Shrinkage porosity prediction empowered by physics-based and data-driven hybrid models

Madyen Nouri, Julien Artozoul, Aude Caillaud, Amine Ammar, Francisco Chinesta, Ole Köser

Shrinkage porosity prediction empowered by physics-based and data-driven hybrid models

Madyen Nouri[1], Julien Artozoul[1], Aude Caillaud[1], Amine Ammar[1], Francisco Chinesta[2] and Ole Köser[3]

[1] LAMPA Laboratory, Arts et Métiers Institute of Technology
Boulevard du Ronceray 2, CEDEX 01, BP 93525
F-49035 Angers, France
e-mail: madyen.nouri@ensam.eu
julien.artozoul@ensam.eu
aude.caillaud@ensam.eu
amine.ammar@ensam.eu

[2] PIMM Laboratory ESI Group Chair, Arts et Métiers Institute of Technology
CNRS, Cnam, HESAM Université
151 boulevard de l'Hôpital, 75013 Paris, France
email: francisco.chinesta@ensam.eu

[3] ESI Group
Gal. Benjamin Constant 1 c/o Fidurba SA, 1003 Lausanne, Switzerland
e-mail: ole.koeser@esi-group.com

ABSTRACT

Several defects might affect a casting part and degrade its quality and the process efficiency. Porosity formation is one of the major defects that can appear in the resulting product. Thus, several research studies aimed at investigating methods that minimize this anomaly. In the present work, a porosity prediction procedure is proposed to assist users at optimizing porosity distribution according to their application. This method is based on a supervised learning approach to predict shrinkage porosity from thermal history. Learning data are generated by a casting simulation software operating for different process parameters. Machine learning was coupled with a modal representation to interpolate thermal history time series for new parameters combinations. By comparing the predicted values of local porosity to the simulated results, it was demonstrated that the proposed model is efficient and can open perspectives in the casting process optimization.

Keywords: Smart manufacturing, Physics-based modelling, Model Order Reduction, PGD, Data-driven modelling, Artificial intelligence, Hybrid Twins, Diagnosis and prognosis, Shrinkage porosity, casting

1 Introduction

In industrial environments, there are several techniques of material forming. Casting is a widely used material forming process to produce a required shape with less raw material consumption. It can be operated to form various metals such as aluminum alloys. The latter is one of the most common metals used in industrial environment thanks to its thermo-physical properties. However, various defects might appear in the resulting part that can contribute on crack initialization and fatigue resistance decreasing [1]. Porosity is one of the major defects that can affect a casting part. This anomaly may appear by two main reasons. During solidification, a negative volume variation occur in general and has to be compensated by liquid flow to avoid porosity. This feeding induces a pressure drop which gives a rise to shrinkage porosity in view of the decreasing temperature [2]. In addition, gas rejection can appear in the interphase region due to the low limit of solubility of dissolved gases and its high concentration in the liquid which leads to a degraded mechanical properties and corrosion resistance [3].This fact highlights the importance of studying the control of porosity according to the application such as obtaining a fine surface condition by avoiding pipe shrinkage or enhance the casting part resistance with circumventing microporosity inside the part. Hence, predicting porosity in design stage is the key to overpass this problem.

Though numerical simulation and experimental approaches have been playing a prominent role in casting process design. However, it would take rounds of simulation with a particular set of parameters to come with an optimized process parameters configuration. Giving a selection of significant parameters that contributes directly on porosity formation, several researchers have been focused on optimizing the overall rate of porosity in a casting part using machine learning techniques. That might minimize the time cost to come with an optimized configuration of the casting process. For this purpose, Tsoukalas [4] and Hsu et al. [5] developped a solution based on multivariable linear regression (MVLR). The latter has often been used in manufacturing processes to fit a linear regression with forcasting a sequence of parameters. In the considered solution, MVLR serves to predict the porosity rate from a weighted sum of the selected parameters. The idea that remains behind is to create an optimization loop that minimizes the overall volume of porosity using Genetic algorithm (GA).The latter is used to generate a new set of values of the selected parameters. Another solution proposed in the works of Anijdan et al. [6] and Gong et al. [7] adopts the same principle using Artificial Neural Network (ANN) instead of MVLR.

These solutions can be considered as efficient in minimizing the overall volume of porosity. This effectiveness has been exemplified in the work of Tsoukalas [4] where the total porosity volume is reduced with a rate of 66% in aluminum alloy pressure die casting. However, it does not predict the porosity distribution in the part. In the present work, a methodology of local porosity prediction is proposed using machine learning techniques.

The remainder of this paper is organized as follow: Section 2 revisits the adopted methodology to develop the proposed solution. Section 3 presents the obtained results and evaluates the effectiveness of the proposed methodology. Finally, section 4 addresses some final conclusions.

2 The proposed methodology

2.1 Basic architecture

After presenting the technical context of this study and discussing the existing works, an overview of our proposed methodology is presented in the flowchart (figure 1) and detailed below.

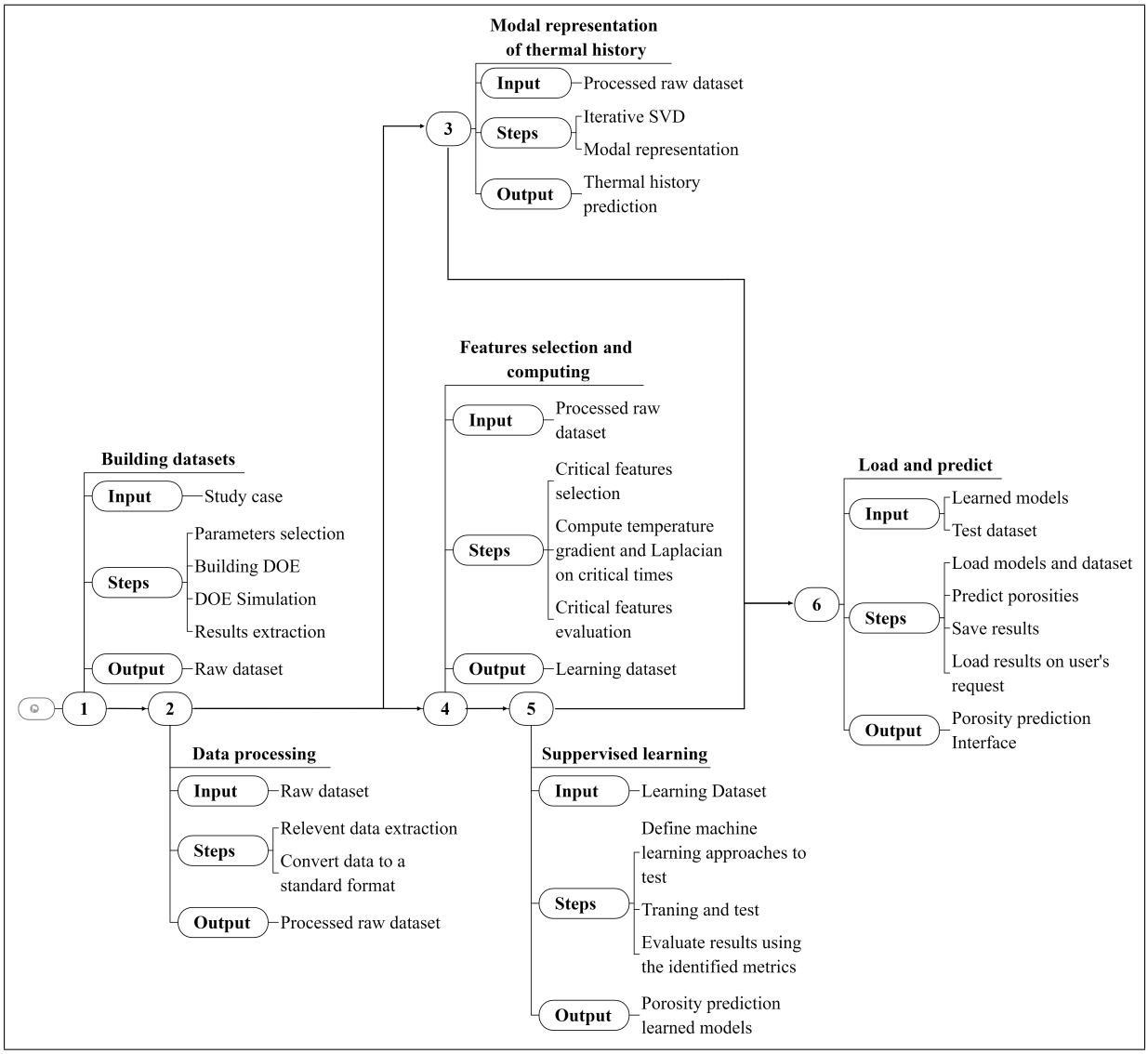

Figure 1: Flowchart of the developed porosity prediction solution

After defining a casting study case, the proposed methodology begins by building a raw dataset (step 1 on figure 1) that serves to implement learning techniques. It is composed mainly of nodal thermal histories and the corresponding porosity values. A preparation step consists of analysing the considered case study and identifying the parameters to be varied in the Design Of Experiment (DOE). Then, a dataset is created from the results of the realized simulations on the casting software (Procast by ESI Group). The latter is post-processed in order to standardize its format (step 2 on figure 1).

