Construction of fatigue criteria through Positive Unlabeled Learning

Olivier Coudray, Philippe Bristiel, Miguel Dinis, Christine Keribin, Patrick Pamphile

Construction of fatigue criteria through Positive Unlabeled Learning

Olivier Coudray[a,b], Philippe Bristiel[a], Miguel Dinis[a], Christine Keribin[b], Patrick Pamphile[b]

[a] *Stellantis, 2-10 Boulevard de l'Europe, 78300, Poissy, France*
[b] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France*

Abstract

Vehicles reliability is a major issue for automotive manufacturers. In particular, mechanical fatigue is an important preoccupation of the design office. In order to accelerate the development of new mechanical parts, car manufacturers want to rely more on numerical simulation and drastically reduce the number of validation tests on prototypes. To do this, they need efficient fatigue criteria, able to correctly identify critical zones on a numerical model. However, the current fatigue criteria used to post process numerical results fail to correlate well on fatigue test rig. In this paper, we first propose a probabilistic Dang Van criterion that accounts for the dispersion of fatigue results in a multiaxial setting. We then introduce a fatigue database built upon numerical results and fatigue test reports on automotive chassis components. A novel approach, based on Positive-Unlabeled learning (PU learning), is developed to leverage this source of data and improve the predictivity of the fatigue criterion. The methodology is applied to the fatigue database to illustrate the interest of the approach.

Keywords: Mechanics of materials, Fatigue design, Fatigue criteria, Supervised machine learning, Positive-Unlabeled learning.

1. Introduction

Metallic structures can wear out over time under the effect of external loads. This phenomenon happens in various industrial applications, including the automotive industry. During its service life, the structure of a vehicle is subjected to various mechanical stresses resulting from external loads transferred by the wheels and suspensions to the whole car. After a long duration of use, the accumulation of stresses combined with stress concentrations (due to the geometry of mechanical components, manufacturing processes...) can cause the initiation of micro-cracks on specific zones of the vehicle. These micro-cracks can progressively lead to the initiation of a macro-crack that propagates until the complete fracture of the mechanical component. This phenomenon, called *mechanical fatigue*, is extremely dangerous as it can lead to the sudden failure of a mechanical part under normal conditions of use without any excessive load. Fatigue is thus a dangerous and complex phenomenon depending on the choice of materials, the manufacturing processes, and the local stresses during the use of the vehicle [22, 3]. It is therefore essential, for car manufacturers, to ensure a good resistance of mechanical parts against fatigue. This is all the more crucial for safety parts of the vehicle (*e.g.* chassis components) where fatigue rupture represents a mortal danger for the driver and the passengers.

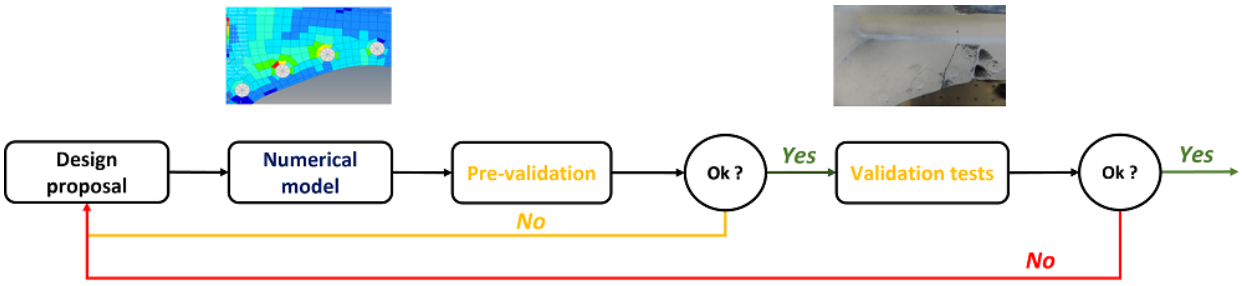

Figure 1: Design workflow from design proposal to validation.

During the conception of a mechanical part, a geometry is defined and materials are chosen. The design proposal is then modeled numerically using a Finite Element Model (FEM) which consists in a meshing of the part. For a given external load applied to the part, the resulting stress field on the part can be simulated. A fatigue criterion (Dang Van for instance) is then applied on each element of the model allowing to identify potential critical zones that are not resistant enough. Thanks to this criterion, the design proposal can be corrected by strengthening the zones subject to fatigue failure. Once the design proposal is satisfying, fatigue validation tests are performed on prototypes. Sometimes, the tests reveal that some critical zones remain. Hence, the design needs to be corrected, which requires an additional iteration over the whole design process (cf. Fig. 1). These iterations between design and validation strongly delay the development of the part and leads to additional costs, due to the additional physical validation tests that need to be performed.

The efficiency of the design criterion used to post process the numerical results is crucial in the design process. The objective of car manufacturers is to drastically reduce the number of physical tests and to tend towards a *full digital* design. Ideally, the design process should only require a unique validation test campaign at the end to check that the durability requirements are met. For that purpose, the car manufacturers seek to improve the efficiency of fatigue criteria in the identification of critical zones.

The modeling of fatigue has been extensively addressed in the literature [23, 3]. The first tool for modeling fatigue phenomena is the S-N curve. An S-N curve is built upon uniaxial fatigue tests on coupon specimens (elementary geometries). It consists in a model relating the number of cycles to failure N (fatigue lifetime) to the amplitude of sinusoidal load S. We can identify three domains for the stress S. *Low Cycle Fatigue* (LCF) refers to situations where the stress is greater or equal to the elastic limit of the material, and the fatigue lifetime is usually below 10^4 cycles. In *High Cycle Fatigue* (HCF), the lifetime is still finite ($10^4 - 10^6$ cycles) but the stress S is below the elastic limit. Finally, the endurance domain represent the stress range over which the fatigue lifetime is infinite. In practice, we consider the endurance limit S_{lim} to be the stress below which the fatigue lifetime is greater than 10^6 cycles, which represents the order of magnitude of a car's lifetime. The second tool for modeling fatigue in more complex situations is the fatigue criterion (cf. [27]). Given a mechanical part subjected to an external load (possibly multiaxial), a fatigue criterion allows to identify the zones that exceed the endurance limit of the material. For these critical zones, there is thus

a risk of failure.

Still, as stated above, the predictions provided by these criteria on FE results do not always correlate well with fatigue test rig results. Several limits can be identified on these criteria and on their calibration. First, most of fatigue criteria remain deterministic. The dispersion inherent to fatigue is not accounted for, or relies on *a priori* expert knowledge. It would thus be beneficial to estimate this variability in a multiaxial framework using experimental data. Second, the material parameters in the fatigue criteria are estimated through coupon uniaxial fatigue tests. These geometries are elementary and not necessarily representative of the complexity of zones encountered on real mechanical parts. Besides, the fatigue criteria only rely on a limited number of physical features (*e.g.* critical shear stress and hydrostatic stress in Dang Van criterion), which are not necessarily sufficient to describe a zone and the stress applied to it. For instance, stress concentrations due to the specific geometry are not well accounted for. Third, during validation test campaigns, tests are carried out according to different severities and durations (Locati test protocol, cf. [4]), whereas the calculations are done for a severity that represents a standardized *objective customer*. This can have an impact on the observability of crack initiation. For instance, a test under lower severity or interrupted too early may not allow some critical zones to fail and thus to be detected at testing.

In addition to the limitations mentioned above, it is crucial to note that the predictions using a fatigue criterion rely on the FEM results. These models do not account for several factors that can affect the fatigue properties of the materials: for instance, manufacturing processes and residual stresses. Moreover, the FEM only allow to approximate the mechanic behavior of a part and the results obtained are subject to errors due to this approximation. In particular, the FEM results on certain complex zones of a mechanical part (welds, edges, holes, corners) may remain uncertain. The improvement of the numerical model is beyond the scope of this article. However, the methodology presented in this article is general and could benefit from advances in FEM, that could also improve predictive performances of fatigue phenomena.

Over the years, automotive manufacturers have accumulated large amounts of data related to the design of mechanical parts: design proposals and physical test results. The history of numerical models and fatigue validation test reports for previous designs represents a rich and sizeable source of data that is currently not completely exploited. In this article, we develop new statistical methods to leverage this source of data in order to estimate fatigue criteria with better predictive performances. The issue is all the more critical in the automotive industry as the transition from thermal to electric cars necessitates the design of new vehicle platforms. In this context, it is crucial to capitalize on the experiments and results acquired on thermal vehicles in order not to restart from scratch.

This article offers a new approach based on Positive-Unlabeled learning (PU learning) to construct fatigue design criteria. The originality of these new criteria is their ability to account for *additional features* to those traditionally considered in fatigue models. Besides, the *impact of testing conditions* (severity, duration of the test) on the initiation of cracks is explicitly modeled as a propensity function that is part of the PU learning

model. Hence, we provide answers to two of the above stated limitations of fatigue criteria currently used in fatigue design.

The article is organized as follows. In Section 2, a probabilistic version of the Dang Van criterion is proposed, which allows a joint estimation of material parameters including dispersion. The methodology is applied to experimental data on welded coupon specimens but the estimated criterion does not generalize on more complex components. In Section 3, we introduce a fatigue database built upon numerical results and reports on fatigue tests carried out for different design proposals of chassis components. We propose a definition of zone to facilitate the analysis and provide new descriptive features to be accounted for in the fatigue criterion (*e.g.* geometry, information about singularities). Now, from a machine learning point of view, in Section 4, we define a PU learning classification model for the fatigue criterion, that integrates a propensity function specific to fatigue. The interest and efficiency of the proposed method are demonstrated on the fatigue database (Section 5).

2. Construction and estimation of a probabilistic Dang Van criterion

Fatigue criteria are deterministic criteria which are calibrated using fatigue tests on elementary structures called *coupon specimens*. Usually, the variability inherent to fatigue is either absent of the model, or integrated based on expert knowledge. In this section, we define a probabilistic fatigue criterion that integrates the variability observed in fatigue test results.

Our probabilistic fatigue criterion relies on Dang Van criterion commonly used in the automotive industry. After recalling the principles of Dang Van criterion (Subsection 2.1), we introduce a probabilistic version of it, which allows a joint estimation of the material parameters and of the dispersion inherent to fatigue (Subsection 2.2). The estimation is carried out using an experimental data set of welded coupon specimens (cf. [16]) and the estimated criterion is compared to experimental results on welds of real-scale chassis components (cf. Subsection 2.3).

