

Machine learning techniques for the ab initio Bravais lattice determination

Esther-Lydia Silva-Ramírez¹ | Inmaculada Cumbreira-Conde² | Rafael Cano-Crespo³ |
Francisco-Luis Cumbreira³ 

¹Department of Computer Science and Engineering, University of Cádiz, Puerto Real, Spain

²Department of Private International Law, Macquarie University, Sydney, New South Wales, Australia

³Departamento de Física de la Materia Condensada, Universidad de Sevilla, Sevilla, Spain

Correspondence

Francisco-Luis Cumbreira, Departamento de Física de la Materia Condensada, Universidad de Sevilla, 41012 Sevilla, Spain.
Email: fcumbreras@us.es

Abstract

Machine learning-based algorithms have been widely applied recently in different areas due to its ability to solve problems in all fields. In this research, machine learning techniques classifying the Bravais lattices from a conventional X-ray diffraction diagram have been applied. Indexing algorithms are an essential tool of the preliminary protocol for the structural determination problem in crystallography. The task of reverting the obtained information in reciprocal lattice to direct space is a complex issue. As an alternative way to afford this problem, different machine learning algorithms have been applied and a comparison between them has been conducted. The obtained accuracy was 95.9% using 10-fold cross-validation (while the best result obtained so far has been 84%). A model based on Bragg positions was our unique predictor, allowing us to obtain the set of the interplanar lattice distances. Our model was successfully checked with a complex example. In addition, our procedure incorporates the following advantages: robustness versus imprecision in data acquisition and reduction of the amount of necessary input data. This is the first time so far that such classification has been carried out in true ab initio condition.

KEYWORDS

Bravais lattices, crystallography, machine learning

1 | INTRODUCTION

Machine learning (ML) methods have achieved recently outstanding contributions in the materials research community as well as in other science domains (Agatonovic-Kustrin & Beresford, 2000; Bhadeshia, 1999; Dai et al., 2020; Scott et al., 2007; Sha & Edwards, 2007; Shetty et al., 1999; Woinaroschy et al., 2000; Zhang et al., 2008). ML developments include methods which simulate brain working, for instance, artificial neural networks (ANN), or simulate human experience and draw conclusions as expert systems. Somehow, we can state that we are facing the solution of traditional complex problems with other different paradigm which offers greater speed and efficiency. The goal of our work is concerned to the solution of an outstanding problem of crystallography: the assignment of the Bravais lattice in structural determination. We will approach that key problem from conventional ML methods. Following, we address the task of laying out the fundamentals of such a problem.

Crystallography is the science that studies the atomic arrangements of crystalline solids. It is an axiomatic discipline that relies on two postulates: The postulates of Bravais and Schoenflies–Fedorov. First of them, known as micro-periodicity principle, states the foundation for

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Expert Systems* published by John Wiley & Sons Ltd.

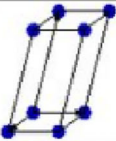
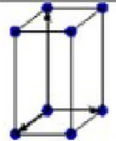
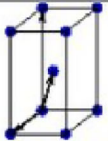
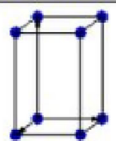
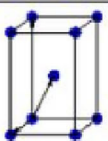
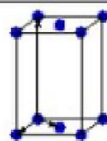
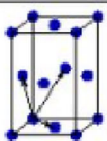
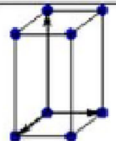
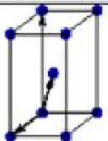
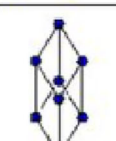
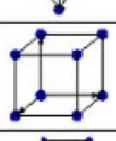
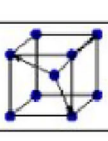
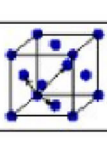
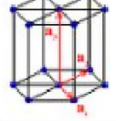
Bravais lattice	Parameters	Simple (P)	Volume centered (I)	Base centered (C)	Face centered (F)
Triclinic	$a_1 \neq a_2 \neq a_3$ $\alpha_{12} \neq \alpha_{23} \neq \alpha_{31}$				
Monoclinic	$a_1 \neq a_2 \neq a_3$ $\alpha_{23} = \alpha_{31} = 90^\circ$ $\alpha_{12} \neq 90^\circ$				
Orthorhombic	$a_1 \neq a_2 \neq a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90^\circ$				
Tetragonal	$a_1 = a_2 \neq a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90^\circ$				
Trigonal	$a_1 = a_2 = a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} < 120^\circ$				
Cubic	$a_1 = a_2 = a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90^\circ$				
Hexagonal	$a_1 = a_2 \neq a_3$ $\alpha_{12} = 120^\circ$ $\alpha_{23} = \alpha_{31} = 90^\circ$				

FIGURE 1 The 14 Bravais lattices in three dimensions. Source: D.V. Anghel, Bravais lattice table, 2003

translational symmetry. All possible arrangement of congruent points in a crystalline solid (those which remain invariants under the operations of an algebraic group of translations) are called Bravais lattices. Group theory states that they are only possible 14 Bravais lattices in 3-D space. Figure 1 shows the 14 possible Bravais Lattices classified according to the seven crystal systems. Rows represent each one of the seven crystal systems. The second column indicates the geometric cell parameters and successive columns a drawing of the unit cell according to the criteria:

- Lattice points only at the vertices (primitive cells).
- Lattice points at the vertices and in the center of the cell (body-centered cells).
- Lattice points at the vertices and in the center of the basal faces (base-centered cells).
- Lattice points at the vertices and in the center of all faces (face-centered cells).

The first step in solving an unknown crystal is the determination of the unit cell (the repeating unit), as a precondition for the subsequent determination of the relative positions of atoms within that cell. In fact, two separate kinds of information can be extracted from diffraction spectra: the first one comes from the geometrical arrangement of the reflections, which gives us the information about the crystal lattice and the symmetry of the crystal. This kind of information source is obtained by means of the indexing algorithms (IA in this article). The second one comes from the intensity of reflections and is concerned with the information about the cell content.

The so called indexing algorithms are an essential part of every data collection software package. The well-known underlying principle of indexing methods is universal: they are based on matching experimental scattering vectors to some vectors of the reciprocal lattice (the algebraic

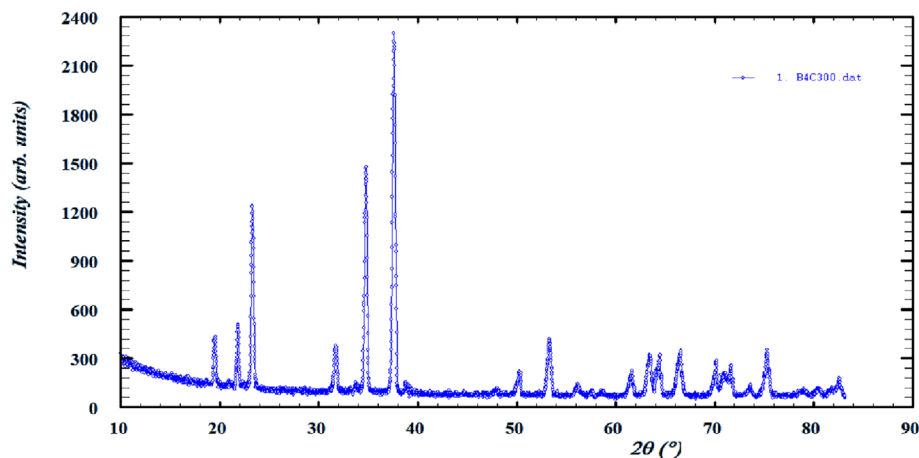


FIGURE 2 Diffraction spectra of boron carbide

dual lattice in the wave vector space). The most extensively used are *TREOR90* (Werner et al., 1985), *DICVOL91* (Boultif & Louër, 2004) and *ITO15* (Visser, 1969), which work on very different approaches to solve the problem. Those current approaches fall into two main categories (Habershon et al., 2004): exhaustive search or deductive recognition of relationship between subsets of data.