Exhaustively, the prediction process concerns two main steps:

- Giving the raw dataset (represented with red dots in figure 2), a modal representation

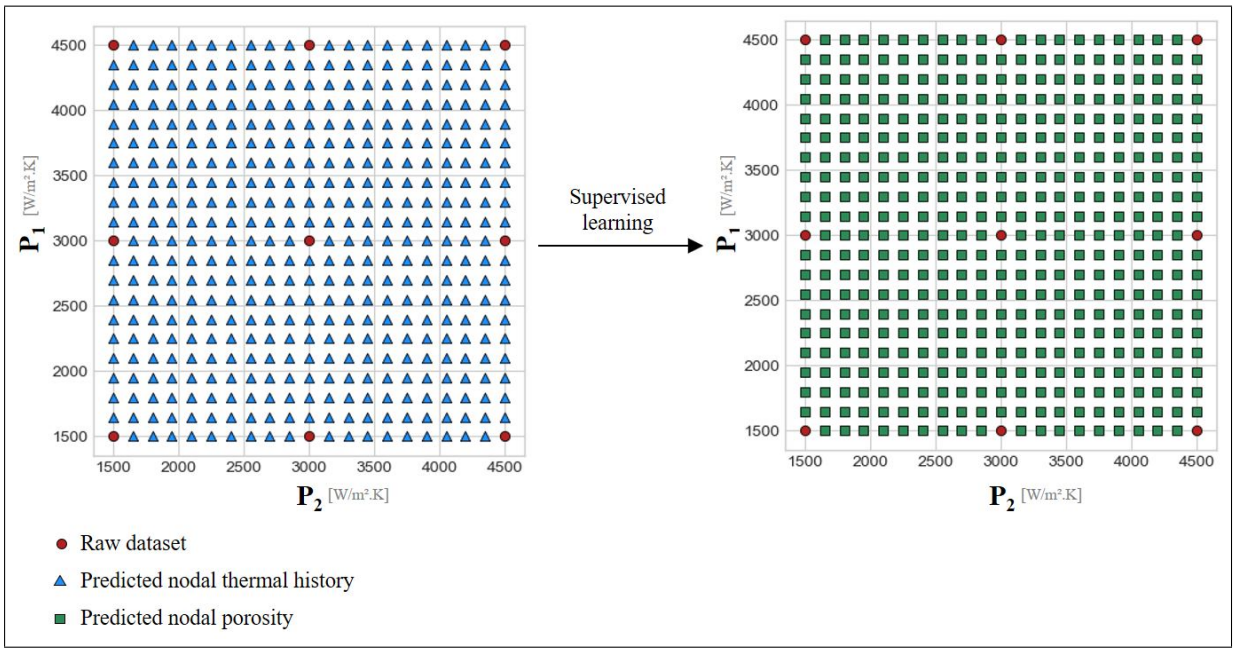

Figure 3: Nodal porosity prediction

combined with model reduction (Singular Value Decomposition) are implemented in order to interpolate the nodal thermal history in the parameter space (blue triangles in figure 2) (step 3 on figure 1).

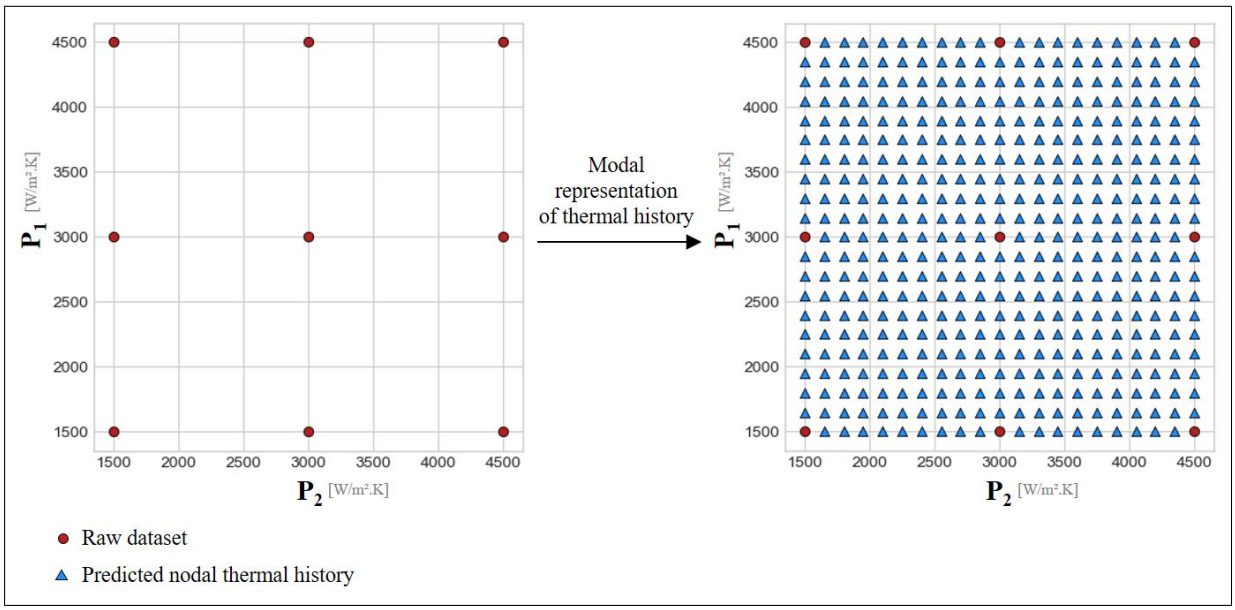

Figure 2: Thermal history prediction

- The raw dataset is enriched by computing some additional features such as gradient and Laplacian of temperature (step 4 on figure 1). Local porosity is predicted from an optimized selection of critical features via supervised learning (step 5 on figure 1). Finally, the learned model can be used to predict local porosity from the interpolated thermal history in the parameter space as presented in figure 3 with green squares (step 6 on figure 1).

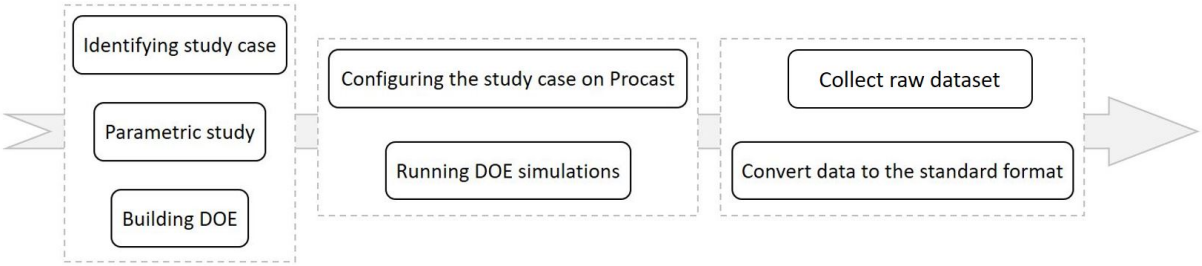

Figure 4: Building dataset process

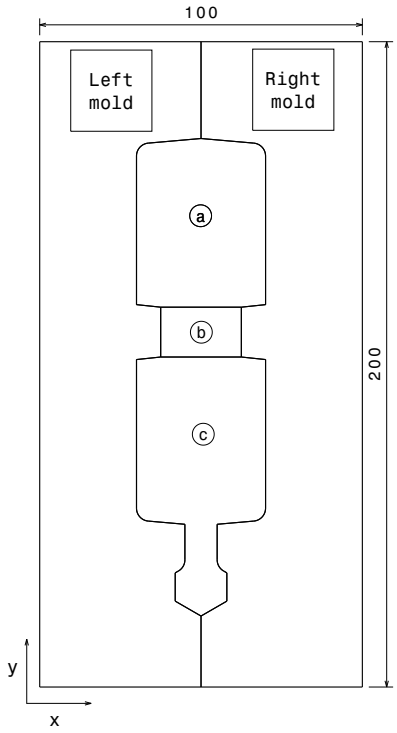

Figure 5: Test case: ”Hourglass”
a: upper area; b: central area (necking); c: lower area

2.2 Building raw dataset

Preparing datasets for training and testing is essential since machine learning based methods are used in this work. An overview of this process is described in figure 4. The resulting dataset is composed of a mesh description of the considered part, time series presenting temperature maps at each instant until reaching 10°C under the solidus temperature of the considered material, nodal porosity and pipe shrinkage (voids).

2.2.1 Test case

Our test case, presented in figure 5, consists of a prismatic part called ”Hourglass” made up of two areas separated by a necking. The width of the latter is set at $25mm$. The height and the width of the part measures respectively $150mm$ and $40mm$. In addition, the defined mesh has an elementary thickness $(2mm)$ to facilitate the conversion to a bi-dimensional problem.

In this study, several parameters are kept unchanged in the DOE such as the casting alloy

AlSi7Mg0.3. The mechanical and thermal characteristics of this alloy are reported in table 1.

Alloy designation	UNI 3599; A356; AlSi7Mg0.3; G Al Si 7
Density [kg/m^3] (at 20°C)	2675
Solidification interval [°C]	613-548
Thermal conductivity [W/mK]	138.171
Specific heat [J/kg.K] (at 20°C)	919
Latent heat [kJ/kg]	431

Table 1: Mechanical and thermal properties of the alloy AlSi7Mg0.30.3

2.2.2 Parametric study

For each study, a parametric analysis should be carried out in order to determine the most significant parameters which influence the resulting temperature map. Since the considered dataset is expected to predict porosity, their significance can be measured with their influence on the resulting temperature and porosity distributions in order to be able to browse a maximum of different learning samples.

For the identified study case, the chosen parameters are:

- P_1: Heat transfer coefficient (HTC) between the area "a" of the casting part and the mold.
- P_2: Heat transfer coefficient between the area "c" of the casting part and the mold.

Bounds should be determined for each parameter in order to help create an optimized DOE. After testing and evaluating these parameters, the bounds are set to [1500, 4500] $W/m^2.K$. The two parameters are varied with a step of 1500 in order to obtain a 3x3 uniform grid that represents the parametric space. Each parameters combination presents an experiment in the DOE.

2.2.3 Numerical simulation

The considered simulation software is Procast by ESI group. The main inputs for the casting simulation process are:

- Thermo-physical properties of the alloy: density, specific heat, thermal conductivity, fraction of solid and viscosity of the alloy.
- Boundary conditions: metallic mold (40CrMnMo8-6) heat transfer coefficients with the alloy and the surrounding environment.
- Process parameters: the mold is considered as initially fully filled. The initial temperatures of the mold and the alloy are set respectively to 100°C and 700°C.