2.1. Definition of the Dang Van criterion

The Dang Van criterion is a critical plane fatigue criterion based on considerations at the microscopic scale of the material [13, 1].

The danger coefficient associated to the Dang Van criterion, denoted DC, characterizes the level of criticality of an element given the stress cycle to which it is subjected. It depends on $P(t)$, the hydrostatic stress, and on $\tau_{mes}(t)$, the maximum mesoscopic shear stress[1] taken over all possible planes. The criterion considers a linear combination of both quantities maximized over the cyclic stress cycle of period T:

$$DC = \max_{0 \leq t < T} \frac{\tau_{mes}(t) + \alpha_M \, P(t)}{\tau_M} - 1 \ . \tag{1}$$

[1] Shear stress at the grain scale of the material.

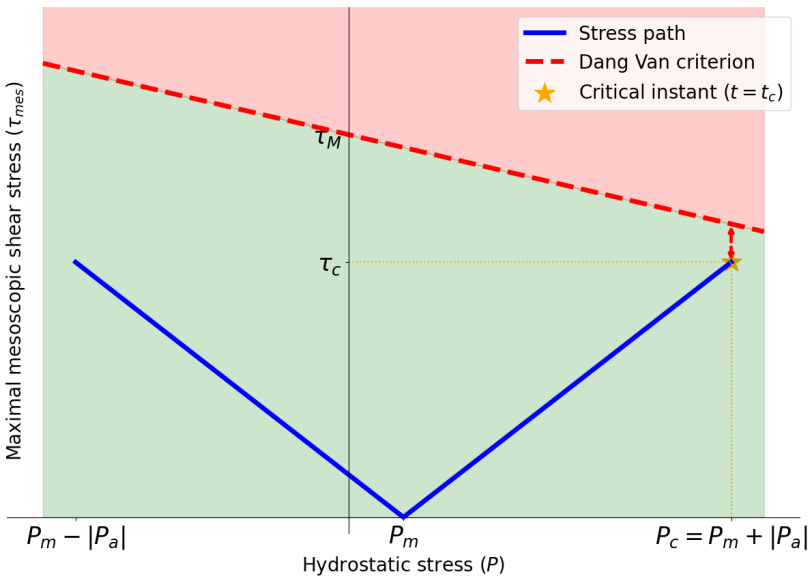

Figure 2: Example of stress path $(P(t), \tau_{mes}(t))_{0 \leq t < T}$ in Dang Van plane (blue curve) and representation of the fatigue criterion (red dashed line, with intercept τ_M and slope $-\alpha_M$). The element is safe if the stress path remains in the green zone, critical if it reaches the red zone.

Here, τ_M and α_M are material parameters that must be calibrated. The fatigue criterion characterizes the fatigue limit of the material M for a given target lifetime N_0. The element is critical if the danger coefficient DC exceeds 0, which means that the fatigue limit is exceeded for this load. Otherwise, the element is safe.

The stress path $(P(t), \tau_{mes}(t))$ can be represented in a Dang Van diagram featuring the hydrostatic stress along X-axis and the mesoscopic shear stress along Y-axis. In the scope of this article, the load cycles considered are sinusoidal and proportional, which implies that the stress paths necessarily have a "V" shape (cf. Fig. 2). The critical instant t_C over the stress cycles is then easily identified and the corresponding stresses $P_C = P(t_C)$ and $\tau_C = \tau_{mes}(t_C)$ are referred to as the critical hydrostatic stress and the critical shear stress. Given a FEM and the calculated stress field, one can compute these features and evaluate the Dang Van danger coefficient DC for each element of the FEM:

$$DC = \frac{\tau_C + \alpha_M P_C}{\tau_M} - 1 \ . \tag{2}$$

2.2. A probabilistic Dang Van criterion

The Dang Van line representing the criterion is adjusted in order to predict whether a zone has a probability lower or greater than 0.5 to be critical. Instead, we want to estimate the probability of criticality for every point in Dang Van diagram. Hence, we construct a probabilistic fatigue criterion based on Dang Van variables: critical hydrostatic and shear stresses (P_c, τ_c). In other words, given a point with coordinates (P_c, τ_c) in Dang Van plane, this probabilistic criterion should output a probability for this point to be critical.

In the uniaxial setting, it is common to deal with the scattering of fatigue results in S-N models [8, 17]. Here, we assume a Basquin model to relate the stress S to the number of cycles to failure N: $S \times N^b$ is

assumed to be constant where b is Basquin slope [2]. In addition, the variations of $\log(N)$ are assumed Gaussian. However, the Basquin model is not suited for the multiaxial setting because the stress is no longer univariate. Still, the S-N model can be extended to the multiaxial setting by replacing S with a multiaxial damage parameter [25, 24, 21, 9]. In the context of the Dang Van criterion, it is natural to consider a linear combination of the critical hydrostatic and shear stresses $\alpha\, P_c + \tau_c$ where α is an unknown parameter representing the slope of Dang Van criterion. This leads to the following regression model:

$$\log(N)\,|_{P_c,\tau_c} = a - b\,\log(\alpha\, P_c + \tau_c) + \sigma\,\varepsilon\ . \tag{3}$$

In the above equation, $\theta = (a, b, \alpha, \sigma)$ is the unknown parameter and ε is a standardized Gaussian noise. The components of θ represent physical fatigue parameters:

- a is related to the intercept of Dang Van criterion;
- b is Basquin slope, an important fatigue parameter;
- α is the slope of Dang Van criterion;
- σ represents the variability of the fatigue lifetime.

The model in Equation 3 leads directly to a probabilistic fatigue criterion. An element with stresses (P_c, τ_c) is considered critical if its lifetime is below N_0 cycles. This happens with probability:

$$p(P_c, \tau_c) = \mathbb{P}(N \leq N_0 \,|\, P_c, \tau_c) = \Phi\left(\frac{\log(N_0) - a + b\,\log(\alpha\, P_c + \tau_c)}{\sigma}\right) \tag{4}$$

where Φ denotes the cumulative distribution function of the standardized normal distribution. The probability $p(P_c, \tau_c)$ is a function of the parameter θ. Hence, by plugging in an estimate of the unknown parameter $\widehat{\theta} = (\widehat{a}, \widehat{b}, \widehat{\alpha}, \widehat{\sigma})$ into Eq. 4, we obtain an estimate of the probability that a zone with stresses (P_c, τ_c) is critical:

$$\widehat{p}(P_c, \tau_c) = \Phi\left(\frac{\log(N_0) - \widehat{a} + \widehat{b}\,\log(\widehat{\alpha}\, P_c + \tau_c)}{\widehat{\sigma}}\right)\ .$$

2.3. Probabilistic criterion calibration on Fayard welded coupon specimens

We can now implement and apply the methodology described in the previous subsection. The data set used for the estimation consists in $n = 144$ independent observations of fatigue lifetimes $(N_i)_{1\leq i\leq n}$ of welded coupon specimens for given values of critical hydrostatic and shear stresses $(P_{c,i}, \tau_{c,i})_{1\leq i\leq n}$ [16]. Several elementary structures and loading conditions are tested (cf. Fig. 3). We use this data set to estimate the parameter θ which characterizes the probabilistic fatigue criterion of Equation 3. The likelihood $\ell(\theta)$ can be expressed as:

$$\ell(\theta) = -\frac{n}{2}\log(2\,\pi\,\sigma^2) - \sum_{i=1}^{n}\frac{\left[\log(N_i) - a + b\log(\alpha P_{c,i} + \tau_{c,i})\right]^2}{2\,\sigma^2}\ . \tag{5}$$

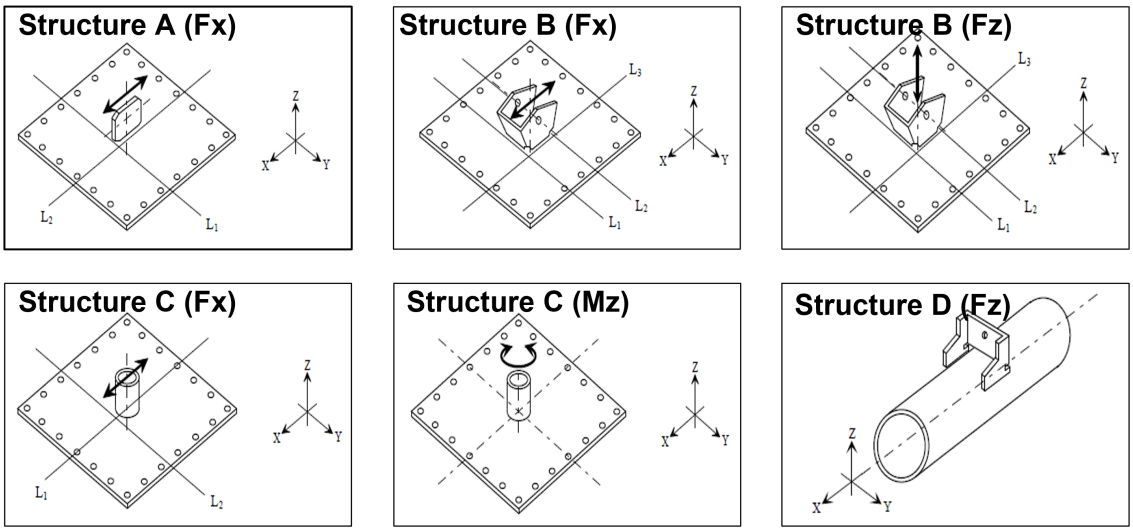

Figure 3: Fayard coupon specimens: geometries and loadings (cf. [16])

An estimator $\widehat{\theta}$ of the parameter is obtained through maximum likelihood. The maximization of the log-likelihood in Eq. 5 cannot be solved analytically because of the non-linear term in the regression. Hence a numerical approximation is found using Newton method. In fact, the main difficulty lies in the estimation of α. Once $\widehat{\alpha}$ is known, the other coefficients have a closed form expression:

$$\widehat{b} = -\frac{\sum_{i=1}^{n}\left(\log(N_i) - \overline{\log(N)}\right)\left(\log(\widehat{\alpha}P_{c,i}+\tau_{c,i}) - \overline{\log(\widehat{\alpha}P_c+\tau_c)}\right)}{\sum_{i=1}^{n}\left(\log(\widehat{\alpha}P_{c,i}+\tau_{c,i}) - \overline{\log(\widehat{\alpha}P_c+\tau_c)}\right)^2}$$

$$\widehat{a} = \overline{\log(N)} + \widehat{b}\,\overline{\log(\widehat{\alpha}P_c+\tau_c)}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\log(N_i) - \widehat{a} + \widehat{b}\log(\widehat{\alpha}P_{c,i}+\tau_{c,i})\right)^2$$

where $\overline{\log(N)}$ and $\overline{\log(\widehat{\alpha}P_c+\tau_c)}$ denote the empirical means:

$$\overline{\log(N)} = \frac{1}{n}\sum_{i=1}^{n}\log(N_i)$$

$$\overline{\log(\widehat{\alpha}P_c+\tau_c)} = \frac{1}{n}\sum_{i=1}^{n}\log(\widehat{\alpha}P_{c,i}+\tau_{c,i})\ .$$

The estimates are presented in Table 1 along with their 95% asymptotic confidence intervals. The estimate of α is very close to the slope found by Fayard. In addition, the value of Basquin fatigue parameter b is perfectly standard for welds [7]. Finally, we provide an estimate of the variability through σ.

