the baseline solution for the application of data driven techniques for quality prognostics.

Next, the second (more ambitious) goal was pursued: the development of a learning agent for quality control prognostics.

During the development of the framework, it was possible to conclude the importance of feature engineering and data pre-processing techniques as key factors to obtain consistent

results. Once again, domain knowledge revealed itself an essential piece on the construction of the learning data-set.

The first binary classification models performed poorly. It was concluded that the imbalance nature of the data was preventing the models to achieve good results. To circumvent the situation, the research goal was particularized to a specific type of anomalies: anomalies ov type III (rivet hole with diameter outside the nominal diameter bounds).

The particularization allowed for a new approach. A new label is built based on the deduction of how probing measurements would imply an anomaly instance. This new label is a continuous variable, allowing for experimentation with regression models, therefore, and more importantly, eliminating the imbalance issue.

However, since the label creation is dependent on measurements that are only performed so often, approximately 90% of the data-set was discarded at this stage, for lacking the data for label construction. The framework validation process was delivered on only 10% of all the available data: the cross validation data-set. This also meant the distribution of anomaly instances according to the original classification label, in the cross validation data-set, become different than the distribution of the conjectured label. The idea that needs to be taken from here is that when evaluating the models performance this effect needs to be accounted for.

Furthermore, it was acquired new domain knowledge regarding domain variables importance and interference with this anomaly type manifestation. For making this possible it was critical to maintain the understandability of the variables in the feature engineering process. To perceive this relations was an underlying objective inside the research goal, and an industrial goal as well.

Regarding the comparison of the regressive machine learning models in terms of MAE and RMSE, the validation performed in this framework lead to the conclusion that regressive support vector machines, decision trees, random forests, extreme gradient boost, and nearest neighbors achieve the lowest prediction errors regression-wise.

All the linear regression models, ordinal least squares, lasso and bayesian ridge, performed poorly under every circumstance considered. From here, the conclusion that this is a non-linear problem can be taken.

Also, from the good performing models, it is perceived that the problem of imbalanced data was in fact removed by the regression approach.

Regarding the extended results, the hunch on the support vector machine model being unstable is immediately confirmed by its poor performance in this test set under both label conditions, followed by the nearest neighbor regressor. The efficiency of the generalization of the remaining three tree based algorithms is divided by the constraint of under which label they are evaluated. It is interesting to observe that for the original anomaly label all three performances are mediocre; while for the conjectured label all are perfect recall-wise.

So, essentially, there are two plausible conclusions:

- "Conjecture X"is confirmed meaning the tree models generalize very well with 100% recall on unseen data, approximately the same precision of 0.2%
- "Conjecture X"is wrong and the the models don't generalize so well, with the more efficient model being extreme gradient boost with 37.5% and 0.3% precision.

The prospect of generalization is in fact of surmountable importance when considering applying the models developed to the real life automated riveting process. In fact, by removing the probing dependence, the model became supportive of performing under different circumstances, e.g. different skin assemblies with different diameter specifications. Furthermore, the implementation cost it may take, is strongly encouraged given the potential decrease on elapsed time in quality control checks. From an optimistic point of view, where the technique developed would be implemented *online*, there is even potential for total elimination of the need for human-based quality control. For the manufacturer, besides reducing the producing time of every skin panel, this solution also encourages a better allocation of human resources.

7. Future Work

Each one of the frameworks developed still has room for improvement, and in the prognostics case, for further research.

Beginning with the diagnostics solution, the improvements are mostly connected to the way data is retrieved. This follows from the fact that the framework mostly depends on probing measurements, that are only taken in 10

The prognostics solution has the clear necessity to validate the conjecture that grants it generalization capabilities. The process of validation of the conjecture is intended as further research for the author. It will mainly consist in applying the model to the machine log data of an active skin riveting process correspondent to the same assembly part of the case study. Otherwise, if it is a different part, the framework can still be applied but will have to wait on the quality assessment in order to have labels to train.

If "Conjecture X" was already proven, there is room to conduct the experiment of, departing only from machine data logs, create the regression label. Next, apply the prognostics framework on the label data-set and test its predictions. This would be a testing phase where the learning agent would operate remotely. Given the strict requirements in aerospace industry, to effectively adopt this technique is process would require extensive testing.

On the other hand, for data-driven techniques to be in fact implemented on line, it would be interesting to do further research on this topic regarding the remaining anomaly types for which a prognostics solution was still not developed.

An even more ambitious research goal is that of combining the envisioned prognostics techniques for all anomaly types using ensemble methods or other data driven techniques.

References

[1] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 657–667, Jan 2015.

[2] D. Braddon, K. Hartley *et al.*, "Aerospace competitiveness: Uk, us and europe," Tech. Rep., 2005.

[3] T. R. Kurfess, *Robotics and automation handbook*.

CRC press, 2004.

[4] M. P. Groover, *Automation, production systems, and computer-integrated manufacturing*. Pearson Education India, 2016.

[5] H. Xu, C. Caramanis, and S. Mannor, "Sparse algorithms are not stable: A no-free-lunch theorem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 1, pp. 187–193, 2012.

[6] I. E. Buchan, J. M. Winn, and C. M. Bishop, "A unified modeling approach to data-intensive healthcare." 2009.