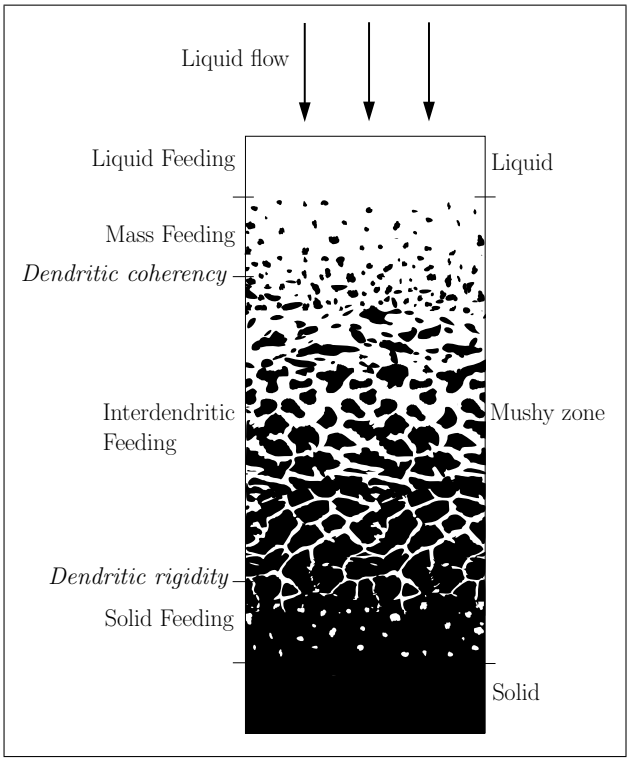

Figure 6: Different feeding mechanisms during casting process

Boundary conditions are varied by defining the HTC of the mold with the areas "a" and "c" according to the DOE. Hence HTC of the mold with the area "b" is kept unchanged at $3000W/m.K$. On the external surface of the mold, an air cooling condition is defined with an ambient temperature of 20°C and a HTC of $10W/m^2.K$. The front and back faces illustrated on figure 5 are defined as adiabatic so that the part can acquire symmetry properties to compensate for the low thickness of the part.

During the simulation, the casting part is segmented into liquid and solid regions separated with an interface called mushy zone. The latter is divided four domains where each feeding mechanism may be active. A schematic representation of the different feeding mechanisms is shown in Figure 6, where four types of feeding are illustrated:

- Liquid feeding which appears at the initial stage of solidification when the solid fraction value is null.
- Mass feeding occurs when solid particles flow with the liquid until reaching Dendritic Coherency (DC).
- Interdendritic feeding refers to the liquid filtering through a coherent network of solid dendritic arms until getting to Dendritic Rigidity (DR).
- Solid feeding is related to distortion of castings and may occur after exceeding the DR phase.

Procast standard model consists of identifying the nodes located in the mushy zone and determining the active feeding mechanism in order to compute the porosity value following its conditional algorithm. The valuable outputs of this simulation model are time dependent thermal fields, shrinkage porosity and solid fraction.

2.3 Thermal history prediction

After computing all the thermo-mechanical fields related to the considered DoE, porosity prediction begins with the interpolation of spacial thermal history. This approach is based on a modal representation of time-space thermal field using Singular Value Decomposition (SVD).

For each parameters combination, temperature field is a time series that is initially defined as a matrix containing time steps (columns) and the space evolution (rows). With the aim of preparing the input data for the SVD, each matrix is converted to a vector that aggregates time and space evolutions. These vectors are assembled in one global matrix $\mathbb{T}$ that contains the temperature fields evolution of all the considered parameters combination in each column. The nodal porosity values that represents the output are superposed likewise on a vector $\mathbf{P}$.

After decomposing the global matrix using SVD, an interpolation is performed in order to predict the thermal history of new parameters combination.

2.3.1 Singular Value Decomposition (SVD)

Using SVD one can approximate a given matrix $\mathbb{T} \in \mathbb{R}^{l \times K}$ that refers to the temperature time series where l rows correspond to the time-space evolution of nodal temperatures and K columns present the considered parameters combinations. The reduced model can be expressed as follows:

$$\mathbb{T} = \mathbb{F}\Sigma\mathbb{G}^T \tag{1}$$

where $\mathbb{F}$ and $\mathbb{G}$ are orthogonal matrices that each column corresponds to left and right singular vectors, respectively. Σ is a diagonal matrix whose diagonal elements are non-negative real singular values. The purpose of a SVD algorithm is to estimate orthogonal matrices $\mathbb{F} \in \mathbb{R}^{l \times m}$, $\mathbb{G} \in \mathbb{R}^{K \times m}$ and a diagonal matrix $\Sigma \in \mathbb{R}_{+}^{m \times m}$. This gives a low-rank approximation of the given matrix $\mathbb{T}$ with the desired rank m [8].

Regarding to the size of the manipulated data, an iterative strategy, presented in appendix A, is used in this work to perform SVD.

For all parameters combination (P_1, P_2) of the raw dataset (represented with the red dots in figure 2), the thermal history can be decomposed in the following form thanks to SVD:

$$\mathbb{T} = \mathbb{F} \otimes \mathbb{G} \tag{2}$$

where $\mathbb{F}$ and $\mathbb{G}$ represent the resulting decomposition functions from SVD and contain the two square roots of Σ.
The components of the thermal history can be expressed by:

$$T_{hj} = \sum_{j=1}^{j=m} F_{hj} G_{ij} \tag{3}$$

where F is the matrix containing m columns of different modes in the time-space description. Each mode has a size equal to l. The different components of this matrix are denoted F_{hj} $(h = 1..l, j = 1..m)$. G_{ij} represents the ith component, that refers to the parameters combination $(P_1^i, P_2^i \ ; \ i = 1..K)$, of the jth column in the matrix $\mathbb{G}$. G_{ij} can also be denoted as $G_j^{P_1^i, P_2^i}$.

In order to establish a continuous representation, the resulting vectors of the SVD decomposition are associated to a piece-wise linear functions. Thus a continuous form of T can be written as:

$$T(P_1,P_2,x,t) = \sum_{j=1}^{j=m} f_j(x,t) \times g_j(P_1,P_2) \tag{4}$$

2.3.2 Modal representation of thermal history

The purpose of our approach is to predict thermal history for new combinations of parameters (P_1',P_2') represented with the blue triangles in figure 2. A new set of values is computed as an interpolation on the resulting functions from SVD:

$$G_j^{P_1',P_2'} = \sum_{i=1}^{i=K} \alpha_i G_j^{P_1^i,P_2^i} \tag{5}$$

where α_i is a coefficient that depends on the interpolation method on the parametric space (piece-wise linear, Chebyshev, ...).

Then, the hth component of the temperature for the new parameters combination (P_1',P_2') is:

$$T_h^{P_1',P_2'} = \sum_{j=1}^{j=m} F_{hj} G_j^{P_1',P_2'} \tag{6}$$

Using the continuous form in the $x-t$ space and the usual time-space piece-wise linear functions, the expression of the temperature reads:

$$T^{P_1',P_2'}(x,t) = \sum_{j=1}^{j=m} f_j(x,t) G_j^{P_1',P_2'} \tag{7}$$

2.4 Critical features selection and computing

At this stage, the temperature field of a given parameters combination can be provided regarding to the modal representation realized on the previous step.

In this section, a parametric study is detailed with an explanation of mathematical operators calculus.

2.4.1 Critical parameters selection

Porosity appearance is generally related to the failure of these feeding mechanisms. During the negative volume variation that occurs in the solidification stage, the existing porosity are meant to be compensated mainly with the interdendritic feeding. However, porosity may appear due to the pressure drop that occurs during this phenomena. This fact highlights the importance of the 4 mentioned solid fraction limits of this feeding mechanisms. The experimental evidence reported by Sigworth et al. [9] suggests that interdendtritic feeding begins at the solid fraction of DC (denoted f_{DC}) that is equal to about 0.2 - 0.3 for large-grained aluminum alloys. Liquid flows in the channels between the dendritic arms until reaching the solid fraction f_{DR} that can

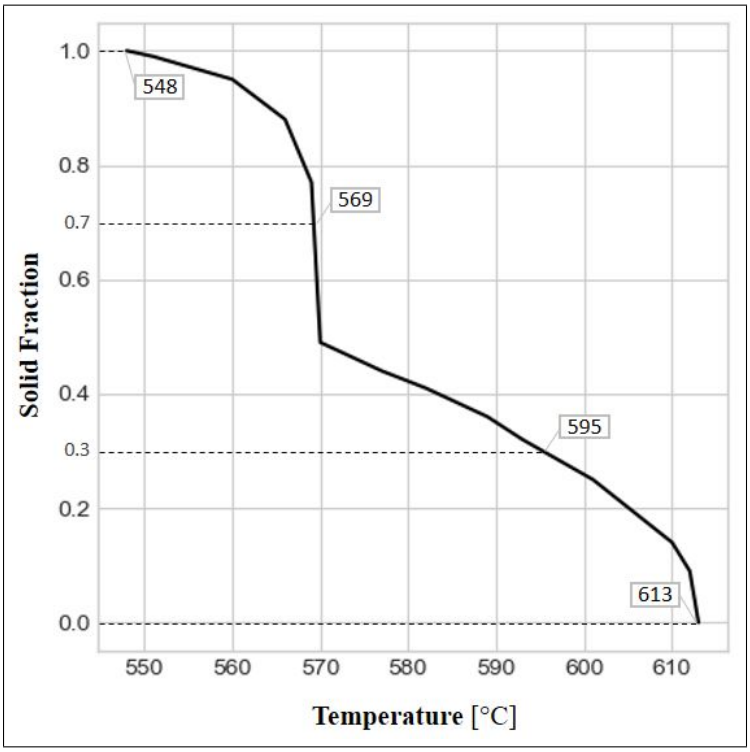

Figure 7: Solid fraction in terms of temperature for the alloy AlSi7Mg0.3

be approximated to 0,55 - 0,7. In our study f_{DC} is fixed on 0,3 and f_{DR} on 0,7. Depending on the used alloy, these solid fractions correspond to different values of temperature. The case of alloy AlSi7Mg0.3 used in this study is illustrated in figure 7.

The respective values of solid fraction null and 1 correspond to the temperature of liquidus (T_L) and the temperature of solidus (T_S). The temperature values of DR and DC are respectively denotes by T_{DR} and T_{DC}. For the rest of this study, these 4 values of temperature are called Critical temperatures.