The fit can be visualized in a "S-N like" diagram by representing the lifetime on the X-axis and $\widehat{\alpha}\,p_c+\tau_c$ on the Y-axis (cf. Fig. 4). The probabilistic Dang Van criterion is represented in Fig. 5. This criterion is no longer represented as a line in Dang Van diagram but as a probability field. It is important to note that the criterion identified by Fayard appears in the 95% confidence interval of our maximum likelihood estimate of

Table 1: Maximum likelihood estimates with the 95% asymptotic confidence intervals.

	inf 95%	mean	sup 95%
a	34.37	37.21	40.06
b	4.39	4.94	5.48
α	0.26	0.35	0.44
σ	0.66	0.75	0.83

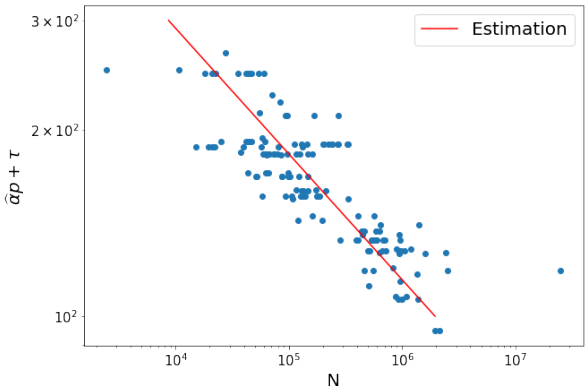

Figure 4: Maximum likelihood estimated regression line in an S-N like diagram (log scale on both axes): Y-axis represents the linear combination of P_c and τ_C with the estimated slope $\widehat{\alpha}$. The experimental points represents the fatigue test results for Fayard coupon specimens.

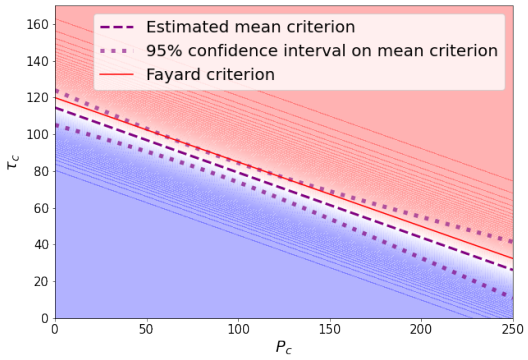

Figure 5: Illustration of probabilistic Dang Van criterion with percentile curves: blue (red) represents low (high) probability of failure before 10^6 cycles. The dashed curves represents the 50% criterion estimated with our method and the 95% confidence interval on this estimate. The red line is the criterion estimated by Fayard [16].

the 50% criterion, which confirms the coherence of our results.

The relevance of this probabilistic fatigue criterion is then assessed by overlaying experimental data points from coupon specimens used in the calibration (Fig. 6, left) but also from real-scale mechanical parts (Fig. 6). If the criterion seems satisfying for explaining crack initiations on the first data set, it clearly fails in generalizing when used for more complex mechanical parts. Indeed, multiple critical welds are poorly identified by the criterion.

3. Fatigue database

The generalization of the probabilistic criterion calibrated on coupon specimens to chassis components faces two limitations. On the one hand, the data set of coupon specimens used for its estimation is not representative of the variety and complexity of zones found in real mechanical parts. On the other hand, Figure 6 shows that the two features accounted for in the criterion (P_C and τ_C) are not sufficient to characterize the criticality of a zone. The probabilistic criterion leads to many false positives (blue points in the red zone) and

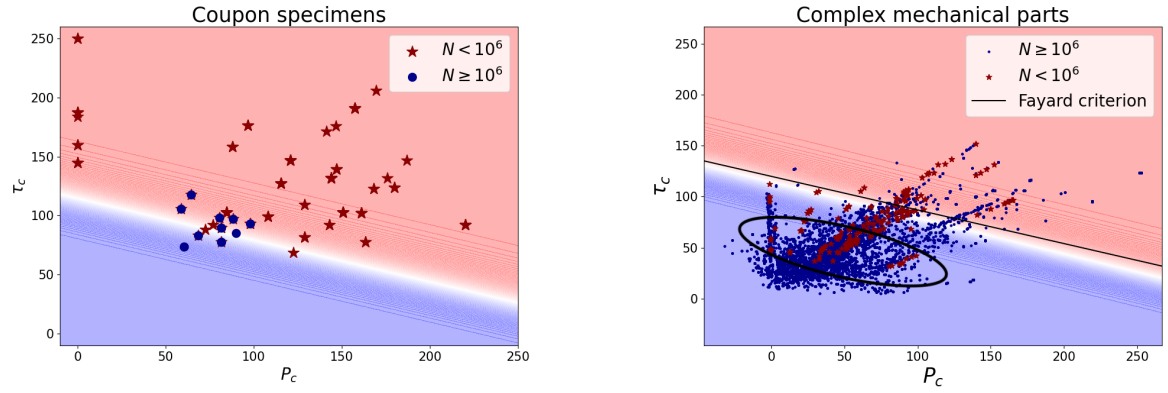

Figure 6: Probabilistic Dang Van criterion and welded zones from coupon specimens (left) and real-scale mechanical components (right). The position of the points depends on the critical hydrostatic and shear stresses calculated through FEM: red stars (blue dots) represent zones with observed lifetime below (over) 10^6 cycles. The background color represents the estimated probabilistic criterion. The blue and red lines are the percentile curves. The critical welds that are poorly identified by the criterion are highlighted in the right figure (ellipse).

false negatives (red points in blue zone). The former represent a safety issue (non-identified critical zones) while the latter may impose unnecessary reinforcement on the design. Hence, we are addressing a supervised classification task. To take into account a larger number of variables, we introduce in this section a fatigue database built upon numerical results and reports of fatigue tests carried out for different design proposals of chassis components (cf. Subsection 3.1). Rather than considering each elements of FEM independently in our approach, we provide a definition of *zone*, which facilitates and enhances the analysis (cf. Subsection 3.2). Each zone of the database can be described by several features that provide additional information to the critical hydrostatic and shear stresses considered in Dang Van criterion (cf. Subsection 3.3). Finally, we have access to the testing conditions and in particular to the severity applied on test rig, which differs from the nominal severity applied on the numerical model (Subsection 3.4).

3.1. Using fatigue data representative for complex components

From now on, we rely on a fatigue database that is built upon different design proposals of cradle and cross-member models under different loading conditions. A case study consists in a numerical model of a mechanical part along with the numerical simulation results for one type of loading. The database contains a total of 39 case studies with two types of mechanical parts (cradles and cross-members), several geometries for each type of part and different types of solicitations (longitudinal, cornering, vertical, transversal).

For each case study, the simulation results give access, for each element of the FEM, to different information describing the coordinates of the element, the material, the type of element (sheet, sheet edge, weld) and the local stresses. In addition, fatigue rig tests are carried out on prototypes under the same loading and the test reports provide information on the crack initiations detected: photos of the cracks, severity of the test, duration... This way, we can identify on the numerical model which elements actually failed during testing (see Fig. 7).

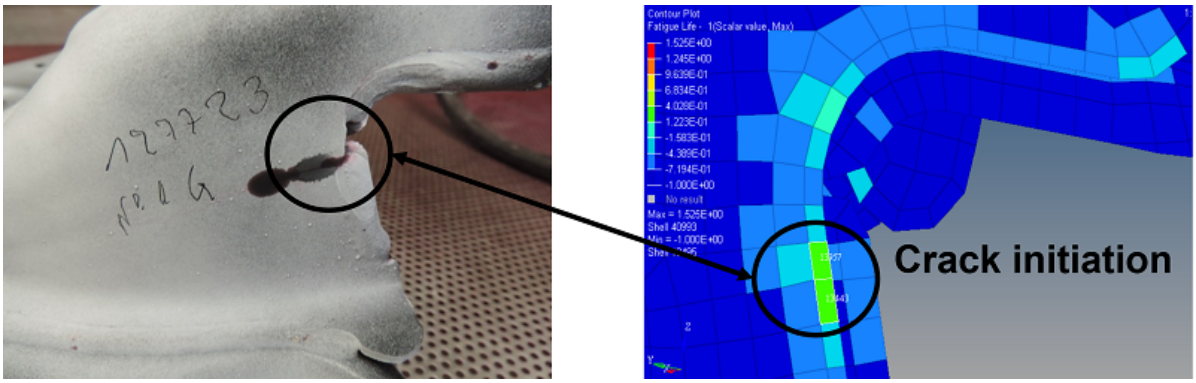

Figure 7: Correspondence between numerical models and test results: on the left, crack initiation location from a test report; on the right, the corresponding zone on the FEM. The color scale on the FEM represents the danger coefficient.

3.2. From elements to zones: changing the unit of analysis

By definition, the fatigue database provides several information describing each element of the FEM. The analysis of this database raises various difficulties. First, the data set is extremely imbalanced. Indeed, each numerical model contains about 10^5 elements and usually no more than 10 crack initiations. Second, the stress remains low for the majority of the elements of the mechanical parts. Hence we have only a few potentially dangerous locations on the part. Third, the stress field over the mechanical part is continuous. Therefore stress values on close elements are highly correlated and provide similar information. Finally, it is not always straightforward to locate precisely the element responsible for the crack initiation. It appears that the large majority of cracks initiate and propagate near singularities (edges, holes, corners, welds). Hence, in order to model the risk of crack initiation, it may be relevant to account for features that describe the whole zone and not just a single element.