In practice, the measurement of the positions of the so called *Bragg peaks* is equivalent to the measurement of the modules of a subset of the vectors of the reciprocal lattice, in fact the shorter ones, without any knowledge about orientation relations. The task of reverting that limited information to direct space, in order to build the correct Bravais lattice, is a complex issue. Figure 2 shows a diffraction spectra of Boron Carbide manufactured and obtained in our laboratory. Horizontal axis is the scattering angle (2θ) and vertical axis is the intensity in arbitrary units. The Bragg peaks are superposed on an inelastic smooth background which lacks of crystallographic relevance.

By means of any peak-hunter algorithm we obtain the positions of the maxima which can be introduced in Bragg law:

$$\lambda = 2 \cdot d_{hkl} \sin \theta, \quad (1)$$

to obtain the interplanar lattice distances of the (hkl) family of planes.

Now, the equation to solve is

$$Q_{hkl} = \frac{1}{d_{hkl}^2} = f(a, b, c, \alpha, \beta, \gamma, h, k, l), \quad (2)$$

where a, b, c, α, β and γ are the linear and angular parameters of the unit cell and the (hkl) Miller indices. Miller indices are unknown three integer numbers that specify the orientation of plane lattice satisfying the Bragg law. All the parameters in $f(a, b, c, \alpha, \beta, \gamma, h, k, l)$ are unknown.

Further obstacles to the correct indexing are related to geometrical ambiguities because while a unit cell defines the lattice, a lattice can be described in infinite different ways. Proof of lack of satisfaction are the new attempts published in bibliography. As a sign of that, Jacobson (1997) proposed a cosine Fourier transform to obtain direct space cell vectors. Coelho (2017) proposed the same way a new indexing algorithm, independent of peak positions, to overcome the issue of inaccuracy in the peaks measurement.

It is tempting to afford the indexing problem by means of ML methods, based on learning (training and validation) from a selected database. Following, we will expose the previous attempts so far. Recently, Liang et al. (2020) made available to public the tool called Crystal Structure Prediction Network (CRYSPNet) which is based on neural networks and can predict several features as Bravais lattice, lattice parameters of inorganic materials based on its chemical composition. The code has been trained and validated over a large number of the Inorganic Crystal Structure Database (ICSD) entries. Despite their results clearly improve other strategies, the higher accuracy reached is about 84% for the metal group set. The limitations were justified on the basis of class imbalance, particularly for the low symmetry compounds. We agree with that observation but we will suggest later that ANN models are by far less efficient than other ML techniques for this kind of classification problem. In any case its algorithm is not universal as it is restricted to groups of inorganic compounds.

In the same direction of thought is the work of Ryan et al. (2018). His approach, based on ANN, demonstrates the ability of deep learning to extract meaningful information from large repository data. Although the evaluation of crystalline structures falls within the scope of their work, it was not contemplated the simple but crucial challenge of identifying the Bravais lattice from the only predictor of Bragg positions obtained from the diffraction spectrum.

Another outstanding contribution includes the work of Lee et al. (2020), which reports a deep learning protocol for phase identification and quantification in multiphase inorganic compounds in the quaternary Sr–Li–Al–O. Finally, we refer to the relevant work of Wang et al. (2020) and Oviedo et al. (2019). In both contributions, the authors, by starting from a large set of predictors, made crystallographic classification limited to metal organic frameworks or seven space groups, respectively.

It is the first time a study related to all sort of materials has been conducted by using the interplanar lattice distances as only predictor. The aim is a fast and reliable method of assignation of Bravais lattice avoiding the traditional indexing algorithms. None of the classic algorithms are conclusive versus complex problems. In fact, practice suggest the combined use of the three algorithms to compare matches and differences.

Our work concerns to the bottleneck task in the preliminary treatment of X-ray data. We have focused our attention in that point (Bravais lattice assignation), once this problem has been solved, the determination of lattice parameters becomes trivial. Then, our challenge is to address the problem by using as only predictor the list of interplanar lattice spacing d_{hkl} . In addition, we intend our method to be applicable to all sort of materials and not restricted to a particular subset as in previous attempts.

Then, the originality of this work lies in the fact that we afford a crucial problem in crystallography based on a single predictor, without the support of large repository data and applicable to all kinds of materials. Previous attempts were restricted to particular subclasses as metals or ceramics as in references: Lee et al. (2020); Liang et al. (2020), and others.

The article sequence is as follows. In Section 2, the applied methodology is described. The main results are presented in Sections 3 and 4. The following section explores the reduction of the input data and the selection of the more significant attributes, Section 5. Section 6 addresses the issue of robustness face to inaccuracies in data acquisition or face to the presence of intruders or missing data. In section 7, the models were tested face to a very complex case. Finally, Section 8 deals with the most prominent conclusions.

2 | METHODOLOGY

In this Section the methodology followed in this work is described. We try to highlight the challenge that the study carried out in this research represents.

The explicit mode of Equation (2) is shown in Equation (3) where the lattice spacings d_{hkl} are related to the unit cell parameters trough the complex relation

$$\frac{1}{d_{hkl}^2} = \left[\begin{array}{c} \frac{h}{a} \left| \begin{array}{ccc} h/a & \cos\gamma & \cos\beta \\ k/b & 1 & \cos\alpha \\ l/c & \cos\alpha & 1 \end{array} \right| + \\ \frac{k}{b} \left| \begin{array}{ccc} 1 & h/a & \cos\alpha \\ \cos\gamma & k/b & \cos\alpha \\ \cos\beta & l/c & 1 \end{array} \right| + \\ \frac{l}{c} \left| \begin{array}{ccc} 1 & \cos\gamma & h/a \\ \cos\gamma & 1 & k/b \\ \cos\beta & \cos\alpha & l/c \end{array} \right| \end{array} \right] \cdot \left| \begin{array}{ccc} 1 & \cos\gamma & \cos\beta \\ \cos\gamma & 1 & \cos\beta \\ \cos\beta & \cos\alpha & 1 \end{array} \right|^{-1}, \quad (3)$$

and in turn these relate to the Bragg position by means of Bragg law, Equation (1). From this equation, we can infer the advantage of choosing lattices spacing as descriptor instead of Bragg positions, since for the latter choice the dependence with lattice parameters is complicated by the presence of an arcsin law.

Although in 3-D space there are 14 Bravais lattices, Figure 1, in practice hexagonal unit cell can be described as rhombohedral by means of:

$$\begin{aligned} a_r = b_r = c_r &= \sqrt{\frac{a_h^2}{3} + \frac{c_h^2}{9}} \\ \alpha_r = \beta_r = \gamma_r &= \cos^{-1} \left(\frac{2c_h^2 - 3a_h^2}{2(c_h^2 - 3a_h^2)} \right), \end{aligned} \quad (4)$$

where a_r , b_r , c_r , α_r , β_r and γ_r are lattice parameters for rhombohedral unit cell while a_h and c_h are the lattice parameters for hexagonal unit cell.