As solidification progresses, the melt ahead of the solidifying front grows towards the cold zone in the neighborhood. This underlines the importance of the gradient of temperature as one of the major controlling parameters of the direction of solidification and the definition of the mushy zone where porosity are produced. Thus gradient and Laplacian are computed at the instants of reaching the critical temperatures and integrated as additional features on the dataset. Hence the selection of critical feature can be as following:

- Critical times: instants of attending the critical temperatures in each nodes (as depicted in figure 8 the example of a random node).
- Gradient of temperature in each node at critical times along X and Y.
- Laplacian of temperature in each node at critical times.
- For each node, the difference between critical temperatures and maximum temperature at critical time.

Other outputs of the casting simulations could be considered as additional features to enrich the input data such as pouring temperature, hydrostatic pressure, etc.

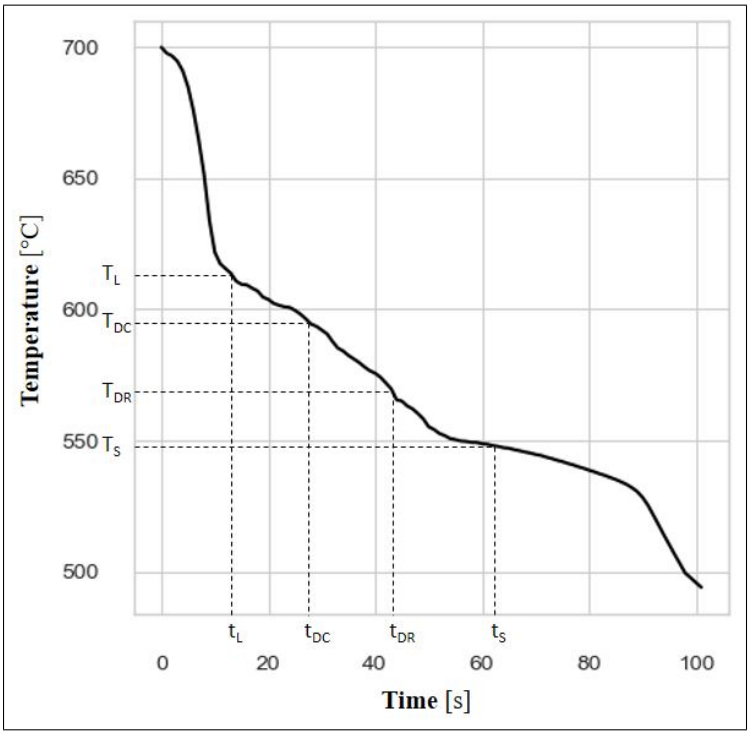

Figure 8: An example of nodal temperature as a function of time

2.4.2 Spacial gradient and Laplacian of thermal history

The geometric elements that allows the transformation from the reference finite element to any real element are defined as follows using a usual finite element interpolation:

$$x(\xi,\eta) = \sum_{i=1}^{n_e} N_i(\xi,\eta)x_i \quad , \quad y(\xi,\eta) = \sum_{i=1}^{n_e} N_i(\xi,\eta)y_i \tag{8}$$

where:

- n_e is the nodes number of the considered element. Its value is fixed on 3 for our bi-dimensional triangulation.
- ξ and η are the coordinates in the reference element.
- $x(\xi,\eta)$ and $y(\xi,\eta)$ are the coordinates of one point of the real element.
- x_i and y_i are the coordinates of the i^{th} node of an element.
- $N_i(\xi,\eta)$ are interpolation functions. Where in our case:

$$N^T = \begin{bmatrix} 1-\xi-\eta \\ \xi \\ \eta \end{bmatrix} \tag{9}$$

The Jacobian matrix of the geometric transformation in the general case writes:

$$\begin{aligned}[J(\xi,\eta)] &= \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n_e} \frac{\partial N}{\partial \xi} x_i & \sum_{i=1}^{n_e} \frac{\partial N}{\partial \xi} y_i \\ \sum_{i=1}^{n_e} \frac{\partial N}{\partial \eta} x_i & \sum_{i=1}^{n_e} \frac{\partial N}{\partial \eta} y_i \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial N_1}{\partial \xi} & \frac{\partial N_2}{\partial \xi} & \frac{\partial N_3}{\partial \xi} \\ \frac{\partial N_1}{\partial \eta} & \frac{\partial N_2}{\partial \eta} & \frac{\partial N_3}{\partial \eta} \end{bmatrix} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \end{bmatrix} \end{aligned} \tag{10}$$

Let us consider the temperature distribution $T(x,y;t)$. It can be expressed as follows:

$$\begin{aligned} T(x,y;t) &= [N_1(x,y), N_2(x,y), N_3(x,y)] \begin{Bmatrix} T_1(t) \\ T_2(t) \\ T_3(t) \end{Bmatrix} \\ &= \mathbf{N}(x,y) T(t) \end{aligned} \tag{11}$$

$$\begin{pmatrix} \frac{\partial T}{\partial x} \\ \frac{\partial T}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial T}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial T}{\partial \eta}\frac{\partial \eta}{\partial x} \\ \frac{\partial T}{\partial \xi}\frac{\partial \xi}{\partial y} + \frac{\partial T}{\partial \eta}\frac{\partial \eta}{\partial y} \end{pmatrix} = J^{-1} \begin{pmatrix} \frac{\partial T}{\partial \xi} \\ \frac{\partial T}{\partial \eta} \end{pmatrix} \tag{12}$$

$$\begin{pmatrix} \frac{\partial T}{\partial \xi} \\ \frac{\partial T}{\partial \eta} \end{pmatrix} = \begin{pmatrix} \frac{\partial N_1}{\partial \xi} & \frac{\partial N_2}{\partial \xi} & \frac{\partial N_3}{\partial \xi} \\ \frac{\partial N_1}{\partial \eta} & \frac{\partial N_2}{\partial \eta} & \frac{\partial N_3}{\partial \eta} \end{pmatrix} \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \end{Bmatrix} = B\, T_e \tag{13}$$

where T_1, T_2 and T_3 are the nodal temperatures of a given element.

Then, the local gradient is expressed as:

$$\nabla T_e = J^{-1} B T_e \tag{14}$$

Let us define the gradient field along the direction X as $\mathbf{U}$:

$$U = (1\ 0)\ \nabla T \tag{15}$$

Equation (15) can be expressed in its weak form on a volume Ω as following:

$$\int_\Omega T^* U d\Omega = \int_\Omega T^* \frac{\partial T}{\partial x} d\Omega \tag{16}$$

The mass and the stiffness of the reference element Ω_r are:

$$\mathbf{m} = \int_{\Omega_r} \mathbf{N}(\xi,\eta)\ \mathbf{N}^T(\xi,\eta)\ \det(\mathbf{J})\ d\Omega_r \tag{17}$$

$$\mathbf{k} = \int_{\Omega_r} \mathbf{N}(\xi,\eta)\ (1\ 0)\ \mathbf{J}^{-1}\ \mathbf{B}\ \det(\mathbf{J})\ d\Omega_r \tag{18}$$

The defined mass and stiffness elements are assembled in order to obtain the global matrices $\mathbb{M}_{(N\times N)}$ and $\mathbb{K}_{(N\times N)}$. A lumped form can be used for $\mathbb{M}$. Then, equation (15) becomes:

$$\mathbb{M}\mathbf{U} = \mathbb{K}T \tag{19}$$

The gradient of T along X can be written as follows:

$$\mathbf{U} = \mathbb{M}^{-1}\mathbb{K}T = \mathcal{D}_x T \tag{20}$$

$\mathcal{D}_x$ denotes the gradient operator along X. A similar development is performed in the direction Y to obtain $\mathcal{D}_y$.
The matrices U, V and $\mathcal{L}$ are defined respectively as the gradient of temperature T along X and Y and its Laplacian.

$$\begin{cases} \mathbf{U} = \mathcal{D}_x T \\ \mathbf{V} = \mathcal{D}_y T \\ \mathcal{L} = (\mathcal{D}_x \mathcal{D}_x + \mathcal{D}_y \mathcal{D}_y) T \end{cases} \tag{21}$$

2.4.3 Learning dataset

Considering a combination of parameters (P_1, P_2), the resulting learning sample a_i from a given data of a node i is:

$$a_i = [t_1^i,..,t_4^i,\ \mathbf{U}^{t_1^i},..,\mathbf{U}^{t_4^i},\ \mathbf{V}^{t_1^i},..,\mathbf{V}^{t_4^i},\ \mathcal{L}^{t_1^i},..,\mathcal{L}^{t_4^i},\ T_1 - T_{max}^{t_1^i},..,\ T_4 - T_{max}^{t_4^i}]$$

where $\{t_1^i,..,t_4^i\}$ correspond to the critical instants $\{t_L, t_{DC}, t_{DR}, t_S\}$ of the considered node i, $\{T_1,..,T_4\}$ are the critical temperatures $\{T_L, T_{DC}, T_{DR}, T_S\}$ related to the used alloy, $T_{max}^{t_j^i}$ is the maximum nodal temperature at the instant t_j^i, $\mathbf{U}^{t_j^i}$, $\mathbf{V}^{t_j^i}$ and $\mathcal{L}^{t_j^i}$ are respectively the gradients along X and Y and the Laplacian of temperature in the considered node at the instant t_j^i.
Each learning sample a of a considered node will be correlated with the corresponding nodal porosity using supervised learning.