To address these challenges, we introduce a new unit of analysis: instead of considering each elements as an individual of the database, we rather focus on zones. The construction of these zones is based on two main principles. The first one is that the analysis should focus only on relevant zones, which means that we will only consider zones with at least one element with a DC greater than -0.8. The second one is that, as fatigue is a local phenomenon, any statistical criterion defined should remain local; hence we limit the radius of each zone to 25 millimeters. Hence, a zone is defined as a group of elements built around a center (element with DC greater than -0.8) and containing all the elements connected to the center through the FEM within a radius of 25 millimeters. The threshold on DC and the size of the zone were chosen empirically. The first parameter is both small enough to select every marked element (with a detected crack initiation) and high enough to limit the number of selected points. The second parameter allows to take into account singularities located near critical points, which are relevant for our study.

A zone is labeled positive ($Y = 1$) if a crack was detected during the test and negative otherwise ($Y = 0$). This pre-processing drastically reduces the number of observations: instead of millions of elements, we have $n = 19\ 367$ zones among which 291 are crack initiations (labeled positive). Meanwhile, it provides access

to richer information: we not only have access to the descriptive features of one element but also for every elements of the zone.

3.3. Descriptive features for zones

In our analysis, each zone i of the database is described by five features $(\tau_{c,i},\, e_i,\, w_i,\, \tau_{mat,i}, h_i)$:

- the critical shear stress of the most critical element of the zone ($\tau_{c,i}$). This feature was already part of the probabilistic criterion.
- the maximum longitudinal stress over the edges of the zone (e_i). This descriptor is calculated by considering a local coordinate system on the edge elements of the zone and projecting the stress tensor in this local basis. Only the maximum is retained, which represents the most dangerous edge of the zone.
- the maximum transversal stress over the welds of the zone (w_i). Similarly, this stress is computed by considering a local coordinate system around each weld element of the zone.
- the material parameter (intercept of the Dang Van line for the material, known through coupon tests) averaged over the elements of the zone ($\tau_{mat,i}$). This descriptor is relevant because it takes into account the heterogeneity of the zone: the majority of zones considered contain at the same time plain sheet, edge and weld elements that have different material parameters.
- the thickness of the metal sheet (h_i).

This information is stored in a covariate vector $x_i = (\tau_{c,i},\, e_i,\, w_i,\, \tau_{mat,i}, h_i)$. These features provide diverse information about the zone: stress on the most critical element of the zone and on the main singularities (main weaknesses of the zone), geometric information, and material information. In particular, the features e_i, w_i and $\tau_{mat,i}$ plainly exploit the notion of zone by accounting for singularities close to the critical element and by considering the spatial average of a physical quantity.

3.4. Testing conditions

For each zone i of the data set, the stresses stored in x_i are obtained through the simulation results on the FEM of the part. The load considered in the simulation has a nominal severity that represents a severe customer (called the *objective customer*). This objective customer severity is defined as part of the Stress-Strength interference method for fatigue design [26].

In fact, the testing conditions on test rig do not exactly reproduce the load simulated on the numerical model: the directions of external forces are respected but the overall severity is different and incremented during the test. The severity is represented by a scalar f which is the percentage of the nominal severity (used on the FEM). During fatigue tests, the severity f is not necessarily 1. This is part of the accelerated test protocol called *Locati* that consists in increasing gradually the severity during the test to bring every

component to failure [18]. A test begins at an initial severity f_0. Then, every n_{inc} cycles, the severity is increased by f_{inc}. For each observation i in the data set, the testing conditions are characterized by a vector $u_i = (f_i, n_i)$ where f_i denotes the initial severity of the test and n_i is its total duration. The increment parameters (f_{inc}, n_{inc}) remain fixed and are identical across the tests. We will see that the testing conditions $(u_i)_{1 \leq i \leq n}$ play a crucial role in the initiation of cracks and must therefore be taken into account when estimating the fatigue criterion.

4. Fatigue criterion under the point of view of PU learning

In the previous section, we introduced a new source of fatigue data representative of the complexity of some automotive mechanical parts. The purpose is now to leverage this data set in order to improve the predictivity of critical zones on new design. It is first crucial to remark that the prediction of critical zones is a classification task (Subsection 4.1). In this sense, the data set at hand can be used to estimate a classification rule able to predict whether a zone with descriptive features x is critical. However, the observed labels (Y_i) (crack initiation or not) only provide limited information on the criticality of the zones: while a crack initiation asserts the criticality of a zone, the absence of crack is not a proof of safety. Hence, the classification task is not standard as there is a selection bias affecting the observations (Subsection 4.2). Therefore, we propose a parametric PU learning model for the fatigue criterion that accounts for this specificity (Subsection 4.3) and develop a method to estimate its parameters (Subsection 4.4).

4.1. Fatigue criteria calibration as a classification task

For a mechanical part to be valid, every zone must be set below the endurance limit. In other words, there should not be any crack initiation over the car lifetime. Therefore, for a zone i with covariates x_i, we seek to predict a binary class Z_i indicating whether the zone may fail over the car lifetime ($Z_i = 1$, critical) or not ($Z_i = 1$, safe).

A solution is then to use the fatigue data set $(x_i, y_i)_{1 \leq i \leq n}$ as a training set to estimate a fatigue criterion. On the one hand, we do not want this criterion to fail in identifying zones that could break: indeed, in that case, the designed model would fail the validation tests and thus require an additional iteration. On the other hand, the criterion should not be too strict in order to avoid useless reinforcements. As Z is a binary target, this problem is a *binary classification task*. A statistical criterion can be estimated using the fatigue data set.

The use of supervised classification techniques as fatigue criteria was first investigated in [11]. By using all the available features, linear and non-linear supervised classification methods (Logistic Regression, Linear discriminant Analysis, Support Vector Machine, Random Forests) were proved to achieve better classification

performances than the Dang Van mechanical criterion. Thus, these methods offer a better identification of potential crack initiations.

However, there is a subtle and crucial difference between the observed outputs of the tests (Y, presence or absence of crack initiation) and the quantity we seek to predict (Z, critical or not). Indeed, the objective of a fatigue criterion is to identify the critical zones of a mechanical part, *i.e.* those that could fail before the target lifetime $N_0 = 10^6$ cycles, under the objective customer load. The fatigue tests only provide a subset of the critical zones that resulted into crack initiation during testing. Therefore, some critical zones ($Z = 1$) remain unlabeled ($Y = 0$, no crack initiation). Conversely, the presence of crack initiation ($Y = 1$) asserts that a zone is critical ($Z = 1$). Accounting for this asymmetric label noise is crucial in the context of fatigue design.

4.2. A PU learning model for the fatigue criterion

The probability that a zone to be critical given its covariate $x = (\tau_c,\, e,\, w,\, \tau_{mat}, h)$ is denoted $\eta(x)$. This quantity does not depend on the testing conditions as it is meant to characterize the criticality of a zone when the part is subjected to the nominal objective customer load.

$$\eta(x) = \mathbb{P}\left(Z = 1 \,|\, X = x\right) \ .$$

As explained in the previous subsection, a crack initiation ($Y = 1$) is necessarily a critical zone ($Z = 1$):

$$\mathbb{P}\left(Z = 1 \,|\, Y = 1, X = x\right) = 1 \ .$$

Reciprocally, the probability that a critical zone will fail during testing is not necessarily 1 and may depend on the features $x = (\tau_c,\, e,\, w,\, \tau_{mat}, h)$ and the testing conditions $u = (f, n)$. This quantity is called the *propensity* and is denoted e:

$$e(x, u) = \mathbb{P}\left(Y = 1 \,|\, Z = 1, X = x, U = u\right) \ .$$

All in all, the probability that the observed output Y is 1 (crack initiation) is modeled as the product

$$\mathbb{P}\left(Y = 1 \,|\, X = x, U = u\right) = \eta(x) \times e(x, u) \ . \tag{6}$$

In this expression, the quantity of interest is the fatigue criterion η. The problem of estimating η in this setting is called *Positive-Unlabeled learning* (PU learning, cf. [15, 6]), as we only have access to an incomplete set of positive data (crack initiations) and unlabeled zones (the rest) for which the class Z is unknown. The propensity e represents the probability that a critical zone will fail under specific testing conditions. As this quantity depends on the covariates, it operates as a selection bias. While most approaches in PU learning assume that the propensity e is constant (Selected Completely At Random assumption, SCAR), in this work we have to consider its dependence in the covariates (x, u) (Selected At Random assumption, SAR). This setting is hence more challenging and has only received few attention in the literature.

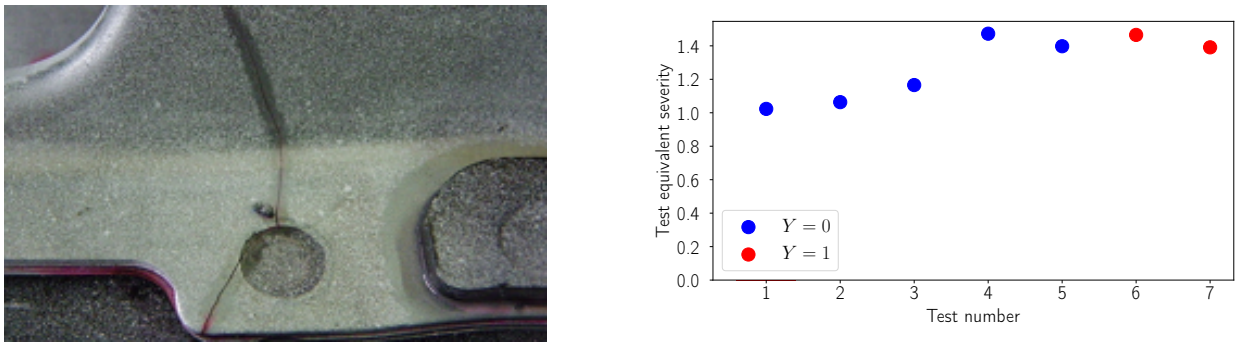

Figure 8: Example of critical zone on a cradle model under longitudinal loading. The picture on the left represents the zone on the sixth specimen tested. The figure on the right represents the severity (multiplicative coefficient of the client objective F_n) of the seven tests performed. Only two (the red ones) resulted into a crack initiation.