In a similar way, rhombohedral lattices can be expressed as hexagonal ones. That is why we have reduced the number of classes to 13 by grouping together the hexagonal and rhombohedral primitive lattices. As a matter of fact, we have established the following correspondence:

Primitive cubic	1
Body centered cubic	2
Face centered cubic	3
Primitive tetragonal	4
Body centered tetragonal	5
Primitive orthorhombic	6
Body centered orthorhombic	7
Base centered orthorhombic	8
Face centered orthorhombic	9
Primitive monoclinic	10
Base centered monoclinic	11
Triclinic	12
Hexagonal and trigonal	13

Our starting set was made from 500 entries, 100 from the ICSD database (Bergerhoff et al., 1983) and the remaining from the crystal open database (COD) data set (Vaitkus et al., 2021). The selection was random so as to avoid any bias, and the only restriction was having at least 3% of each class. The histogram of the starting data set is shown in Figure 3. This graph represents how many instances are classified in each class. For example, for class 1 there are 29 records, 32 instances are classified in class 2, and so on.

At this respect, the percentage of each Bravais lattice in the ICSD data set can be read in Liang et al. (2020).

Two facts are well known:

1. The suite of lattice distances is monotonously decreasing with the scattering angle.
2. Traditional methods need at least the first 20 reflections as input data, to avoid ambiguities.

With regard to both points, all lattice distances were divided by the largest one which is the first. This way we have a sequence of real numbers starting from 1 and monotonously decreasing afterwards. According to the usual methods, we have chosen the first 20 lattice distances, although the normalization procedure reduces by one unit the number of significant input data. In this way, we have 20 relationships with 66 unknown: 6 defining the metric of the unit cell and 20 triples of integer numbers (hkl). Once we solved Equation (3), we obtained the unit cell and the set of (hkl) integers which allow us to build the corrected Bravais lattice. In Table 1 an extract of data set is shown. The first 19 columns represent the described attributes and the last column represents the target variable for the classification problem.

For the particular case of Bravais lattices of the cubic system, Equation (3) reduces to:

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}}. \quad (5)$$

For this particular case of cubic symmetry, the sequence of normalized data does not depend on lattice parameters, but only on the integer Miller indexes. This leads to an easily identifiable pattern. For the rest of cases, Equation (3) shows the complexity of the problem and the inability

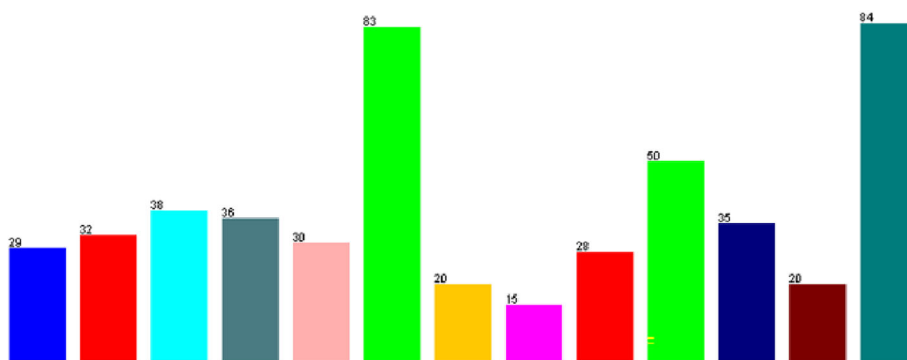


FIGURE 3 Classes distribution in the data set

TABLE 1 An extract of data set

Attributes	Target class																	
	0.894	0.748	0.639	0.599	0.573	0.555	0.501	0.498	0.459	0.445	0.445	0.428	0.413	0.407	0.385	0.382	0.373	7
1	0.863	0.691	0.640	0.540	0.507	0.476	0.448	0.443	0.432	0.410	0.367	0.363	0.346	0.340	0.322	0.317	0.313	6
1	0.707	0.632	0.500	0.471	0.447	0.426	0.392	0.378	0.354	0.343	0.333	0.324	0.301	0.283	0.277	0.272	0.263	4
1	0.760	0.667	0.613	0.555	0.496	0.432	0.411	0.392	0.369	0.349	0.329	0.308	0.285	0.273	0.264	0.255	0.245	13

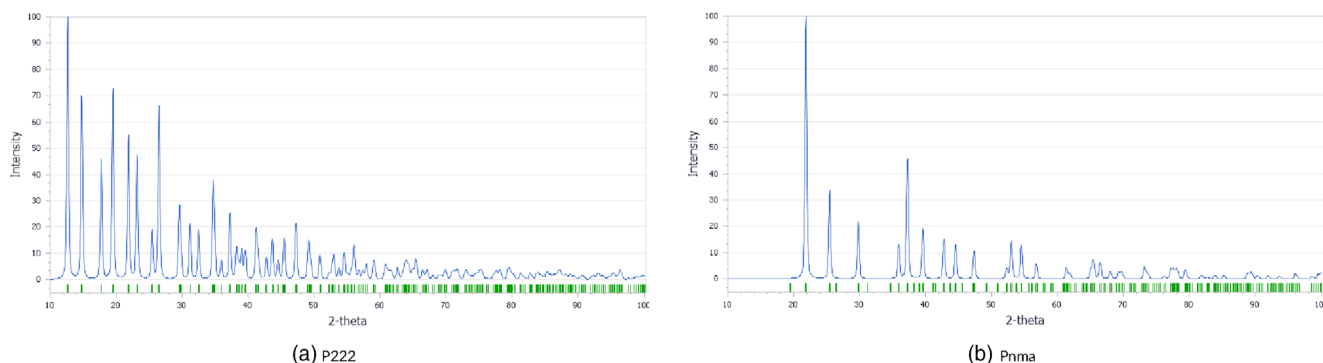


FIGURE 4 Simulated X-ray diffraction pattern for two materials with the same Bravais lattice but different space group

to obtain predictable patterns. Another additional consideration to take into account is that every Bravais lattice is associated with a subset of the 230 Space Groups of Schoenflies-Fedorov. For each group of symmetry certain operations leave a trace in the diffraction pattern systematically cancelling certain intensities, following rules called systematic absences.

In Figure 4a and b, we show powder X-ray diffraction diagrams for compounds with two different space groups of the same Bravais lattice (primitive orthorhombic). The first is a symmorphic group (P222) whereas the second is Pnma which show systematic absences due to screw axes and glide planes (in fact Pnma is a supergroup of $P2_12_12_1$). In both figures, horizontal axis is the scattering angle (2θ) and vertical axis is the intensity in arbitrary units. The small markers at the bottom indicate the Bragg positions.

If we remark the Bragg peak positions, regardless its intensities, patterns are very different despite coming from the same Bravais lattice. Therefore, we can state we are facing up a formidable challenge without the hope of any predictable pattern.

3 | COMPUTATIONAL TOOLS

With regard to ANN we have used the multiple back propagation (MBP) software developed by Lopes and Ribeiro (2010). This program incorporates an architecture *multiple feed forward network* which allows a partition in sub-spaces for the mapping function relating to inputs and outputs. It also implements neurons with selective activation. Experimental results on benchmarks showed the superiority, in most cases, over classical multilayer perceptron networks (Lopes & Ribeiro, 2011).