2.5 Machine learning approaches and experiments

2.5.1 Machine learning approach

The main purpose is to create a model that predicts the value of local porosity by learning the decision rules inferred from the simulated data. Decision Tree (DT) is one of the available solutions that can be used for this purpose. A tree can be seen as a piece-wise constant approximation and could mimic the conditional algorithms of porosity simulation. Starting from a root node, a tree estimator is built by selecting the variable that best splits the data. At each step, a score is computed per variable to evaluate the quality of the split node until getting to the leafs

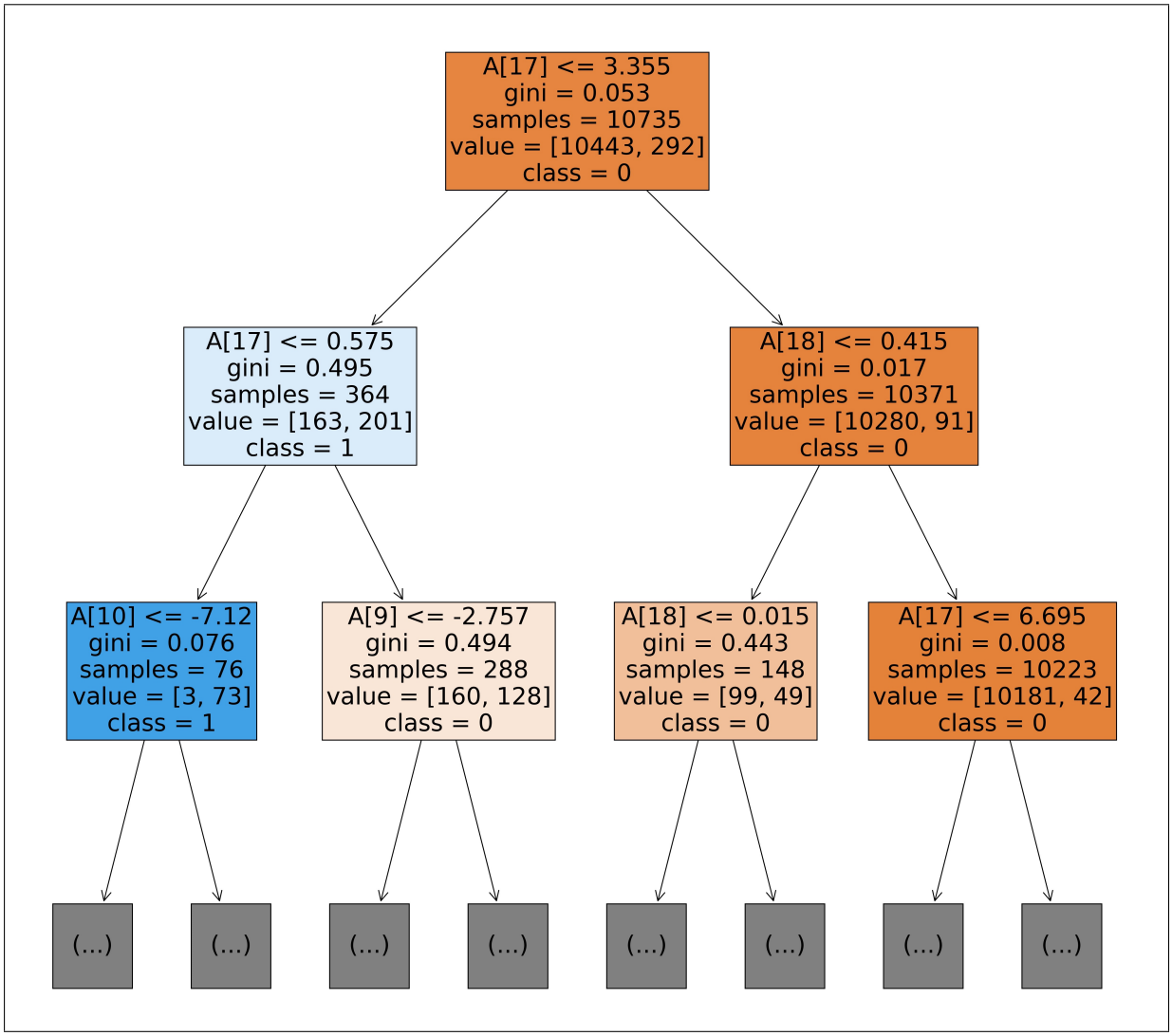

Figure 9: First 3 layers of a classification tree

or the maximum depth D_{max} imposed by the user. In this work, Gini index is used as the evaluation score which measures the probability of misclassifying an input sample. An example of a classification tree extracted from our work is illustrated in figure 9 where $A[i]$ denotes the ith feature of the input data, *gini* is the value of the Gini index, *sample* is the considered number of samples in the node, *value* is the number of samples per category and *class* represents the corresponding category.

Derived from this general approach, Ensemble Trees is a family of methods that can be used in our solution for more accuracy. The driving principle is to train N_e independent tree estimators and aggregate their results to yield the final predictions, by majority vote in classification problems and arithmetic average in regression. The datasets used to train these estimators depend on the adopted Ensemble Trees method. For Extremely randomized trees (ET) [10], the entire dataset is used to train each tree. However, in Random Forest method (RF) [11], a subset is built for each estimator using sampling with replacement, known as Bootstrapping. The cited Ensemble Trees methods (ET and RF) are tested in our solution along with the DT method. The minimum number of samples required to build a split node and a leaf, denoted respectively $N_{s_{min}}$ and $N_{l_{min}}$, should be defined in the hyperparameters configuration as well as D_{max} and N_e for Ensemble Trees.

In our bi-dimensional case, the results of Procast simulation shows that porosity locates in a small region of the whole domain (around 3% of the total area for the Hourglass test case). Consequently, a strategy is adopted to predict porosity incrementally in two steps (as illustrated in figure 10):

- Porosity localization using a classifier that gives a binary output per node to evaluate the presence of porosity in the considered point.
- Porosity value prediction on the previously localized nodes.

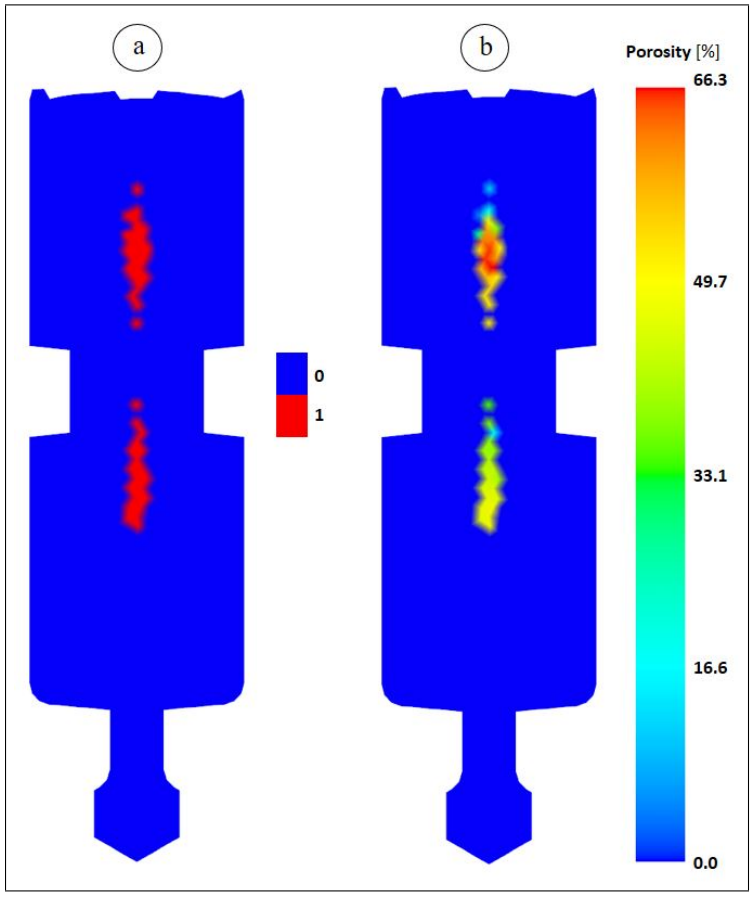

Figure 10: Classifier and regressor predictions
a: classifier prediction ; b: regressor prediction

The considered machine learning methods are tested on the prepared datasets with the hyperparameters configuration presented in table 2. The purpose of this evaluation is to find the most suitable method to predict shrinkage porosity.

	Classifier			Regressor		
	DT	RF	ET	DT	RF	ET
D_{max}	50	50	50	50	50	50
$N_{s_{min}}$	2	2	2	2	2	2
$N_{l_{min}}$	1	1	1	1	1	1
N_e	N/A	100	20	N/A	100	100

Table 2: Hyperparameters configuration of the tested methods

2.5.2 Results of machine learning

The prepared data is splitted to two portions and shuffled to break its order. The first subset is used to perform a training (80% of the entire data in our case).Then the learned model pass over an evaluation step with the rest of the data where two metrics are computed:

- Root mean square error (RMSE): an evaluation of the accuracy by computing the mean difference of the n samples between the target y and the predicted value $\hat{y}$:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{22}$$

- R^2 score: the proportion of the variance in the dependent variable that is predicted from the independent variables:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{23}$$

  where $\bar{y}$ is the mean of the predicted values $\hat{y}$.

The errors are computed also during the training to measure the ability of the method to learn the desired property and during the testing to evaluate its efficiency.

	Classifier errors				Regressor errors			
	Train		Test		Train		Test	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
DT	0.00	1.00	0.079	0.798	0.00	1.00	0.105	0.178
RF	0.009	0.996	0.066	0.857	0.036	0.942	0.041	0.872
ET	0.00	1.00	0.057	0.893	0.00	1.00	0.045	0.849

Table 3: Metrics computed on machine learning identified methods

Regarding to the computed errors repported in table 3, the evaluation revealed that decision trees based methods are efficient to predict local porosity since they resulted a satisfying *RMSE*

and R^2 scores for classification and regression. Especially the promising results obtained by ET can prove the ability of the method to learn efficiently porosity prediction.

Based on the field of applicability, this approach builds a set of unpruned classification or regression trees. The fully randomized generation of cut point and attribute candidates combined with ensemble averaging contributes to reduce variance which helps to improve the accuracy of the resulting predictions.

2.5.3 Features relevance and redundancy evaluation

After preparing the learning dataset, an evaluation of the selected features should be performed. "minimal-Redundancy-Maximal-Relevance" (mRMR) is used to minimize the redundancy and maximize the relevance between a feature set and a target in order to come with an optimal dataset. This method consists of finding an optimal set of features that is mutually and maximally dissimilar and can represent the response variable effectively.