This mislabeling is observable considering the multiplicity of the tests performed for a same design. For instance, Fig. 8 (left) represents a crack initiation detected on a cradle part, that asserts the criticality of the zone. Among the seven identical prototypes tested, only two resulted in a crack initiation at this specific location (see Fig. 8, right). This also means that if only the first five prototypes had been tested, the critical zone would not have been labeled. It is then likely that several critical zones (for which $Z = 1$) remain unlabeled ($Y = 0$). It is important to note that this label noise is completely asymmetric: we only have false negatives (unlabeled positive instances) but no false positive (labeled negative instance).

The testing conditions u can influence the propensity in several ways. A higher severity can accelerate the initiation of a crack in a critical zone. Furthermore, increasing the duration of the test will leave more time for a crack to initiate and propagate enough to be observable. Hence, the severity and the number of cycles both have an influence on the propensity. Usually, we rely on a single variable to represent the testing conditions: the equivalent severity that depends on the initial severity of the test and the total number of cycles. This notion will be explained in more details in the next subsection. Figure 8 already illustrates the effect of equivalent severity on propensity as we clearly see that the critical zone broke for two of the most severe tests. We can confirm this statement looking at the severity for every known critical zone of the data set, *i.e.* those that broke at least for one test among the repetitions (cf. Fig. 9, left). Even if the two histograms seem close, a rough estimate of the propensity for each bin (Fig. 9, right) asserts the increasing trend of propensity when the equivalent severity increases.

The propensity also depends on the observability of a crack. A crack in a hidden location on the mechanical part is less likely to be detected. Likewise, the size of the crack and the effort to detect crack initiations on the part have an influence on the mislabeling. In some testing experiments, penetrant inspection is used to help detect crack initiations. This makes the detection of cracks easier and thus increases the propensity. Finally, let us recall that several cracks can initiate on a same part during testing. Sometimes, different cracks initiate on close locations. Hence, there can be a dependence effect that facilitates the initiation of cracks around an already broken zone or making it harder. Unfortunately, these parameters are not easily

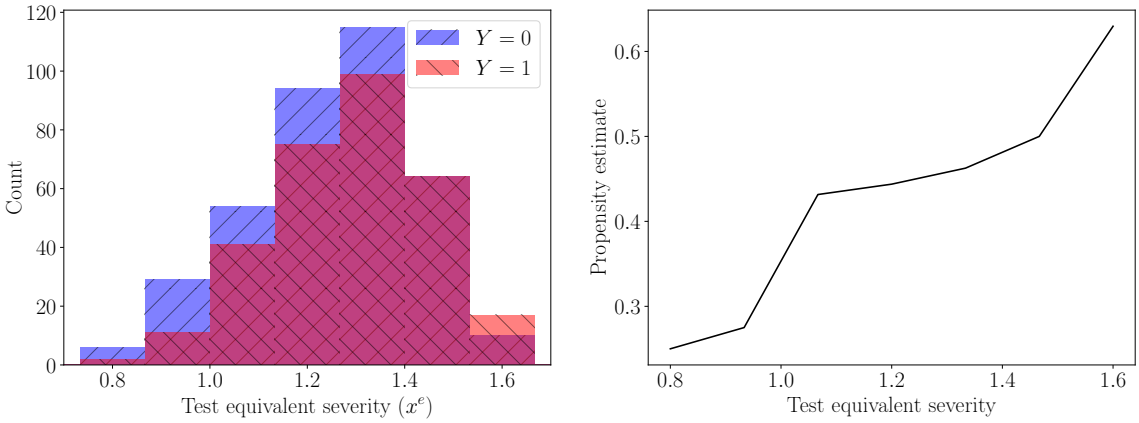

Figure 9: Influence of test severity on propensity. On the left, blue (red) histogram represents the empirical distribution of test severity for unbroken (broken) critical zones. Each bin leads to an estimate of the propensity (as the frequency of crack initiations among the total number of observations in the bin) represented on the right.

accessible and thus cannot be properly accounted for. Hence we will stick to a propensity that depends only on the available information, *i.e.* the testing covariates u.

4.3. Parametric PU learning model

Our objective is to predict the criticality Z of a zone with covariates x. For that purpose, we need to provide a model for η. As η represents the design criterion, it should only depend on the covariates x and not on the testing conditions u that only affects the results of the test. Moreover, since only $(y_i)_{1\leq i\leq n}$ are observed and not $(z_i)_{1\leq i\leq n}$, we also need to provide a model on the propensity e. Contrary to the design criterion, the propensity depends on the testing conditions u: for two zones equally critical (*i.e.* for which $\eta(x)$ is similar), the zone tested under the higher severity or over a longer time is more likely to break. This way, we can provide the following decomposition of the probability of crack initiation given the covariates:

$$\begin{aligned}\mathbb{P}\left(Y=1\,|\,X=x,U=u\right) &= \mathbb{P}\left(Z=1\,|\,X=x\right)\times\mathbb{P}\left(Y=1\,|\,Z=1,U=u\right)\\ &= \eta(x)\times e(u)\ .\end{aligned} \tag{7}$$

Contrary to Eq. 6, the propensity now depends only on the testing conditions u.

We choose parametric models for both. Hence, from now on, a PU learning model will consist in a couple of parametric models described by parameters (θ,ϕ), where θ characterizes the fatigue criterion and ϕ the propensity. The conditional distribution of Y given $X=x$ and $U=u$ is denoted $\mathbb{P}_{\theta,\phi}$:

$$\mathbb{P}_{\theta,\phi}\left(Y=1\,|\,X=x,U=u\right)=\eta_\theta(x)\times e_\phi(u)\ . \tag{8}$$

We now provide explicit parametric models for the classification rule and the propensity.

4.3.1. Fatigue criterion

To model the fatigue criterion, we provide a parametric model of the conditional probability of Z given x. The probability for a zone to be critical is assumed to be a function of a linear combination of the descriptive features $x = (\tau_c,\, e,\, w,\, \tau_{mat}, h)$. We choose a linear logistic regression to model the probability for a zone to be critical $\eta_\theta(x)$:

$$\eta_\theta(x) = \frac{1}{1 + e^{-\alpha_0 - \alpha^T x}} \tag{9}$$

where $\theta = (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{R}^5$.

4.3.2. Propensity models

Since the propensity represents a probability of crack initiation for a critical zone (a zone with finite lifetime), S-N models can provide useful ways to model it. However, the situation more complex than a standard S-N model because the tests are not carried out at constant amplitude: the severity is incremented by f_{inc} every N_{inc} cycles according to a Locati protocol (cf. Fig. 10). A solution to this issue is to seek for a constant severity $f_{eq}(u)$ that induces a fatigue damage equivalent to the testing conditions u. The computation of this so-called *fatigue equivalent* relies on two ingredients:

1. a cumulative damage law (in our case, we consider Miner rule, which is the simplest and the most frequently used, [20, 19]);
2. an S-N curve for the material (in our case, the S-N curve is assumed to follow a Basquin model with a known slope b).

We recall that $u = (f, n)$ where f denotes the initial severity and n is the total number of cycles. Let q and r denote the quotient and remainder of the Euclidean division of n by N_{inc}, the fatigue equivalent at N_0 cycles, $f_{eq}(u)$, is given by the following formula (cf. [4]):

$$f_{eq}(u) = \left[\frac{1}{N_0} \sum_{j=0}^{q-1} \left[N_{inc}\, (f + j\, f_{inc})^b \right] + \frac{r}{N_0}\, (f + q\, f_{inc})^b \right]^{\frac{1}{b}} . \tag{10}$$

The interpretation of $f_{eq}(u)$ is as follows: the Locati test under conditions u is equivalent, in terms of fatigue damage, to a constant-amplitude test over N_0 cycles at constant severity $f_{eq}(u)$ (cf. Fig. 10). Hence, the probability of observing a crack initiation before the end of the test is equal to the probability of observing a crack initiation before N_0 cycles at constant amplitude $f_{eq}(u)$. This probability can be modeled as:

$$e_\phi(u) = F\left(\log\left(\beta\, f_{eq}(u)\right)\right) \tag{11}$$

where F belongs to a parametric family of cumulative distribution functions and $\beta \in \mathbb{R}_+$ is a scalar parameter.

There are several choices for the parametric family to which F belongs: we have selected a *log-normal fatigue model*, a *Weibull fatigue model* and a logistic model.

In the log-normal fatigue model, the fatigue lifetime follows a log-normal distribution, which means that the logarithm of fatigue lifetime follows a normal distribution, hence F belongs to the family $(F_\sigma)_{\sigma \in \mathbb{R}_+^*}$,

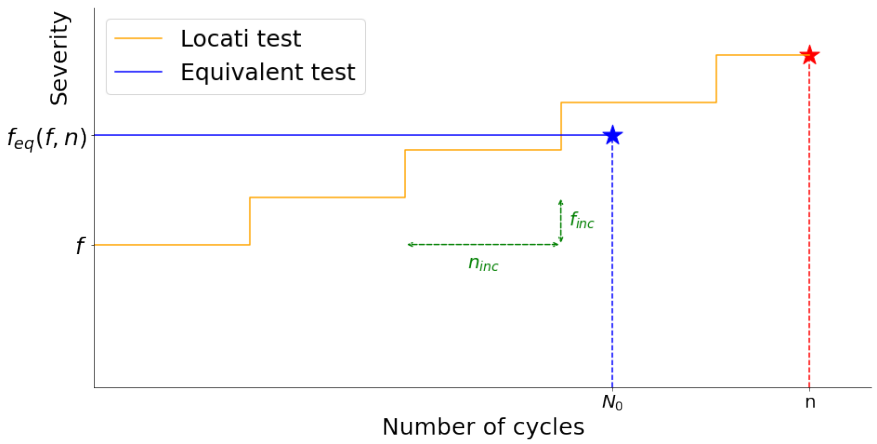

Figure 10: Illustration of a Locati test with conditions $u = (f, n)$. The orange line represents the Locati test: severity (Y-axis) as a function of the number of cycles (X-axis). The blue line represents the equivalent test at constant severity over N_0 cycles. In terms of fatigue damage, the equivalent test is equivalent to the Locati test.

where F_σ denotes the cumulative distribution function of a centered normal distribution with variance σ^2. The propensity e_ϕ is then represented by the set of parameters $\phi = (\beta, \sigma)$.