The other ML models were accessed through the graphical interface Weka (Frank et al., 2016). This software allows you to preprocess a big set of data. It also applies different ML algorithms and compares various outputs. Weka was developed by the University of Waikato in New Zealand, this term stands for *Waikato Environment for Knowledge Analysis*. The methods used in the experiments were:

- Naïve Bayes** A probabilistic classifier applying Bayes theorem.
- k-NN** A clustering supervised algorithm.
- Chirp** Method based on Hypercube description regions (HDR).
- Furia** A fuzzy rule learner.
- J48** A decision tree implementing the ID3 algorithm (iterative dichotomiser).
- Meta-estimators** random tree, random forest and extra tree.

4 | PRELIMINARY RESULTS

Two kind of experiments were done, using different methods for validation in order to measure the ability to predict new data avoiding selection bias and overfitting:

- Train/test split: data are shuffled randomly and a split 2/3 is used for training and the other split 1/3 for testing the model.
- k-fold cross-validation (CV): the data are split into k folds, so the model is trained on $k - 1$ folds and then it is tested on the other fold. The process is repeated until each unique group has been used as the test set. Exhaustive performances were performed with a partition parameter equal to 10.

TABLE 2 Results from preliminary experiments

Models	Train/test split	10-fold CV
MBP	37.0	22.0
Naïve Bayes	14.1	13.6
k-NN	27.1	32.4
J48	39.5	37.3
Chirp	31.5	29.8
Furia	41.7	29.9
Random Tree	50.2	59.6
Extra Tree	50.3	59.5
Random Forest	50.0	61.3

TABLE 3 Results from experiments with preprocessed data

Models	Train/test split	10-fold CV	k-statistic
Naïve Bayes	28.6	32.6	0.19
k-NN	82.4	89.6	0.89
J48	79.3	85.7	0.79
Chirp	76.5	75.2	0.81
Furia	71.8	81.0	0.76
Random Tree	92.5	95.5	0.93
Extra Tree	90.0	83.5	0.90
Random Forest	84.7	90.6	0.92

The results with the percentages of correctly classified records are shown in Table 2. The first column represents the applied model, the second the percentage of correctly classified instances using the HoldOut validation method versus the 10-fold Cross-Validation technique in the last column.

We draw several conclusions:

- Superiority of meta-estimators against ANN model for these kinds of problems. From now on we will no longer consider ANN methods.
- While ML models are promising, however the test accuracy is not significant for both splitting and cross validation experiments.

All of above leads us to make the decision of pre-process the data to overcome the imbalance between classes or overfitting. With such a goal in mind we have used two supervised filters already included in Weka: Smote and ClassBalancer. The first one is a synthetic minority oversampling method. On the other hand, the second one reweights the instances so that each class has the same total weight.

After this we have applied again the ML models, the obtained results are described in Table 3, including the results of κ -statistic from 10-fold CV experiments accounting for the possibility of agreement by chance (Smeeton, 1985).

In Figure 5 a sketch of these results is shown. The x-axis represents the different studied models and the y-axis represents the accuracy (%) of the classification for both splitting and cross validation experiments. It is observed that most of the models present a higher accuracy when working with the 10-fold CV validation method. Except for the Naïve Bayes model, all models present an accuracy greater than 75%, highlighting the Random Tree model, which exceeds 95%.

5 | REDUCTION IN INPUT SPACE

The following Section explores the selection of the more significant attributes and the reduction of the input data. We intend to determine if there is an improvement in the performance of the models when a reduction of the data set is carried out.

5.1 | Improvement by attribute selection

For the process of classification, we selected 20 attributes which were eventually reduced to 19 because of the normalization task. Afterwards, we raised the question of its relevance since some of them might be irrelevant or superfluous. In fact, attribute selection is a mining data method

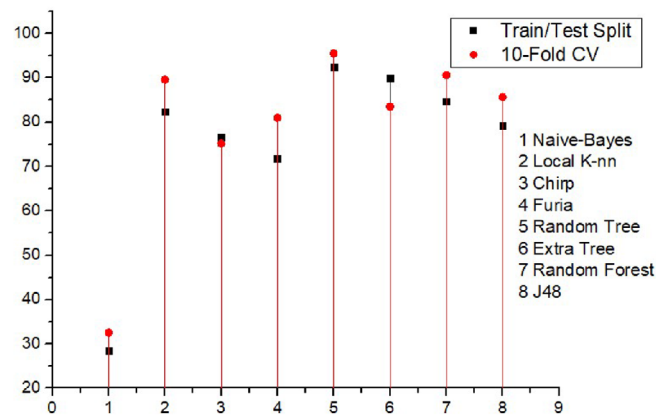


FIGURE 5 Graphic representation of performance from the experiments

TABLE 4 Selected attributes according to acceptance criteria

Attributes	Significance (%)	Acceptance
1	0	
2	100	X
3	60	X
4	40	
5	40	
6	80	X
7	20	
8	60	X
9	20	
10	60	X
11	20	
12	20	
13	60	X
14	20	
15	0	
16	60	X
17	60	X
18	0	
19	20	
20	20	

to extract the ranking of attributes. The goal is not only to reduce the processing time, but also to improve the metrics of the classification algorithms. The underlying philosophy is very well explained in the article of Gnanambal et al. (2018).

Attribute selection is a two-step process, the first one is a searching process and the following one deals with ranking. The PART algorithm was first applied according to the recommendations in Gnanambal et al. (2018). For the second step we applied the heuristic Greedy stepwise method (Butterworth et al., 2004). The results are shown in Table 4. We have highlighted those attributes that exceed 55% significance which it is clearly observed in Figure 6. The y-axis represents the percentage of significance, while the x-axis represents each of the attributes. The threshold value has been marked with a horizontal line. All those attributes that exceeded this value were selected.

After this task we have repeated the classification based only on the highlighted attributes and the most successful algorithm (Random Tree). Thus, we improved our results from 95.5 to 95.9 in 10-fold CV. While this improvement seems slight, we showed nonetheless effectiveness of attribute selection.

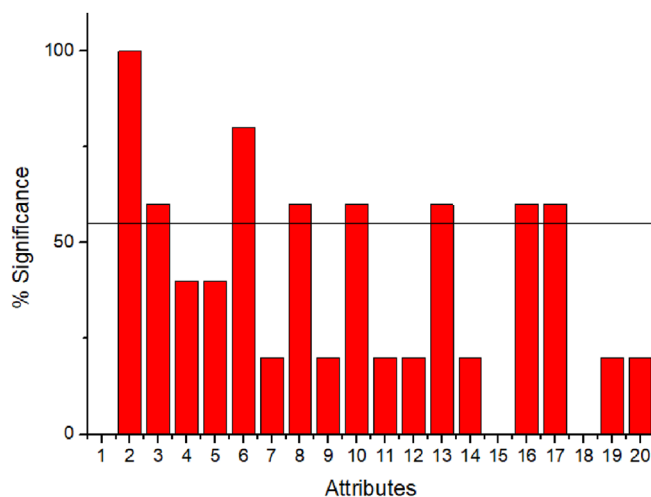


FIGURE 6 Acceptance criteria to select attributes which exceed the threshold value