Let us consider the matrix $\mathbb{A}_{(N\times N_f)}$ as the input dataset where $N = n.K$ rows contain the prepared learning samples with n is the number of nodes in the mesh and K is the number of the considered parameters combinations (P_1, P_2). N_f denotes the number of features in the dataset. $\mathbf{P}$ is defined as the target vector that contains the N values of nodal porosity.

The relevance score of the ith feature represented by the column $\mathbf{A_{.i}}$ is the F-statistic F_i of this column with the target vector $\mathbf{P}$. In regression, the latter is computed as follows:

$$F_i = \frac{\gamma^2_{\mathbf{A_{.i}},\mathbf{P}}}{N(1-\gamma^2_{\mathbf{A_{.i}},\mathbf{P}})} \tag{24}$$

where $\gamma^2_{\mathbf{A_{.i}},\mathbf{P}}$ denotes the cross correlation coefficient between the evaluated feature $\mathbf{A_{.i}}$ and the target $\mathbf{P}$.

In classification, F-statistic can be expressed as:

$$F_i = \frac{\sum_{c=1}^{2} n_c(\bar{\mathbf{A}}^{\mathbf{c}}_{\mathbf{.i}} - \bar{\mathbf{A}}_{\mathbf{.i}})^2}{\sigma^2} \tag{25}$$

where $\sigma^2 = \sum_{c=1}^{2} \frac{(n_c-1)\sigma_c^2}{n-1}$ is the pooled variance and σ_c^2 is the variance of the evaluated feature within the cth class. $\bar{\mathbf{A}}_{\mathbf{.i}}$ is the mean value of the column $\mathbf{A_{.i}}$ and $\bar{\mathbf{A}}^{\mathbf{c}}_{\mathbf{.i}}$ is the mean value of $\mathbf{A_{.i}}$ within the cth class. n_c is the number of training samples of the cth class [12].

The relevance score of a subset of features S can be formulated as:

$$D_S = \frac{1}{|S|}\sum_{i\in S} F_i \tag{26}$$

The results of feature relevance scores in regression and in classification are reported in figure 11.

The second criteria in features selection is the redundancy. It is expressed as a variable—pairwise score that evaluates the redundant information between two variables. The redundancy of a subset of features S is:

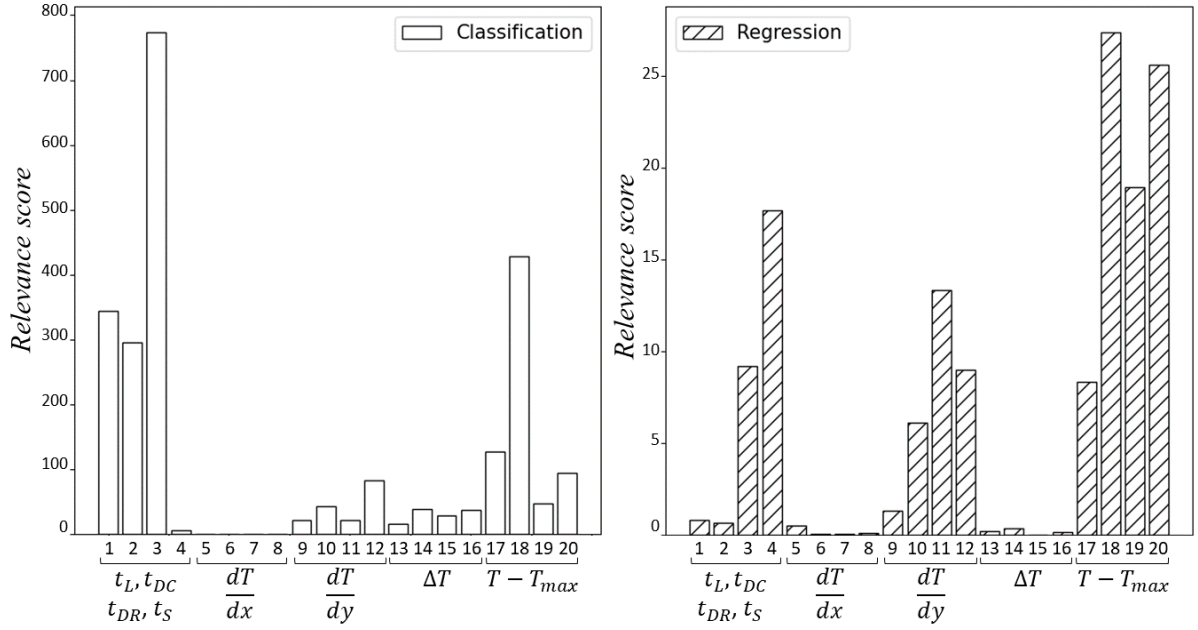

Figure 11: Importance score of the selected features candidates

$$R_S = \frac{1}{|S|^2} \sum_{i,j \in S} \rho_{i,j} \tag{27}$$

where $\rho_{i,j}$ denotes the Pearson correlation coefficient between the ith and the jth features.

The redundancy the considered dataset $\mathbb{A}$ is represented in a triangular matrix as illustrated in figure 12.

mRMR is an iterative process that employs sequential forward selection scheme seeking to maximize a score m_f that considers the relevance of the considered feature f and its redundancy with the previously selected features. At an iteration i, the score of the feature f is calculated using the following formula:

$$m_i(f) = \frac{D_{f,p}}{\frac{\sum_{f_s \in s} R_{f,f_s}}{i-1}} \tag{28}$$

where s is the subset of the selected features until the iteration $i-1$ and p is the target.

The iterative process stops when reaching the number of most relevant features N_f that is defined a priori.

In the purpose of determining the most fine selection of features, 15 training are realized using the selected methods on sets of features that are resulted from mRMR process. In this experiment N_f is varied from 5 to 20 in order to build the tested feature sets.

The results of the evaluation showed in figure 13 prove that Extra-Trees (ET) is the most efficient approach in comparison with the tested methods. The most successful combination is composed of 18 features with eliminating these components from the initial dataset:

- $\frac{\partial T_{t_L}}{\partial x}$ and $\frac{\partial T_{t_S}}{\partial x}$ from the classification training set.