In the Weibull fatigue model, the fatigue lifetime follows a Weibull distribution, hence its logarithm follows a Gumbel distribution: F belongs to $\left\{F_a,\ a \in \mathbb{R}_+^*\right\}$, where F_a denotes the cumulative distribution function of a Gumbel distribution with parameter a:

$$F_a(v) = 1 - e^{-\exp\left(\frac{v}{a}\right)} \ .$$

The propensity e_ϕ is then represented by the set of parameters $\phi = (\beta, a)$.

The above two models are derived from classical S-N models in the literature and are thus well suited for fatigue applications [cf. 8]. In addition, we also consider a general statistical model consisting of a logistic regression based on the equivalent severity $f_{eq}(u)$ (*logistic propensity*):

$$e_\phi(u) = \frac{1}{1 + e^{-\beta_0 - \beta f_{eq}(u)}} \ .$$

The set of parameters is $\phi = (\beta_0, \beta) \in \mathbb{R}^2$.

4.4. Parameter estimation

The parametric model $(\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi\in\Phi}$ introduced in the previous subsection (cf. Eq. 8, 9 and 11) is identifiable (cf. [10], Chap. 4). In this subsection, we deal with the estimation of the parameters on training observations through maximum likelihood. Let us consider the n independent observations $(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}) = (x_i, u_i, y_i)_{1\leq i\leq n}$ and denote $\ell(\theta, \phi \,|\, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y})$ the log-likelihood. The objective is to find the parameters $(\widehat{\theta}, \widehat{\phi})$ maximizing the log-likelihood:

$$\left(\widehat{\theta}, \widehat{\phi}\right) = \underset{\theta,\phi}{\text{Argmax}}\ \ell\Big(\theta, \phi \,|\, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}\Big) \ . \tag{12}$$

Since we are in the presence of missing data (latent classes $(z_i)_{1\leq i\leq n}$), we will use the Expectation Maximization (EM) algorithm to maximize the likelihood. After presenting the general principle of the EM algorithm (Paragraph 4.4.1), we provide the details concerning its implementation in the context of our PU learning model (Paragraph 4.4.2).

4.4.1. EM algorithm

The EM algorithm, introduced by [14], enables the calculation of maximum of likelihood estimates for models with latent variables. The algorithm consists in iterating through an expectation step (E step) and a maximization step (M step) (cf. Algorithm 1). The likelihood increases at each iteration, and the algorithm stops when it reaches a local maximum.

Algorithm 1 General EM algorithm for PU learning

Initialization: start with initial parameters $\left(\theta^{(0)}, \phi^{(0)}\right)$

Iterate until convergence:

E step Given the parameters $\left(\theta^{(c)}, \phi^{(c)}\right)$ obtained at step c, compute the conditional expectation of the complete log-likelihood:

$$Q^{(c)}\left(\theta, \phi\right) = \mathbb{E}\left[\ell\big(\theta, \phi \,|\, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Z}, \boldsymbol{Y}\big) \,\Big|\, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}, \widehat{\theta}^{(c)}, \widehat{\phi}^{(c)}\right]$$

M step Maximize the conditional expectation over (θ, ϕ):

$$\left(\widehat{\theta}^{(c+1)}, \widehat{\phi}^{(c+1)}\right) = \underset{\theta,\phi}{\text{Argmax}}\ Q^{(c)}\left(\theta, \phi\right)$$

4.4.2. Estimation of the fatigue criterion

First, let us recall the model given by Eq. 8:

$$\mathbb{P}_{\theta,\phi}\left(Y = 1 \,|\, X = x, U = u\right) = \eta_\theta(x) \times e_\phi(u)\ .$$

The likelihood of the observations is:

$$\ell\Big(\theta, \phi \,|\, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}\Big) = \sum_{i=1}^{n} \Big[y_i \log\Big(\eta_\theta(x_i)\, e_\phi(\boldsymbol{u}_i)\Big) + (1 - y_i) \log\Big(1 - \eta_\theta(x_i)\, e_\phi(u_i)\Big)\Big]\ . \tag{13}$$

The use of the EM algorithm in the estimation of parametric PU learning models was first proposed by [5]. We now provide the detailed calculations regarding the expectation and maximization steps in our specific setting, with a propensity given by one of the models listed in Paragraph 4.3.2.

Expectation step. Let us assume that the parameter after iteration c in the EM algorithm is $\left(\theta^{(c)}, \phi^{(c)}\right)$. We want to calculate $Q^{(c)}\left(\theta, \phi\right)$. Thanks to the linearity of the expectation, we only need to compute the posterior probabilities $\gamma_i^{(c)}$:

$$\gamma_i^{(c)} = \mathbb{E}\left[Z_i \,\big|\, x_i, u_i, y_i, \theta^{(c)}, \phi^{(c)}\right] = \mathbb{P}_{\theta^{(c)}, \phi^{(c)}}\left(Z_i = 1 \,\big|\, x_i, u_i, y_i\right)\ .$$

We recall that, in PU learning, a labeled instance ($y_i = 1$) is necessarily positive, hence:

$$\mathbb{P}_{\theta^{(c)},\phi^{(c)}}\left(Z_i = 1 \,\middle|\, x_i, u_i, Y_i = 1\right) = 1 \ .$$

The posterior probability for unlabeled instances can be computed using the Bayes theorem:

$$\begin{aligned}
\mathbb{P}_{\theta^{(c)},\phi^{(c)}}\left(Z_i = 1 \,\middle|\, x_i, u_i, Y_i = 0\right) &= \mathbb{P}_{\theta^{(c)},\phi^{(c)}}\left(Z_i = 1 \,|\, x_i, u_i, Y_i = 0\right) \\
&= \frac{\mathbb{P}_{\theta^{(c)},\phi^{(c)}}\left(Z_i = 1, Y_i = 0 \,|\, x_i, u_i\right)}{\mathbb{P}_{\theta^{(c)},\phi^{(c)}}\left(Y_i = 0 \,|\, x_i, u_i\right)} \\
&= \frac{\eta_{\theta^{(c)}}(x_i)\,\left(1 - e_{\phi^{(c)}}(u_i)\right)}{1 - \eta_{\theta^{(c)}}(x_i)\, e_{\phi^{(c)}}(u_i)} \ .
\end{aligned}$$

The conditional expectation of the log-likelihood is then:

$$\begin{aligned}
Q^{(c)}\left(\theta, \phi\right) = &\sum_{i=1}^{n} \left[\gamma_i^{(c)} \log\left(\eta_\theta(x_i)\right) + (1 - \gamma_i^{(c)}) \log\left(1 - \eta_\theta(x_i)\right)\right] \\
&+ \sum_{i=1}^{n} \gamma_i^{(c)} \left[y_i \log\left(e_\phi(u_i)\right) + (1 - y_i) \log\left(1 - e_\phi(u_i)\right)\right] \ .
\end{aligned}$$

Maximization step. The conditional expectation of the log-likelihood can be separated into two terms, one involving only θ, and the other involving only ϕ. Hence, the maximization step consists in solving two separate maximization problems:

$$\theta^{(c+1)} \in \underset{\theta}{\text{Argmax}} \sum_{i=1}^{n} \left[\gamma_i^{(c)} \log\left(\eta_\theta(x_i)\right) + (1 - \gamma_i^{(c)}) \log\left(1 - \eta_\theta(x_i)\right)\right] \tag{14}$$

$$\phi^{(c+1)} \in \underset{\phi}{\text{Argmax}} \sum_{i=1}^{n} \gamma_i^{(c)} \left[y_i \log\left(e_\phi(u_i)\right) + (1 - y_i) \log\left(1 - e_\phi(u_i)\right)\right] \tag{15}$$

The first maximization problem (Eq. 14) is a weighted logistic regression where each observation is considered as positive with weight $\gamma_i^{(c)}$ and negative with weight $1 - \gamma_i^{(c)}$. The second in Eq. 15 is also a weighted logistic regression based on the observed labels $(y_i)_{1 \leq i \leq n}$.

We implemented the methodology on synthetic data and the results confirm and illustrate its effectiveness. We refer the reader to the Chapter 4 of [10] for a detailed presentation of these results. Similar experiments were conducted in [12] to study empirically the generalization performances of PU learning given the propensity. These experiments quantify the difficulty of PU learning when the propensity decreases.

5. Application and results

We now apply the PU learning methodology to the fatigue data set $(x_i, u_i, y_i)_{1 \leq i \leq n}$. Subsection 5.1 details the experimental setup and the performance metrics used to evaluate the estimated models. Then, the results are presented and analyzed in Subsection 5.2.

5.1. Estimation and performance evaluation

The fatigue data set is split in two sub-samples with equal sizes: one will be used for training (estimation), the other one for testing (evaluation). In the training phase, the parameters of the PU learning model are estimated on the training data. The PU learning model used is the one introduced in Subsection 4.3 with a logistic propensity. This PU learning model is compared with two other methods:

1. a baseline model corresponding to the standard danger coefficient from the Dang Van criterion;
2. a standard linear logistic regression, which ignores the selection bias and therefore assumes that the observed labels $(y_i)_{1\leq i\leq n}$ are identical to the true classes $(z_i)_{1\leq i\leq n}$. The parameters of this model are estimated on the training set.

The choice of linear logistic regression as the benchmark classification model is motivated by two reasons. On the one hand, it is identical to the classification model of the PU learning parametric model; hence we can compare PU and non PU approaches on a similar basis. On the other hand, we demonstrated that even non-linear classification models such as Random Forests do not outperform significantly linear logistic regression (cf. [11]).

Once the estimations are done, we evaluate the performances of the models on the test set. As the test data is itself affected by the same PU label noise, it does not provide the ground truth about the classes. Hence, we cannot properly assess the performances of the estimated classifier $\widehat{\eta}$. We have two alternative ways to evaluate the PU learning model.