[7] A. McAfee, E. Brynjolfsson, T. H. Davenport *et al.*, "Big data: the management revolution," *Harvard business review*, vol. 90, no. 10, pp. 60–68, 2012.

[8] T. Hey, S. Tansley, K. M. Tolle *et al.*, *The fourth paradigm: data-intensive scientific discovery*. Microsoft research Redmond, WA, 2009, vol. 1.

[9] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.

[10] G. Piateski and W. Frawley, *Knowledge discovery in databases*. MIT press, 1991.

[11] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning," in *Machine learning*. Springer, 1983, pp. 3–23.

[12] I. Y. Tumer and E. M. Huff, "Analysis of triaxial vibration data for health monitoring of helicopter gearboxes," *Transactions-American Society of Mechanical Engineers Journal of Vibration and Acoustics*, vol. 125, no. 1, pp. 120–128, 2003.

[13] T. Brotherton, G. Jahns, J. Jacobs, and D. Wroblewski, "Prognosis of faults in gas turbine engines," in *Aerospace Conference Proceedings, 2000 IEEE*, vol. 6. IEEE, 2000, pp. 163–171.

[14] S. W. Doebling, C. R. Farrar, M. B. Prime *et al.*, "A summary review of vibration-based damage identification methods," *Shock and vibration digest*, vol. 30, no. 2, pp. 91–105, 1998.

[15] T. Brotherton, P. Grabill, D. Wroblewski, R. Friend, B. Sotomayer, and J. Berry, "A testbed for data fusion for engine diagnostics and prognostics," in *Aerospace Conference Proceedings, 2002. IEEE*, vol. 6. IEEE, 2002, pp. 6–6.

[16] B. E. Parker Jr, T. M. Nigro, M. P. Carley, R. L. Barron, D. G. Ward, H. V. Poor, D. Rock, and T. A. DuBois, "Helicopter gearbox diagnostics and prognostics using vibration signature analysis," in *Optical Engineering and Photonics in Aerospace Sensing*. International Society for Optics and Photonics, 1993, pp. 531–542.

[17] J. Xu and L. Xu, "Health management based on fusion prognostics for avionics systems," *Journal of Systems Engineering and Electronics*, vol. 22, no. 3, pp. 428–436, 2011.

[18] A. Hess, G. Calvello, and T. Dabney, "Phm a key enabler for the jsf autonomic logistics support concept," in *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, vol. 6. IEEE, 2004, pp. 3543–3550.

[19] D. Scrimieri, N. Antzoulatos, E. Castro, and S. M. Ratchev, "Automated experience-based learning for plug and produce assembly systems," *International Journal of Production Research*, vol. 55, no. 13, pp. 3674–3685, 2017.

[20] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Computers & Chemical Engineering*, vol. 33, no. 4, pp. 795–814, 2009.

[21] L. Wang, M. Törngren, and M. Onori, "Current status and advancement of cyber-physical systems in manufacturing," *Journal of Manufacturing Systems*, vol. 37, no. Part 2, pp. 517 – 527, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278612515000400

[22] D. Wu, S. Liu, L. Zhang, J. Terpenny, R. X. Gao, T. Kurfess, and J. A. Guzzo, "A fog computing-based framework for process monitoring and prognosis in cyber-manufacturing," *Journal of Manufacturing Systems*, vol. 43, no. Part 1, pp. 25 – 34, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278612517300237

[23] G. W. Vogl, B. A. Weiss, and M. Helu, "A review of diagnostic and prognostic capabilities and best practices for manufacturing," *Journal of Intelligent Manufacturing*, Jun 2016. [Online]. Available: https://doi.org/10.1007/s10845-016-1228-8

[24] X. Jian-Xin and H. Zhong-Sheng, "Notes on data-driven system approaches," *Acta Automatica Sinica*, vol. 35, no. 6, pp. 668–675, 2009.

[25] A. Thakur and A. Krohn-Grimberghe, "Autocompete: A framework for machine learning competition," *arXiv preprint arXiv:1507.02188*, 2015.

[26] M. Zinkevich, "Rules of machine learning: Best practices for ml engineering," 2017.

[27] P. Domingos, *Master Algorithm*. Penguin Books, 2016.

[28] ——, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[29] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sept 2009.

[30] R. Ranawana and V. Palade, "Optimized precision - a new measure for classifier performance evaluation," in *2006 IEEE International Conference on Evolutionary Computation*, 2006, pp. 2254–2261.

[31] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[32] L. Valiant, *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ), 2013.

[33] M. Dobrska, H. Wang, and W. Blackburn, "Data-driven rank ordering—a preference-based comparison study," *International Journal of Computational Intelligence Systems*, vol. 4, no. 2, pp. 142–152, 2011.

[34] C. Hayes, "Coated rivet dies: A dramatic improvement in rivet interference profile," *SAE Technical Paper*, vol. 1, p. 2, September 2016.

[35] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[36] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Stanford, CA, 1995, pp. 1137–1145.

[37] E. Keogh and A. Mueen, *Curse of Dimensionality*. Boston, MA: Springer US, 2010, pp. 257–258. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_192

[38] C. J. Van Rijsbergen, *The geometry of information retrieval*. Cambridge University Press, 2004.

[39] M. Khondoker, R. Dobson, C. Skirrow, A. Simmons, and D. Stahl, "A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies," *Statistical methods in medical research*, vol. 25, no. 5, pp. 1804–1823, 2016.