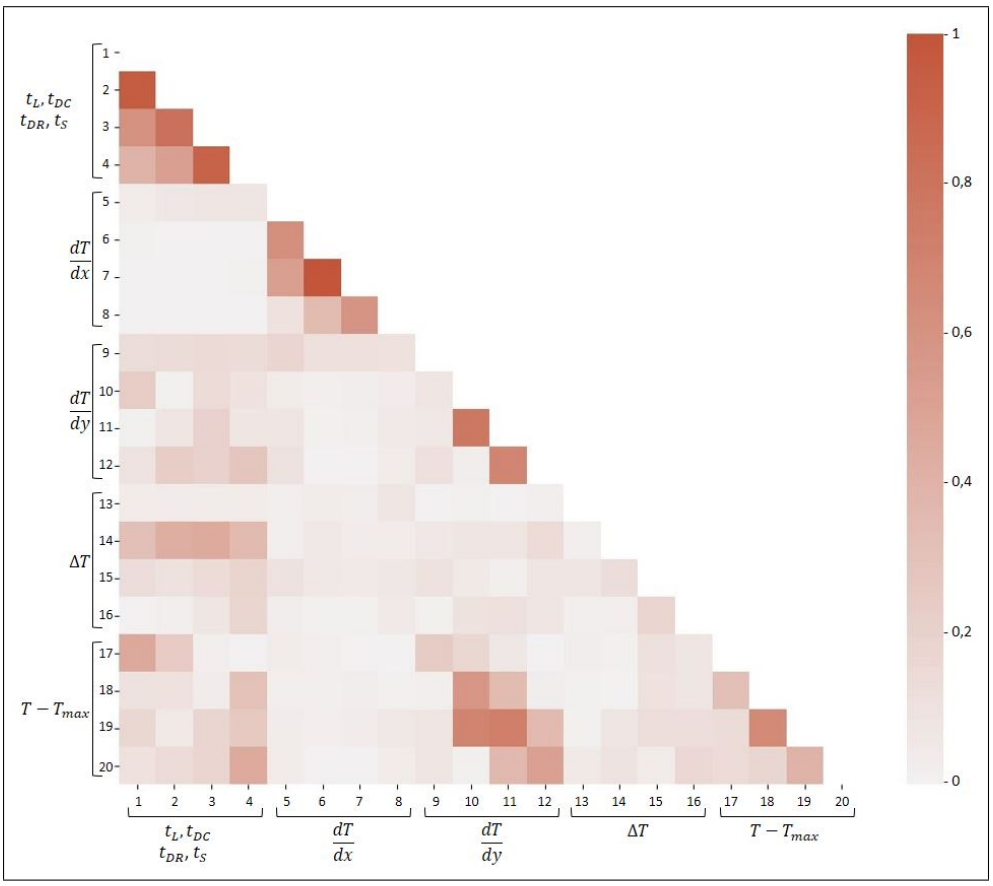

Figure 12: Redundancy matrix of the selected features candidates

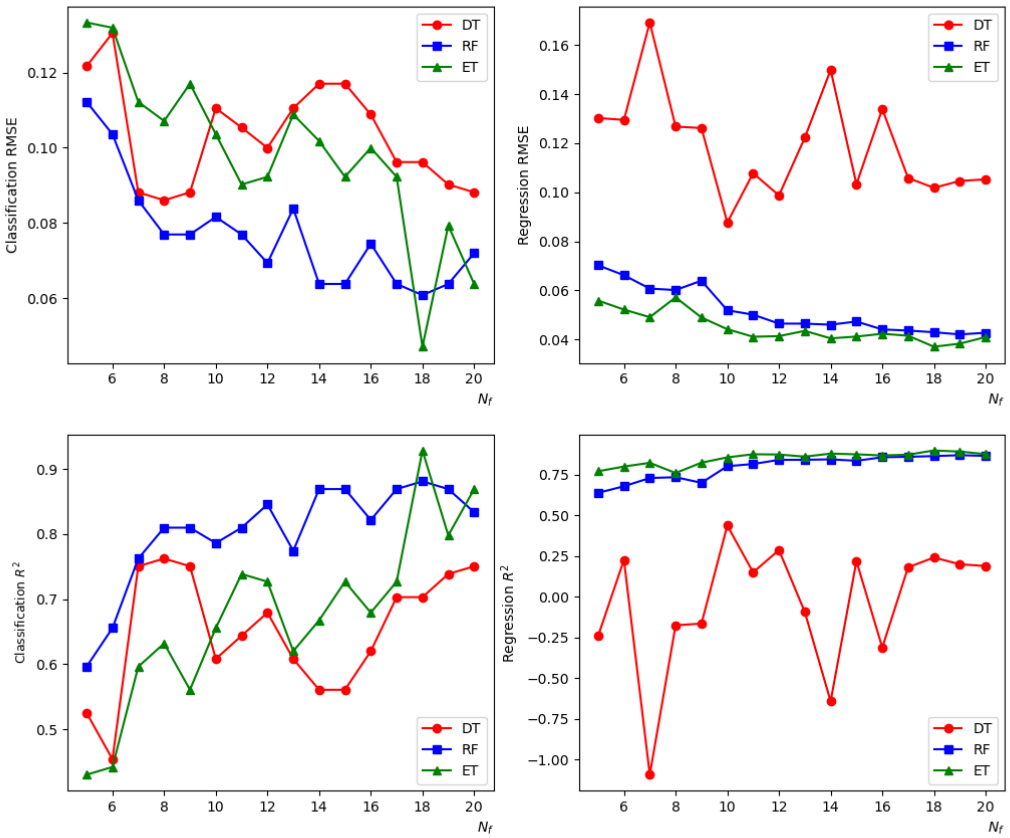

Figure 13: Metrics in terms of the number of selected features N_f

- t_S and ΔT_{t_S} from the regression training set.

Regarding to the limited resistance to liquid and mass feedings since they occur at low solid fraction, interdendritic feeding is considered to be the most important stage for producing porosity defects. This fact can explain the elimination of features related to solidus from the regression training set in the result of mRMR. The elimination of solidus and liquidus temperature gradients along X from the classification training set might be related to geometry of our study case.

3 Results

In this section, results of application of the proposed solution on our study case "Hourglass" are presented.

Following the detailed process, the thermal history interpolation is required as a first prediction step. The approach had been tested on the identified study case with 9 modes. The results of four parameter combinations are presented below in figure 14. The latter represent the projection of the 5 dimensional space (composed of (X,Y) geometrical space, time step and parameters P_1 and P_2).

During the building of the dataset, the mesh is processed to eliminate voids due to their unsettled result of temperature map in order to avoid noise in the training set.

Examples of four predicted temperature maps with different parameters combinations and time steps are illustrated in figure 14. The corresponding L2 norm of the difference between the simulated temperature history and the interpolated fields (denoted $\|T - \hat{T}\|_2$) are presented in the same figure. One parameter is varied in each combination to show the sensibility of the used approach to each parameter. For example, when P_2 is lower than P_1 like in case "b", the higher area cools down first as it is visible in the figure.

After proving the performances of decision trees based methods and choosing Extra Trees for our porosity prediction solution regarding to its satisfying metrics, the method is tested on our study case and result of porosity predictions are presented below. Predictions and training data are assembled in the same figure 15 to evaluate the integrity of the obtained results. A visual comparison can be performed between the simulated shrinkage porosity (represented with outlined parts in figure 15) and the rest of the illustrated results that comes from our solution prediction of porosity distribution.

The obtained result of porosity prediction seems to be corresponding to the target. The results of porosity distribution, is following the logic of parameters variation. Porosity spots tend to get higher with increasing P_1 or decreasing P_2. The results of the ET regressor predictions with the 18 selected features against the simulated porosity are represented in figure 16 and confirm the prediction accuracy.

4 Conclusion

The wider adoption of artificial intelligence in strategic industrial sectors to improve the quality of the products and avoid microstructure anomalies like shrinkage porosity is nowadays confirmed. In the present work, a porosity prediction model is proposed using supervised learning.

Major findings of this work are summarized below:

- Machine learning approach Extra-Trees is applied to predict porosity distribution from a

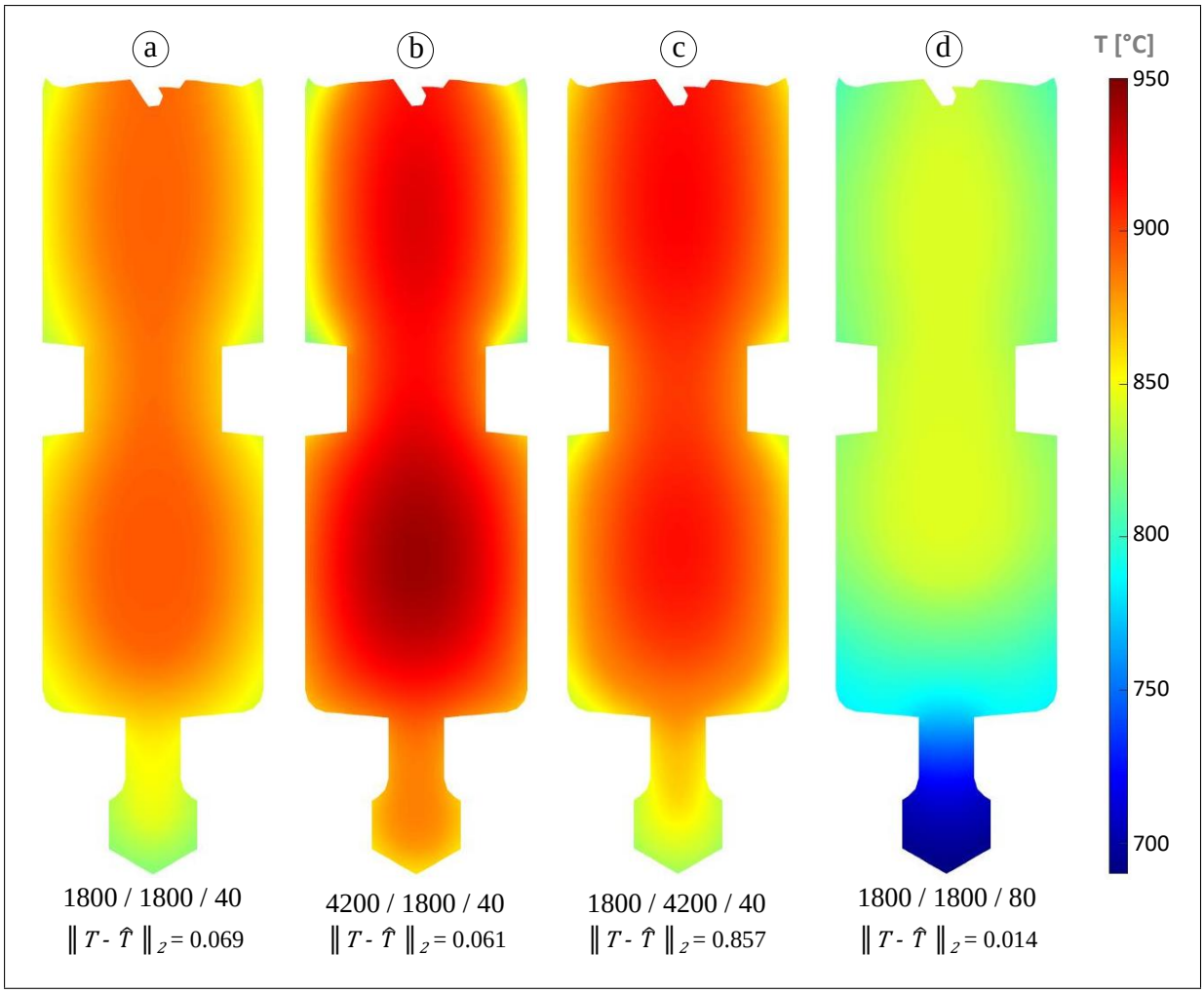

Figure 14: Temperature field predictions on the ”Hourglass” with random parameters combinations
each parameters combination composed of:
P_1 value $[W/m^2.K]$ / P_2 value $[W/m^2.K]$ / Time step [s]