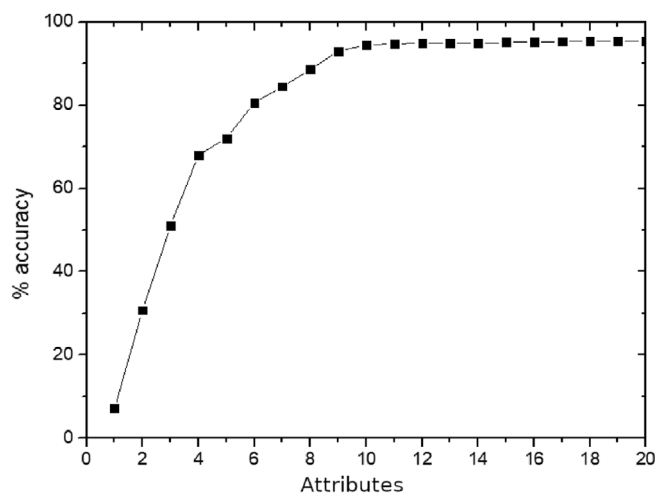


FIGURE 7 Accuracy with reduction of input data

5.2 | Reduction of input data

Data acquisition for indexing purposes requires, at least, three times the number of parameters that fix the unit cell. If the symmetry is unknown, we need six unit-cell parameters and consequently, most automatic indexing programs require a minimum of 20 reflections as input data. In a natural way our attributes are sorted in a descending order. After this, we raised the question that perhaps a lower number might be needed for ML methods. This way, and starting from our whole set of 20 attributes we will sequentially remove attributes in an ascending order, that is, starting from the smallest one.

The results are represented in Figure 7, in which the y-axis represents the percentage of accuracy achieved with the reduction of input data and the x-axis represents the attributes. We can deduce that it can be maintained an accuracy over 90% with the first 10 terms only. For any further reduction the decrease is very prominent, decaying quickly to a random probability of 100/13% ($\approx 7.7\%$). In a previous subsection we already put across that only eight attributes were really significant (at least for ML models). Thus, ML methods allow us to simplify the input data, versus the strict demand of traditional approaches.

6 | ROBUSTNESS

This section addresses the issue of robustness face to inaccuracies in data acquisition or face to the presence of intruders or missing data.

6.1 | Robustness versus imprecise data acquisition

It is a common fact that some diffraction diagrams contain heavily overlapped domains where it is very difficult to determine the Bragg peak positions accurately. The consequences of these inaccuracies may be critical, especially in the low angle domain (first attributes). In order to check the robustness of the methodology we have modified randomly our attributes from the second onwards. To this end, we have generated 19 random numbers in the interval $(-1,1)$ that we have multiplied by a number ε which adopts successively the values (0.01, 0.02, 0.03 and so on). For a given value of ε , the set of attributes, from the second onwards, are shifted multiplying its actual value by the factor $(1 + \varepsilon)$. The coefficient ε represents the percentage of inaccuracy measurement. In Figure 8 we show the worsening of accuracy relative to the random displacements. On the horizontal axis we represent the inaccuracy (noise) in the data acquisition while the vertical axis shows the accuracy in the classification.

We can appreciate that the algorithm is robust, at least until a level of noise of $\varepsilon = 8(8\%)$. A severe relative error for Bragg positions is about 1% and from the Bragg equation this is equivalent to an inaccuracy of 1% in lattice distances. Accordingly, we can confirm outright that our procedure is robust face to noise in the peaks hunting data acquisition.

6.2 | Pitfalls in the low angle domain. Omitted or intrusive peaks

It is well known that photon counting statistics follows a Poisson noise distribution which degrades quality of spectra. The signal to noise ratio is strongly dependent on the data acquisition time. We often meet difficult choices concerning the doubt of whether an observed peak corresponds to a true Bragg reflection or to an impurity peak, a noise signal or a Kalfa2 shoulder (a superfluous right-sided companion due to characteristic X-rays). Two possible wrong outcomes might happen: omission or intrusion. Besides, it is possible to ignore some faint reflections having intrinsically low structure factor moduli.

Experience shows that consequences are less important in the high angle domain but are extremely critical in the low angle region (large interplanar lattice distances). To check the consequences of each wrong choices, we have.

- omitted the second attribute.
- introduced an intruder value in that second position and so moving to one single place the others.

For the last case, the intrusive position was a random real number between one and the old second attribute value. In Table 5, we illustrate the obtained results, the true positive ratio (TPR) is showed for each Bravais lattice omitting the second attribute as well as adding an intruder value. This is, the correctly predicted positive values which means that the value actual class and the value of predicted class are the same. And Table 6, the accuracy (%) of correctly classified observations and k-statistic are described in both cases, omitting the second attribute as well as adding an intruder value.

The first obvious conclusion is that the method is not robust facing this situation, as in traditional methods. In any case we can say that false inclusion gives rise to a more pronounced worsening of results. In fact, crystallographers know the rule: in doubt it is better to omit. The third

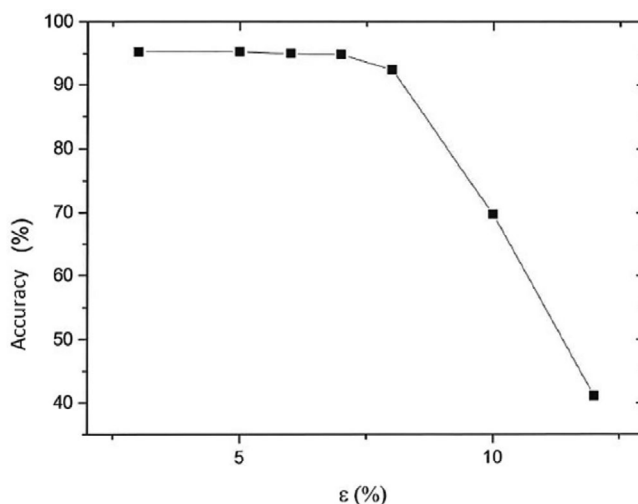


FIGURE 8 Accuracy with imprecise data

TABLE 5 TPR for each Bravais lattice

Attributes	Omission	Intrusion
1	0.008	0.012
2	0.143	0.071
3	0.833	0.583
4	0.007	0.178
5	0.429	0.075
6	0.101	0.012
7	0.099	0.111
8	0.002	0.007
9	0.167	0.011
10	0.171	0.070
11	0.005	0.005
12	0.073	0.230
13	0.016	0.002

Note: Bold values represents the third class which stands out from the rest of Bravais lattices.

TABLE 6 Results according to omission or intrusion

Metric	Omission	Intrusion
Accuracy	16.700	9.400
k-statistic	0.089	0.012

class, corresponding to the resilient fcc lattice (face centered cubic), was highlighted in Table 5 because it stands out singularly from the rest of Bravais lattices with regard to the TPR factor.

7 | RESULTS AND DISCUSSION

In this work, we approach the problem of the assignation of the Bravais lattice in structural determination applying ML methods. So, the different ML-based algorithms were applied to the data set manufactured and obtained in our laboratory, whose extract was shown in Table 1. In the present research, the accuracy was calculated, measuring the classification ability of the models. As mentioned above, Random Tree model offers the best results, exceeding 95%, as observed in Figure 5. In general, all the models present an good behaviour with an accuracy greater than 75% with the 10-fold CV validation method.