1. *Label predictions:* for a covariate vector (x, u), the probability of crack initiation $(Y = 1)$ is modeled as a product $\eta(x) \times e(u)$. Thus, using the estimated classifier $\widehat{\eta}$ and the propensity $\widehat{e}$, it is possible to estimate this probability $\widehat{p}(x, u)$:
$$\widehat{p}(x, u) = \widehat{\eta}(x) \times \widehat{e}(u) \ .$$
Therefore, comparing these posterior probabilities to the labels on the test set provides a first set of performance indicators.
2. *Class predictions:* Even if the true classes $(Z_i)_{1\leq i\leq n}$ are not available, we have access at least to a sub-sample of positive instances. Indeed, we already know that labeled instances ($Y = 1$, crack initiations) are critical ($Z = 1$). Besides, we have access to several test outputs for each zone (usually 3 to 7). In particular, a zone with at least one crack initiation among the tests performed is critical. Hence, we know that these instances share the same class ($Z = 1$), even those that did not result in crack initiation. It is worth noting that the corresponding individuals in the data set are not strictly identical as the test severities are different. Therefore, we have access to the knowledge of an extended subset of positive instances (critical zones). We denote $\tilde{Z}$ the variable indicating whether a crack was initiated at least once ($\tilde{Z} = 1$) or never ($\tilde{Z} = 0$). These "approximate classes" $\tilde{Z}_i$ can be compared to the classification predictions $\widehat{\eta}(x_i)$, allowing to assess the performances of the PU classifier. We insist though that these evaluations may be biased since $\tilde{Z}_i$ does not provide the ground truth about the classes.

The performances are computed on the three models: the PU learning model of interest, the standard classifier ignoring the PU label noise and the Dang Van fatigue criterion. The performances are evaluated in terms of area under Receiver Operating Characteristics curve (ROC AUC) and Precision-Recall curve (PR AUC). The advantage of these metrics is that they do not require the specification of a decision threshold on the predictions (*i.e.* the probability value above which a zone should be predicted as positive). Instead, they are global performance metrics that integrate the performances of a classifier over all the possible thresholds. When classes are separable, the ROC AUC and PR AUC of an optimal classifier are both equal to 1. The ROC AUC of a random classifier (no-skill classifier) is 0.5 and its PR AUC is equal to the proportion of positive instances.

In our case, performance assessment is particularly difficult because of the importance of variance in performance estimation. In order to evaluate our models with more consistency, we repeat $B = 100$ times the procedure described above. For each repetition, the train-test split is randomly chosen. This gives us access to the distribution of the performances and will allow us to compare the models.

Remark:. The fatigue data set is very imbalanced: only 2% of the observations are crack initiations. Even if the proportion of critical zones is higher (as some of them are unlabeled), there are still far more safe zones ($Z = 0$) than critical ones ($Z = 1$). This imbalance is accounted for in the performance metrics used to evaluate the predictive performances of our models. It is well known that the accuracy metric, which measures the proportion of well classified instances can be misleading when evaluating performances on imbalanced data sets: in our example, a naive classifier that predicts every zone as safe ($Z = 0$) would get an accuracy close to perfection even if the model is meaningless. Instead, the ROC and PR AUC used here are not sensitive to imbalance. Besides, these performance metrics only depend on the ranking of the predicted probabilities. In other words, they measure the ability of the criterion to correctly rank the zones from the less critical to the more critical.

5.2. Results

We carry out the experiment described in the previous paragraph and compare the prediction performances of PU learning, standard classification and Dang Van criterion. We recall that the Dang Van criterion only relies on two variables whereas both PU learning and standard classification use the five variables presented in Subsection 3.3. The results are presented in Figure 11.

We notice that both statistical methods get higher performances compared to the Dang Van criterion. For label prediction, PU learning outperforms standard logistic regression in terms of PR AUC score (Fig. 11, top right). This is less obvious for the ROC AUC (Fig. 11, top left).

In order to compare more efficiently PU learning to the two other methods, let us look at the difference of performances achieved experiment by experiment. For j between 1 and 100, denote s_j^{PU}, s_j^{LR} and s_j^{DV} the scores of PU learning, standard logistic regression and Dang Van criterion for the j^{th} experiment. We perform several paired difference t-tests to compare the mean performances. More particularly, for each evaluation

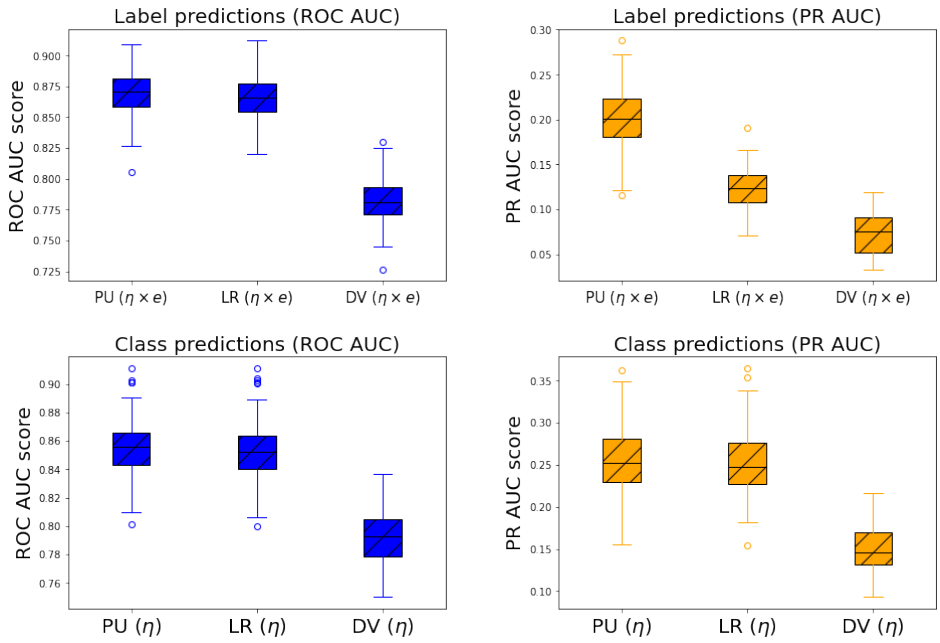

Figure 11: Quantitative performances (ROC AUC) of models with five variables: prediction performances on the labels, *i.e.* using the first evaluation method (top, $\eta \times e$); and on the classes, *i.e.* using the second evaluation method (bottom, η). The metrics used are ROC AUC (left) and PR AUC (right). Each boxplot represents the distribution of the prediction performances. Each series of three boxplots corresponds to (from left to right): PU-LR, the standard logistic regression and the mechanical Dang Van criterion.

		$H_0 : s_m^{PU} \leq s_m^{DV}$		$H_0 : s_m^{PU} \leq s_m^{LR}$	
		Statistics	p-value	Statistics	p-value
1. Label predictions (η)	ROC AUC	48.36	5.46×10^{-71}	2.95	1.98×10^{-03}
	PR AUC	46.97	8.73×10^{-70}	34.96	8.11×10^{-58}
2. Class predictions ($\eta \times e$)	ROC AUC	43.55	1.10×10^{-66}	6.20	6.54×10^{-09}
	PR AUC	38.65	7.82×10^{-62}	3.35	5.76×10^{-04}

Table 2: Comparison of mean performances (PU vs DV and PU vs LR) using paired difference t-tests.

methods (label predictions, class predictions) and performance metrics (ROC and PR AUC), we test the null hypothesis H_0: $s_m^{PU} \leq s_m^{DV}$ against the alternative H_1: $s_m^{PU} > s_m^{DV}$, where s_m^{PU} (s_m^{DV}) is the mean score for PU learning (mean score for the Dang Van criterion). The same method is used to test the superiority of PU learning s_m^{PU} over the standard logistic regression mean score s_m^{LR}. The use of paired difference t-tests is legitimate as for each j, s_j^{PU}, s_j^{DV} and s_j^{NT} are calculated over the same test set for models estimated over the same training sets. The results are presented in Table 5.2: for each test performed, we reject the null hypothesis at 5% level and the p-values confirm the statistical significance of the superiority of PU learning over the other approaches.

We represent in Figure 12 the empirical distribution of the difference of performances between PU learning

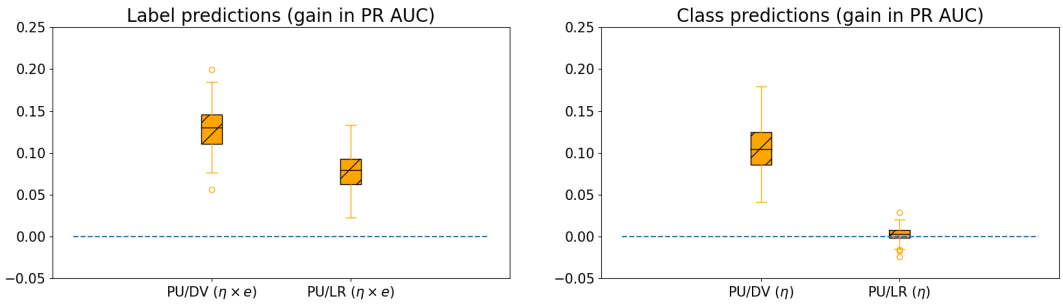

Figure 12: Performance comparisons (PR AUC): label predictions on the left (first evaluation method), class predictions on the right (second evaluation method). For each figure, the boxplot on the left represents the difference of performances between PU learning and the Dang Van criterion, the second compares PU learning and the standard logistic regression.

and Dang Van criterion $(s_j^{PU} - s_j^{DV})_{1\leq j\leq 100}$, and between PU learning and standard logistic regression $(s_j^{PU} - s_j^{NT})_{1\leq j\leq 100}$. Looking at the performances on the label and class predictions, PU learning clearly outperforms the Dang Van criterion for every experiment with a significant median gain of approximately 0.14 over PR AUC. For label predictions, PU learning achieves a median gain of about 0.08 over standard classification. This visually confirms the results of Table 5.2. However, when comparing PU learning to standard classification for class predictions, it appears that the average gain in PR AUC is small. This can seem surprising: indeed, as the goal of PU learning is to better estimate the classifier by accounting for PU label noise, we may expect better performances on class predictions. However, as explained earlier, the performances of the PU classifier may remain biased as we do not know all the critical zones in the test set. Hence, the results only tell that PU learning classifier is as good as its standard counterpart in characterizing already labeled critical zones. Moreover, another added value of the PU learning approach is that it provides an estimate of the propensity function at the same time as the design criterion, whereas the standard classification does not.

5.3. Illustration

Once the classifiers are estimated, they can be easily deployed in Finite Element software to provide predictions on different zones of a conception. As for the danger coefficient of the Dang Van criterion, it is straightforward to evaluate the probability for each zone to be critical.