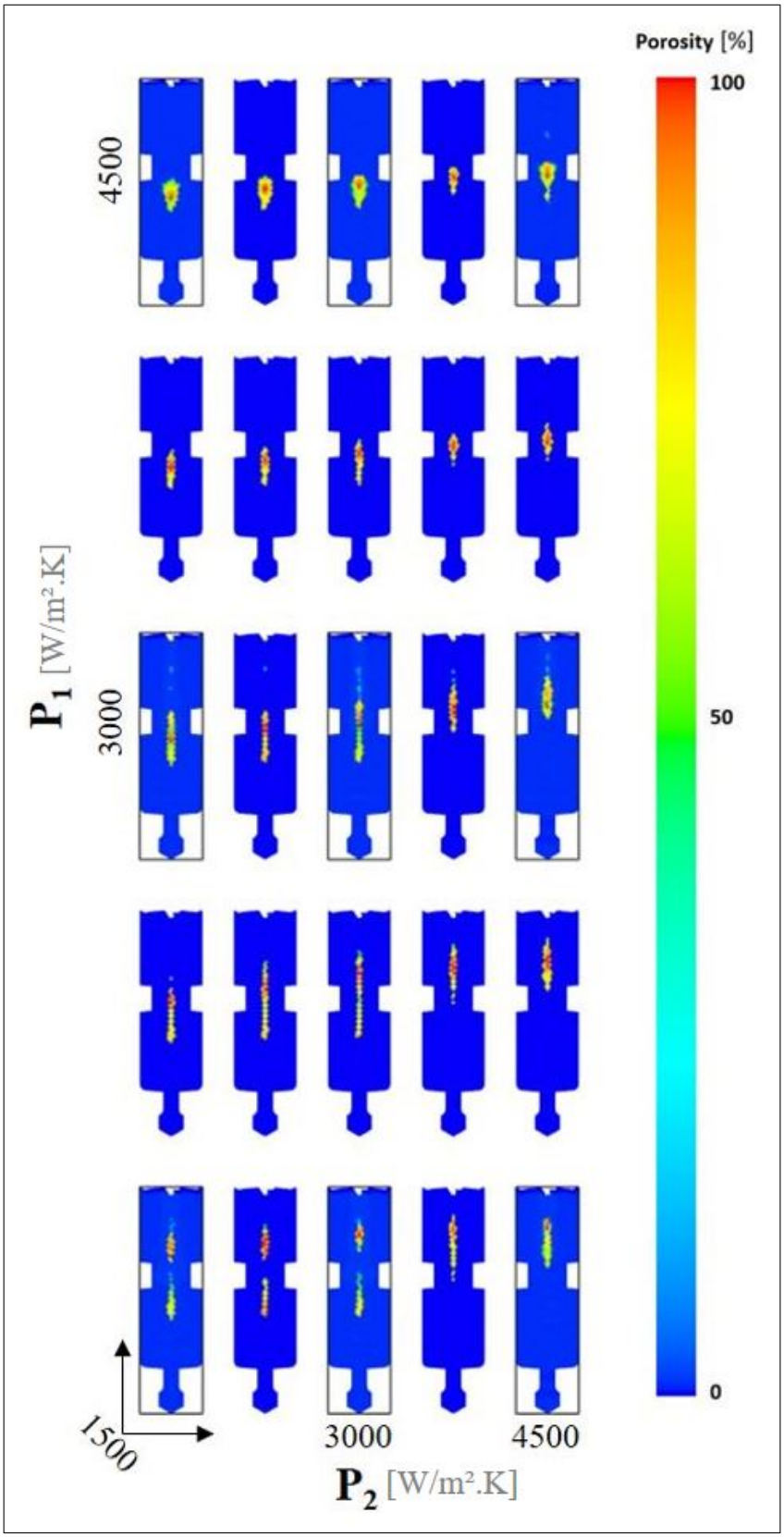

Figure 15: Our solution prediction and simulated shrinkage porosity

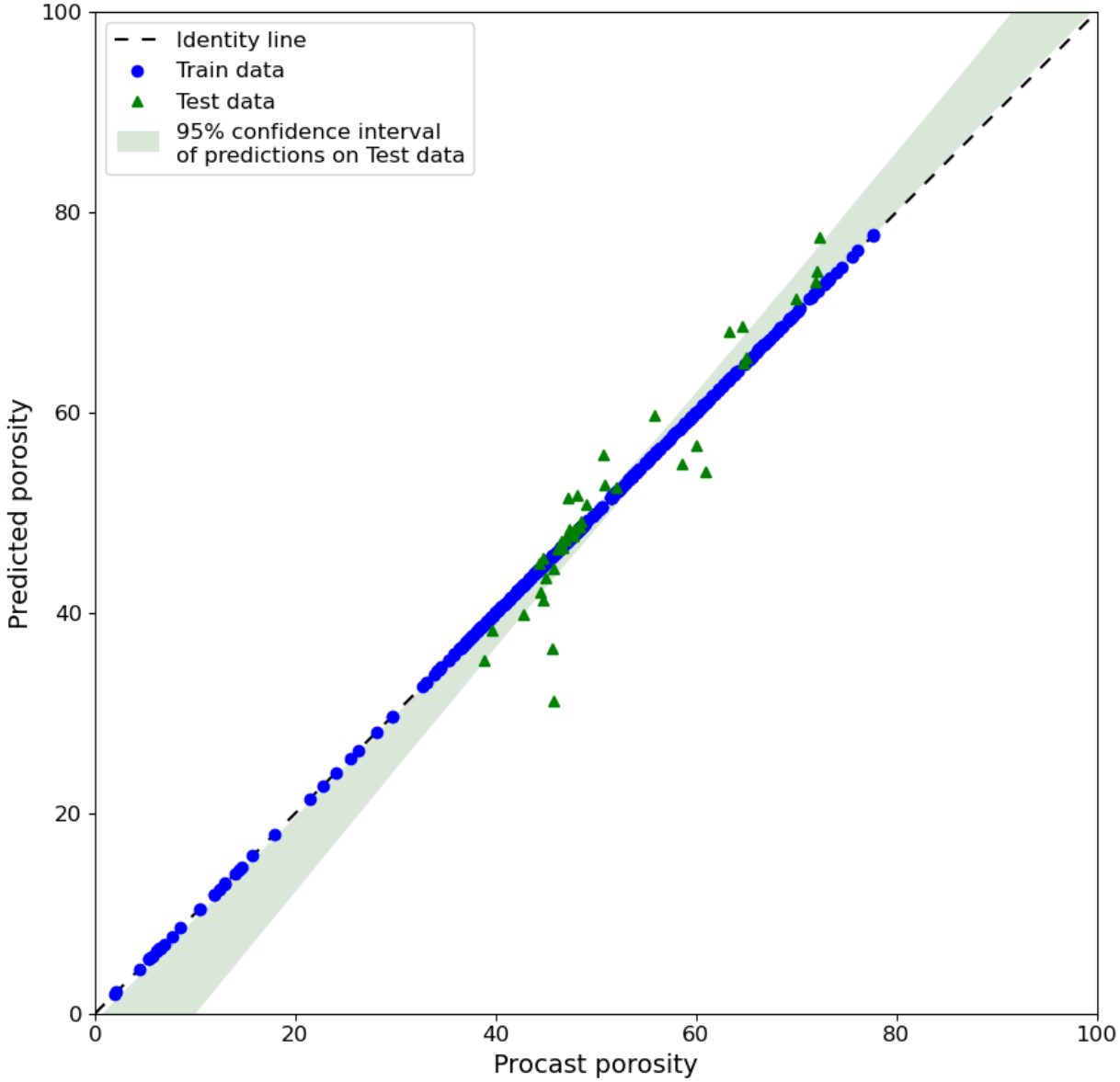

Figure 16: Porosity prediction of our solution against Procast simulation

set of 18 features. The obtained results are satisfying with respect to the metrics (around 4,7% of RMSE and 0,92 of $R^2$for the classification and 3,7% of RMSE and 0,9 of $R^2$for the regression).

- The learning is realized on critical features that aggregates each time serie of more than 13000 components in a vector of 18 scalars including critical temperature, gradients and Laplacian.
- SVD combined with a modal representation of thermal history serve to interpolate its values on new parameters combination.
- The proposed methodology allows porosity prediction from identified parameters. This method could assist the user to control porosity distribution according the application and improve casting design experience.

Appendix A Iterative SVD

There is an alternative approach that is used in this work to perform SVD. The solution is searched in a separated representation. After calculating the first m terms of the finite sum decomposition, T is expressed as:

$$T = \sum_{i=1}^{m} F_i \otimes G_i + R_i \otimes S_i \tag{29}$$

where $F \in \mathbb{R}^l$ and $G \in \mathbb{R}^K$ contains the computed modes of the rank-one greedy constructor and m is the number of performed modes. $R \in \mathbb{R}^l$ and $S \in \mathbb{R}^m$ are the computed functions, at each iteration, that approximates T:

$$\begin{aligned} T &= RS^T \\ TS &= R\langle S,S\rangle \\ R &= \frac{TS}{\langle S,S\rangle} \end{aligned} \tag{30}$$

The same approach is used to define S.
At the first iteration, F and G are initialized using the following expression:

$$\begin{cases} F_1 = \dfrac{R\sqrt{\|R\|\|S\|}}{\|R\|} \\ G_1 = \dfrac{S\sqrt{\|S\|\|R\|}}{\|S\|} \end{cases} \tag{31}$$

The approximated functions are optimized in an iterative process that recomputes R and S and updates F and G at each iteration. At the enrichment step m, R and S are calculated as follows:

$$\begin{cases} R = \dfrac{TS - \sum_{i=1}^{i=m} F_i G_i^T S}{\langle S,S\rangle} \\ S = \dfrac{T^T R - \sum_{i=1}^{i=m} G_i F_i^T R}{\langle R,R\rangle} \end{cases} \tag{32}$$

The process stops when reaching the predefined error or the maximum number of iterations.

Acknowledgements. Authors acknowledge the support of ESI forms its chair at Arts et Metiers Institute of Technology as well as Angers Loire Métropole that co-financed this work.

Conflict of Interest. The authors declare that they have no conflict of interest.

REFERENCES

[1] M. J. Couper, A. E. Neeson, and J. R. Griffiths. Casting defects and the fatigue behaviour of an aluminium casting alloy. 13(3):213–227.

[2] Ch. Pequet, M. Rappaz, and M. Gremaud. Modeling of microporosity, macroporosity, and pipe-shrinkage formation during the solidification of alloys using a mushy-zone refinement method: Applications to aluminum alloys. 33(7):2095–2106.

[3] Manas Dash and Makhlouf Makhlouf. Effect of key alloying elements on the feeding characteristics of aluminum–silicon casting alloys. 1:251–265.

[4] V. D. Tsoukalas. Optimization of porosity formation in AlSi9cu3 pressure die castings using genetic algorithm analysis. 29(10):2027–2033.

[5] Quang-Cherng Hsu and Anh Tuan Do. Minimum porosity formation in pressure die casting by taguchi method. 2013:e920865. Publisher: Hindawi.

[6] S. H. Mousavi Anijdan, A. Bahrami, H. R. Madaah Hosseini, and A. Shafyei. Using genetic algorithm and artificial neural network analyses to design an al–si casting alloy of minimum porosity. 27(7):605–609.

[7] Xue-dan Gong, Dun-ming Liao, Tao Chen, Jian-xin Zhou, and Ya-jun Yin. Optimization of steel casting feeding system based on BP neural network and genetic algorithm. 13(3):182–190.

[8] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. 14(5):403–420.

[9] Geoffrey K. Sigworth and Chengming Wang. Mechanisms of porosity formation during solidification: A theoretical analysis. 24(2):349–364.

[10] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[11] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[12] Xiaoxing Yang, Ke Tang, and Xin Yao. The minimum redundancy – maximum relevance approach to building sparse support vector machines. pages 184–190, 2009.