From the results, it is noticeable that the accuracy of the Random Tree is very high. Focusing on this classification model, the rules were analysed, as well as the extensive tree obtained. A total of 482 classification rules were extracted from decision tree, so a part of the constructed Random Tree in Weka is shown in Figure 9 and the corresponding rules generated for this tree branch are represented in appendix. The rules serve as a condition that when it is satisfied, it returns an equivalent predicted classification.

In Random Tree, each node is split using the best among the subset of predictors chosen at that node, using the gain to split them. So, we can observe that the first dichotomy (the greatest gain of information) refers to attribute 13. The following dichotomies concern the attributes 2 and 6. The alternative for the second attribute is critical (< 0.69 or ≥ 0.69) and results in a branch to either the Bravais lattices with high symmetry or the opposite, respectively.

After the creation of the tree, the training observations are grouped at the terminal nodes. To predict a new observation, the tree is traversed based on the value of its predictors until one of the terminal nodes is reached. In the case of classification, the mode of the response variable is used as the prediction value, that is, the most frequent class of the node. The leaf or terminal nodes are represented by rectangle and decision nodes are represented by oval.

The tree obtained greatly facilitates the tasks of classifying the Bravais lattices from a conventional X-ray diffraction diagram. It is only necessary to take the values of the attributes of the new instance to classify and traverse the tree to obtain a correct classification.

In addition, we have gone further in trying to test the model in extreme cases and observe its behaviour. In this way, a complex case was tested which is described below.

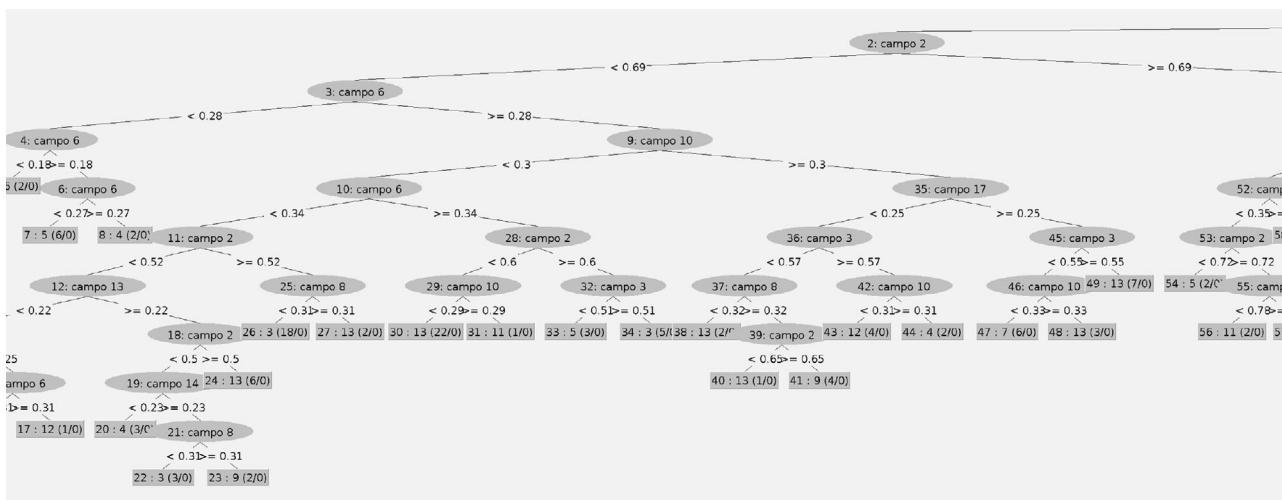


FIGURE 9 Tree branches for random tree

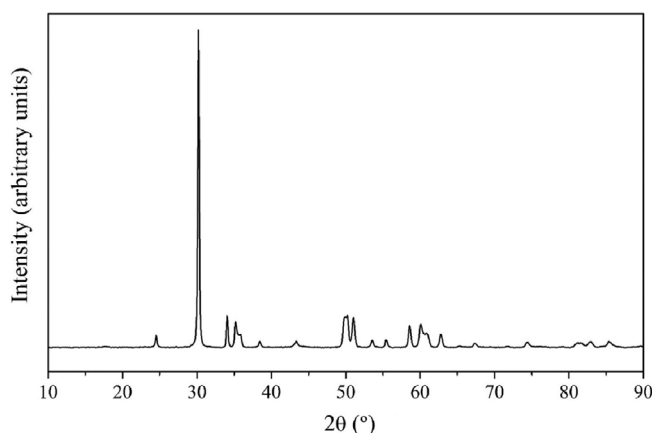


FIGURE 10 X-ray pattern diffraction of Ta-doped zirconia

In the Ta-doped zirconia system, Cumbreira et al. (2018) observed, for the first time, a new orthorhombic crystal structure coming from a symmetry-breaking distortion of the tetragonal polymorph of zirconia. It is well-known that pure ZrO_2 exists at ambient pressure in three polymorphic variants depending on the temperature:

1. Monoclinic ZrO_2 (m- ZrO_2 ; space group P21/c) from room temperature up to 1205°C.
2. Tetragonal ZrO_2 (t- ZrO_2 ; space group P42/nmc) in the range 1170–2370°C.
3. Cubic ZrO_2 (c- ZrO_2 , space group Fm3m) from 2370°C up to the melting point at 2680°C.

In practice, all the variants may be stabilized at room temperature through alloying oxides (e.g., CaO, MgO, Y_2O_3 , CeO_2 and others). In our work, by using Ta-doping, we obtained a new orthorhombic ceramic compound which seems to be immune to aging and shows an interesting ferroelectric behaviour.

Indexation of the X-ray diffraction pattern led to two possible alternative solutions. The first one is a centric crystal structure with a monoclinic cell and a space group C2/c, and the other is an acentric crystal structure with an orthorhombic cell and a space group Pca2₁ (which is a maximal subgroup of the tetragonal P42/nmc space group). In this case, the ferroelectric behaviour of the material allowed to rule out the monoclinic base centered option because of its incompatibility with symmetry centers. Figure 10 shows the corresponding diffraction diagram. The x-axis is the scattering angle 2θ (the double of the Bragg angle), while the y-axis is the diffracted intensity in arbitrary units.

Thus, we have tested our best trained model face to that crucial proof. After introducing in the testing set the experimental input attributes regarding to Ta-doped zirconia, we also introduced as target the incorrect choice of monoclinic base centered lattice (class 11). The model was able to detect and correct that wrong assignation, so the model repositioned the nominal output class as the correct primitive orthorhombic (class

```

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  <-- classified as
0  0  0  0  0  0  0  0  0  0  0  0  0  | a = 1
0  0  0  0  0  0  0  0  0  0  0  0  0  | b = 2
0  0  0  0  0  0  0  0  0  0  0  0  0  | c = 3
0  0  0  0  0  0  0  0  0  0  0  0  0  | d = 4
0  0  0  0  0  0  0  0  0  0  0  0  0  | e = 5
0  0  0  0  0  0  0  0  0  0  0  0  0  | f = 6
0  0  0  0  0  0  0  0  0  0  0  0  0  | g = 7
0  0  0  0  0  0  0  0  0  0  0  0  0  | h = 8
0  0  0  0  0  0  0  0  0  0  0  0  0  | i = 9
0  0  0  0  0  0  0  0  0  0  0  0  0  | j = 10
0  0  0  0  0  1  0  0  0  0  0  0  0  | k = 11 ←
0  0  0  0  0  0  0  0  0  0  0  0  0  | l = 12
0  0  0  0  0  0  0  0  0  0  0  0  0  | m = 13
    ↑

```

FIGURE 11 Confusion matrix in Weka for Ta-doped zirconia

6), as observed in confusion matrix supplied by Weka, shown in Figure 11. The incorrect class 11 was correctly classified as the correct one, class 6.