The deployment of the methodology is represented in Figure 13. On this test example of a cradle model under longitudinal loading, the predictions of the Dang Van criterion and those of PU learning are compared. The danger coefficient of a zone is defined as the maximum danger coefficient among the elements of the zone. First, it is important to note that PU learning criterion highlights only a few critical zones, while the majority remains blue (safe). Conversely, the Dang Van danger coefficient is high (values close to the threshold 0 or above) for a large number of zones. Hence, following the predictions of Dang Van criterion would require many reinforcements on the part: most of these reinforcements may be unnecessary. Instead, PU learning has less false alarms and is more likely to guide the design teams into focusing on the most

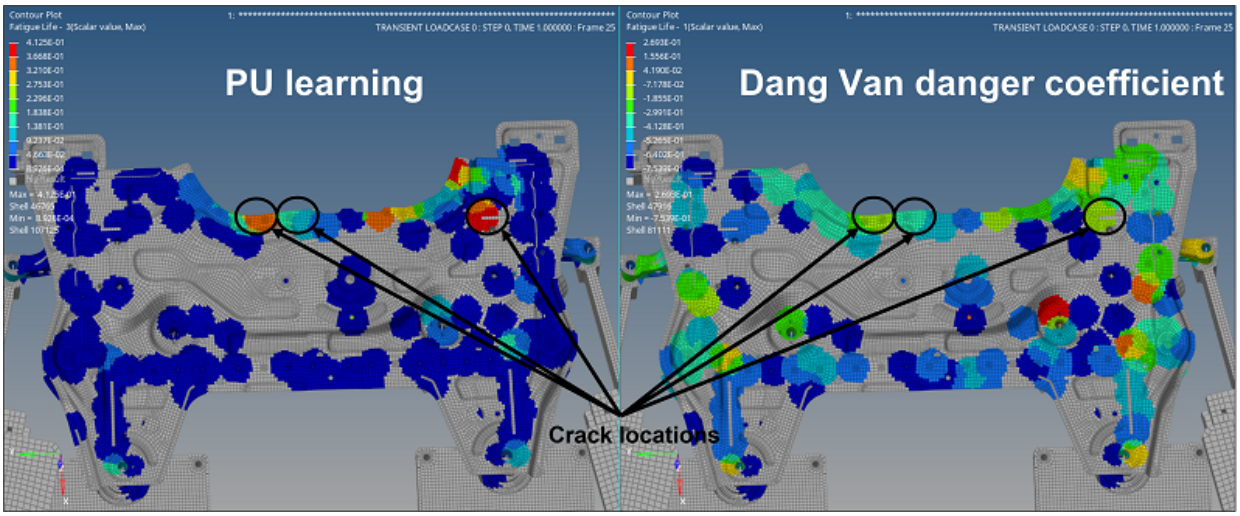

Figure 13: PU learning predictions (left) and Dang Van criterion (right) on a FEM of cradle model under longitudinal loading (in the test set, not used for estimating the model): dangerous zones in red, safe zones in blue. The gray zones on the FEM were not selected during the pre-processing (very low stress): they should be considered as blue. The crack initiation locations on this side of the part are highlighted as dark circles.

critical locations (those that are more likely to fail during testing). Moreover, the known crack initiation locations are better identified by PU learning criterion than through the Dang Van danger coefficient. Overall, these observations confirm the substantial gain in performances achieved by our statistical fatigue criteria compared to the classical Dang Van criterion.

6. Conclusion and perspectives

In this article, we presented a statistical point of view on fatigue criteria. A probabilistic Dang Van criterion was proposed, which allows the joint estimation of the material and dispersion parameters. This criterion was estimated on an auxiliary data set gathering numerical results and fatigue test data on welded elementary structures (coupon tests). However, this probabilistic criterion does not generalize well to the identification of critical zones on real-scale and complex automotive parts. Therefore, we relied directly on a fatigue database of complex automotive components (cradle and cross-member designs) to construct a fatigue criterion using statistical methods. We introduced a concept of zone to facilitate the analysis of the database and developed a novel approach, based on PU learning, to build a fatigue criterion. This criterion combines several advantages. First, it is estimated on a data set that is representative of the complexity of real-scale mechanical components. Second, it takes into account various features in addition to those traditionally considered in Dang Van criterion. Finally, the PU learning approach allows to model the effects of the testing conditions on the test output (observed labels). The results confirm the interest of the method and highlight a substantial gain in predictive performances compared to the standard Dang Van criterion.

This work paved the way for the development and deployment of statistical fatigue criteria to improve the numerical design of safety parts of vehicles. Several research directions remain open and could lead to

significant improvements in the methodology and its efficiency. In fact, despite the improved performances, these statistical criteria are still far from being perfectly predictive. Indeed, their accuracy is limited by the uncertainties of the underlying numerical model. In particular, several important phenomena such as manufacturing processes and residual stresses are still not well accounted for in FEM computations. As the improvement of FEM is an active topic in the automotive industry, it is expected that FEM will become more accurate in the near future. As a consequence, the fatigue database could benefit from higher quality data sets, which would also improve the characterization of fatigue phenomena using the methodologies developed in this article. As a future perspective, it would also be interesting to increase the size of the database. The potential of available data is huge and augmenting the diversity of mechanical parts and loading conditions could contribute to the improvement of the statistical criterion. In addition, increasing the number of descriptive features may improve the efficiency of the fatigue criteria: for example, better consideration of geometry may help in the identification of critical zones. Even if manufacturing process effects and residual stresses cannot be calculated through FEM, two zones with similar geometry and materials will have similar fatigue resistance. Hence, adapted deep learning architectures may be helpful to account for the entire geometry of the zone and extract high-level features.

Acknowledgements

This work was carried out within the framework of the partnership between Stellantis and the OpenLab AI with the financial support of the ANRT for the CIFRE contract n°2019/1131.

References

[1] P Ballard, K Dang Van, A Deperrois, and YV Papadopoulos. High cycle fatigue and a finite element analysis. *Fatigue & Fracture of Engineering Materials & Structures*, 18(3):397–411, 1995.

[2] OH Basquin. The exponential law of endurance tests. In *Proc Am Soc Test Mater*, volume 10, pages 625–630, 1910.

[3] Claude Bathias and André Pineau. *Fatigue of materials and structures*. Wiley Online Library, 2010.

[4] Pauline Beaumont. *Optimisation des plans d'essais accélérés Application à la tenue en fatigue de pièces métalliques de liaison au sol*. PhD thesis, Université d'Angers, 2013.

[5] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data under the selected at random assumption. In *Proceedings of The Learning with Imbalanced domains: Theory and Application Workshop @ ECML 2018*, volume 94, pages 8–22. Journal of Machine Learning Research, 2018.

[6] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719–760, April 2020.

[7] S Bergamo, P Schimmerling, F Triboulet, P Wilson, M L Facchinetti, M. Monin, Lefebvre F., and B Weber. Préconisations pour les caractéristiques statistiques de résistance en fatigue. *SIA*, 2017.

[8] Enrique Castillo and Alfonso Fernández-Canteli. *A unified statistical methodology for modeling fatigue damage*. Springer Science & Business Media, 2009.

[9] José Correia, Nicole Apetre, Attilio Arcari, Abílio De Jesus, Miguel Muñiz-Calvente, Rui Calçada, Filippo Berto, and Alfonso Fernández-Canteli. Generalized probabilistic model allowing for various fatigue damage variables. *International Journal of Fatigue*, 100:187–194, 2017.

[10] Olivier Coudray. *A statistical point of view on fatigue criteria : from supervised classification to positive-unlabeled learning*. Theses, Université Paris-Saclay, December 2022.

[11] Olivier Coudray, Philippe Bristiel, Miguel Dinis, Christine Keribin, and Patrick Pamphile. Fatigue data-based design: statistical methods for the identification of critical zones. In *SIA Simulation Numérique*, 2021.

[12] Olivier Coudray, Christine Keribin, and Patrick Pamphile. Convergence rates for positive-unlabeled learning under selected at random assumption: sensitivity analysis with respect to propensity. In *Conférence sur l'Apprentissage automatique*, 2022.

[13] K Dang Van and B Griveau. On a new multiaxial fatigue limit criterion- theory and application. *Biaxial and multiaxial fatigue(A 90-16739 05-39). London, Mechanical Engineering Publications, Ltd., 1989,*, pages 479–496, 1989.

[14] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[15] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 213, Las Vegas, Nevada, USA, 2008. ACM Press.

[16] Jean-Luc Fayard. *Dimensionnement à la fatigue polycyclique de structures soudées*. PhD thesis, Ecole Polytechnique, 1996.

[17] Rémy Fouchereau, Gilles Celeux, and Patrick Pamphile. Probabilistic modeling of s–n curves. *International Journal of Fatigue*, 68:217–223, 2014.

[18] L Locati. Le prove di fatica come ausilio alla progettazione ed alla produzione. *La Metallurgia Italiana*, 9:301, 1955.

[19] Milton A Miner. Cumulative damage in fatigue. *Journal of Applied Mechanics*, 1945.

[20] Arvid Palmgren. Die lebensdauer von kugellargern. *Zeitshrift des Vereines Duetsher Ingenieure*, 68(4):339, 1924.

[21] Clément Roux, Xavier Lorang, Habibou Maitournam, and ML Nguyen-Tajan. Fatigue design of railway wheels: a probabilistic approach. *Fatigue & Fracture of Engineering Materials & Structures*, 37(10):1136–1145, 2014.

[22] J Schijve. Statistical distribution functions and fatigue of structures. *international Journal of Fatigue*, 27(9):1031–1039, 2005.

[23] Jaap Schijve. *Fatigue of structures and materials*. Springer, 2009.

[24] R Ben Sghaier, Ch Bouraoui, R Fathallah, T Hassine, and A Dogui. Probabilistic high cycle fatigue behaviour prediction based on global approach criteria. *International journal of fatigue*, 29(2):209–221, 2007.

[25] Luca Susmel and P Lazzarin. A bi-parametric wöhler curve for high cycle multiaxial fatigue assessment. *Fatigue & Fracture of Engineering Materials & Structures*, 25(1):63–78, 2002.

[26] Jean-jacques Thomas, T.M.L. Nguyen-Tajan, and P Burry. Structural durability in automotive design. *Mat.-wiss. u. Werkstofftech*, 36(11), 2005.

[27] Bastien Weber. *Fatigue multiaxiale des structures industrielles sous chargement quelconque*. PhD thesis, Lyon, INSA, 1999.