8 | CONCLUSIONS

To the best of our knowledge this is the first ab initio successful attempt to classify Bravais lattices starting with interplanar lattice distances as the only predictor. This new procedure for a quick assignment of the Bravais lattice is a very significant step forward for the crystallographer's routine. The procedure reliability allows us to avoid the tedious time-consuming trials and errors associated to classical approaches when applied to new unknown structures.

The methodology is not restricted to any kind of compounds neither any subset of space groups of symmetry. The algorithm random tree, with the help of data preprocessing, yields a quick classification of Bravais lattice providing a low error margin, despite not having any a priori knowledge (ab initio). The obtained accuracy was 95.5%. The performance of ANN results was found clearly inferior to other ML methods. In addition, it was shown by means of data mining methods that the amount of input data may be reduced to eight attributes instead of the usual 20 attributes demanded by the classic software applications. With this reduced set as a start point, for which the least significant data were suppressed, the obtained accuracy improves to 95.9%.

Moreover, it was proved that, with the present methodology, we can ignore the last 10 smallest interplanar lattice distances while maintaining an accuracy over 90%. Likewise, the method was shown robust versus random noise in the d_{hkl} set coming from errors in the Bragg peak hunting procedure. In addition, not only the problem has been solved with ML-based methods with high accuracy, but by selecting the best model and subjecting it to a complex and unknown case, Random Tree achieved notable success.

Finally, our trained model was able to overcome a severe test of Bravais lattice assignment in the domain of advanced ceramic materials. The success in solving this difficult problem by means of ML methods will undoubtedly contribute to rise interest in the community of crystallographers. We are currently expanding our focus of attention to other complex crystallographic systems as defective spinels and non-stoichiometric boron carbides.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Microsoft SharePoint at https://uses0-my.sharepoint.com/:u:/g/personal/fcumberras_us_es/Ec_is23q7oIHkcreWmvDEzcBu4JEXLWwSgWnyANdjjE82A?e=P9g0cT https://uses0-my.sharepoint.com/:u:/g/personal/fcumberras_us_es/EbGUPfBB0EIEq1Xq_lfs2fkB84TIWhS1W2tI9ImqjdBoEg?e=bPYHo0.

ORCID

Francisco-Luis Cumbra  <https://orcid.org/0000-0003-4666-0998>

REFERENCES

- Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22(5), 717–727. [https://doi.org/10.1016/S0731-7085\(99\)00272-1](https://doi.org/10.1016/S0731-7085(99)00272-1)
- Bergerhoff, G., Hundt, R., & Sievers, R. (1983). The inorganic crystal structure data base. *Journal of Chemical Information and Computer Sciences*, 23, 66–69.
- Bhadeshia, H. K. D. H. (1999). Neural networks in materials science. *ISIJ International*, 39(10), 966–979. <https://doi.org/10.2355/isijinternational.39.966>
- Boultif, A., & Louër, D. (2004, Oct). Powder pattern indexing with the dichotomy method. *Journal of Applied Crystallography*, 37(5), 724–731. <https://doi.org/10.1107/S0021889804014876>
- Butterworth, R., Simovici, D., Santos, G., & Ohno-Machado, L. (2004). A greedy algorithm for supervised discretization. *Journal of Biomedical Informatics*, 37(4), 285–292.
- Coelho, A. (2017). An indexing algorithm independent of peak position extraction for X-ray powder diffraction patterns. *Journal of Applied Crystallography*, 50(5), 1323–1330. <https://doi.org/10.1107/S1600576717011359>
- Cumbra, F., Sponchia, G., Benedetti, A., Riello, P., Pérez, J., & Ortiz, A. (2018). Some crystallographic considerations on the novel orthorhombic zro2 stabilized with ta doping. *Ceramics International*, 44(9), 10362–10366.
- Dai, D., Xu, T., Wei, X., Ding, G., Xu, Y., Zhang, J., & Zhang, H. (2020). Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. *Computational Materials Science*, 175, 109618. <https://doi.org/10.1016/j.commatsci.2020.109618>
- Frank, E., Hall, M., & Witten, I. (2016). *The weka workbench*. Morgan Kaufmann.
- Gnanambal, S., Thangaraj, M., Meenatchi, V., & Gayathri, V. (2018). Classification algorithms with attribute selection: An evaluation study using weka. *International Journal of Advanced Networking and Applications*, 9(6), 3640–3644.
- Habershon, S., Cheung, E., Harris, K., & Johnston, R. (2004). Powder diffraction indexing as a pattern recognition problem: A new approach for unit cell determination based on an artificial neural network. *The Journal of Physical Chemistry A*, 108(5), 711–716. <https://doi.org/10.1021/jp0310596>
- Jacobson, R. (1997). A Monte Carlo method for indexing. *Zeitschrift Fur Kristallographie*, 212(2), 99–102.
- Lee, J., Park, W., Lee, J., Singh, S., & Sohn, K. (2020). A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns. *Nature Communications*, 11(86), 1–11. <https://doi.org/10.1038/s41467-019-13749-3>
- Liang, H., Stanev, V., Kusne, A., & Takeuchi, I. (2020). Cryspnet: Crystal structure predictions via neural networks. *Physical Review Materials*, 4(12), 123802. <https://doi.org/10.1103/PhysRevMaterials.4.123802>
- Lopes, N., & Ribeiro, B. (2010). Stochastic gpu-based multithread implementation of multiple back-propagation. In *Proceedings of the 2nd international conference on agents and artificial intelligence*, vol. 1: ICAART (pp. 271–276). SciTePress. doi: <https://doi.org/10.5220/0002722102710276>
- Lopes, N., & Ribeiro, B. (2011). An evaluation of multiple feed-forward networks on gpus. *International Journal of Neural Systems*, 21(1), 31–47. <https://doi.org/10.1107/S1600576720016532>
- Oviedo, F., Ren, Z., Sun, S., Settens, C., Liu, Z., Hartono, N., Ramasamy, S., DeCost, B. L., Tian, S. I. P., Romano, G., Kusne, A. G., & Buonassisi, T. (2019). Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *NPJ Computational Materials*, 5(60), 1–9. <https://doi.org/10.1038/s41524-019-0196-x>
- Ryan, K., Lengyel, J., & Shatruk, M. (2018). Crystal structure prediction via deep learning. *Journal of the American Chemical Society*, 140(32), 10158–10168. <https://doi.org/10.1021/jacs.8b03913>
- Scott, D., Coveney, P., Kilner, J., Rossiny, J., & Alford, N. N. (2007). Prediction of the functional properties of ceramic materials from composition using artificial neural networks. *Journal of the European Ceramic Society*, 27(16), 4425–4435. <https://doi.org/10.1016/j.jeurceramsoc.2007.02.212>
- Sha, W., & Edwards, K. (2007). The use of artificial neural networks in materials science based research. *Materials & Design*, 28(6), 1747–1752. <https://doi.org/10.1016/j.matdes.2007.02.009>
- Shetty, K. R., Rao, A., & Gopala, K. (1999). A neural network approach to crystal structure classification. *Current Science*, 76(5), 670–676.
- Smeeton, N. (1985). Early history of the kappa statistic. *Biometrics*, 41(3), 795.
- Vaitkus, A., Merkys, A., & Gražulis, S. (2021). Validation of the crystallography open database using the crystallographic information framework. *Journal of Applied Crystallography*, 54(2), 661–672. <https://doi.org/10.1107/S1600576720016532>
- Visser, J. (1969). A fully automatic program for finding the unit cell from powder data. *Journal of Applied Crystallography*, 2(3), 89–95. <https://doi.org/10.1107/S0021889869006649>
- Wang, H., Xie, Y., Li, D., Deng, H., Zhao, Y., Xin, M., & Lin, J. (2020). Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *Journal of Chemical Information and Modeling*, 60(4), 2004–2011. <https://doi.org/10.1021/acs.jcim.0c00020>
- Werner, P., Eriksson, L., & Westdahl, M. (1985). TREOR, a semi-exhaustive trial-and-error powder indexing program for all symmetries. *Journal of Applied Crystallography*, 18(5), 367–370. <https://doi.org/10.1107/S0021889885010512>
- Woinaroschy, A., Isopescu, R., & Filipescu, L. (2000). X-ray patterns identification of crystallized sodiumdisilicates mixtures. *Crystal Research and Technology*, 35(8), 969–977. <https://doi.org/10.2355/isijinternational.39.966>
- Zhang, Y., Yang, S., & Evans, J. (2008). Revisiting Hume-Rothery's rules with artificial neural networks. *Acta Materialia*, 56, 1094–1105. <https://doi.org/10.1016/j.actamat.2007.10.059>

AUTHOR BIOGRAPHIES

Esther-Lydia Silva-Ramírez received the Bachelor's degree in Computer Science and Engineering from the University of Cádiz and the Engineer's and the PhD degrees in Computer Science and Engineering from the University of Sevilla. She is currently an Associate Professor in the

Department of Computer Science and Engineering at the University of Cádiz. Her major research interests include Computational Intelligence, Data Processing and Machine Learning-based techniques applied to different areas.

Inmaculada Cumbreira-Conde has a Bachelor of Business Administration and law degrees in Spain and Australia (University of Cadiz, Catholic University of Lyon, University of Technology, Sydney and University of Sydney). She is working towards a PhD with the Road to Research Scholarship - RTP and MRES Scholarship from Macquarie University. Her research interests include Project Management, AI and Machine Learning, and Private International Law. She is the Chair of Spanish Researchers Abroad in Australia-Pacific. She has broad experience in law and business and is a Seasonal Teacher at Macquarie University.

Rafael Cano-Crespo received the degree in Physics in 2012, Master's degree in Science and Technology of New Materials in 2014 and PhD in Science and Technology of New Materials from the University of Sevilla in 2018. He is currently an Assistant Professor and researcher in the Department of Condensed Matter Physics of the Faculty of Physics at the University of Sevilla. His research interests include the processing, spark plasma sintering and mechanical behavior of advanced ceramic materials (alumina, zirconia and ceramic composites) using different experimental characterization techniques such as: scanning electron microscopy, transmission electron microscopy, creep at high temperatures, Raman spectroscopy, Vickers hardness, nanoindentation.

Francisco-Luis Cumbreira received the degree in Physics from the University of Sevilla. He obtained the PhD in the same university working on the Crystallization of Metallic Glasses. He made a stay of four years in the Laboratoire de Microscopie Electronique in the University of Haute Normandie. He has made long stays in the CNRS laboratory of Bellevue (Paris) and also in the laboratory of Cristallographie et Mineralogie (Université Pierre et Marie Curie, Paris). He is currently Full Professor in the Department of Condensed Matter Physics in the University of Sevilla. His research interests include X-rays, intermetallic compounds and advanced ceramics.

How to cite this article: Silva-Ramírez, E.-L., Cumbreira-Conde, I., Cano-Crespo, R., & Cumbreira, F.-L. (2023). Machine learning techniques for the ab initio Bravais lattice determination. *Expert Systems*, 40(2), e13160. <https://doi.org/10.1111/exsy.13160>



APPENDIX A

An extract of the rules set generated by Random Tree model is shown in this section:

```

campo 13 < 0.29
| campo 2 < 0.69
| | campo 6 < 0.28
| | | campo 6 < 0.18: 6 (2/0)
| | | campo 6 ≥ 0.18
| | | | campo 6 < 0.27: 5 (6/0)
| | | | campo 6 ≥ 0.27: 4 (2/0)
| | campo 6 ≥ 0.28
| | | campo 10 < 0.3
| | | | campo 6 < 0.34
| | | | | campo 2 < 0.52
| | | | | | campo 13 < 0.22
| | | | | | | campo 10 < 0.25: 13 (11/0)
| | | | | | | campo 10 ≥ 0.25
| | | | | | | | campo 6 < 0.31: 11 (1/0)
| | | | | | | | campo 6 ≥ 0.31: 12 (1/0)
| | | | | | | | | campo 13 ≥ 0.22
| | | | | | | | | campo 2 < 0.5
| | | | | | | | | | campo 14 < 0.23: 4 (3/0)
| | | | | | | | | | campo 14 ≥ 0.23
| | | | | | | | | | | campo 8 < 0.31: 3 (3/0)
| | | | | | | | | | | campo 8 ≥ 0.31: 9 (2/0)
| | | | | | | | | | | | campo 2 ≥ 0.5: 13 (6/0)
| | | | | | | | | | | | | campo 2 ≥ 0.52
| | | | | | | | | | | | | | campo 8 < 0.31: 3 (18/0)
| | | | | | | | | | | | | | campo 8 ≥ 0.31: 13 (2/0)
| | | | | | | | | | | | | | | campo 6 ≥ 0.34
| | | | | | | | | | | | | | | | campo 2 < 0.6
| | | | | | | | | | | | | | | | | campo 10 < 0.29: 13 (22/0)
| | | | | | | | | | | | | | | | | | campo 10 ≥ 0.29: 11 (1/0)
| | | | | | | | | | | | | | | | | | | campo 2 ≥ 0.6
| | | | | | | | | | | | | | | | | | | | campo 3 < 0.51: 5 (3/0)
| | | | | | | | | | | | | | | | | | | | campo 3 ≥ 0.51: 3 (5/0)
| | | | | | | | | | | | | | | | | | | | | campo 10 ≥ 0.3
| | | | | | | | | | | | | | | | | | | | | | campo 17 < 0.25
| | | | | | | | | | | | | | | | | | | | | | | campo 3 < 0.57
| | | | | | | | | | | | | | | | | | | | | | | | campo 8 < 0.32: 13 (2/0)
| | | | | | | | | | | | | | | | | | | | | | | | | campo 8 ≥ 0.32
| | | | | | | | | | | | | | | | | | | | | | | | | | campo 2 < 0.65: 13 (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | campo 2 ≥ 0.65: 9 (4/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 3 ≥ 0.57.
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 10 < 0.31: 12 (4/0).
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 10 ≥ 0.31: 4 (2/0).
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 17 ≥ 0.25
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 3 < 0.55
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 10 < 0.33: 7 (6/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 10 ≥ 0.33: 13 (3/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | campo 3 ≥ 0.55: 13 (7/0)

```