buildings

MDPI

Article

Advancing Efficiency in Mineral Construction Materials Recycling: A Comprehensive Approach Integrating Machine Learning and X-ray Diffraction Analysis

Markus Wilhelm [1], Frank Lotter [2], Christian Scherdel [2] and Jan Schmitt [1,*]

[1] Institute of Digital Engineering, Technical University of Applied Sciences Würzburg-Schweinfurt, Ignaz-Schön-Straße 11, 97421 Schweinfurt, Germany; markus.wilhelm@thws.de
[2] Center for Applied Energy Research, Magdalene-Schoch-Straße 3, 97074 Würzburg, Germany; frank.lotter@cae-zerocarbon.de (F.L.); christian.scherdel@cae-zerocarbon.de (C.S.)
* Correspondence: jan.schmitt@thws.de

Abstract: In the context of environmental protection, the construction industry plays a key role with significant CO_2 emissions from mineral-based construction materials. Recycling these materials is crucial, but the presence of hazardous substances, i.e., in older building materials, complicates this effort. To be able to legally introduce substances into a circular economy, reliable predictions within minimal possible time are necessary. This work introduces a machine learning approach for detecting trace quantities ($\geq$0.06 wt%) of minerals, exemplified by siderite in calcium carbonate mixtures. The model, trained on 1680 X-ray powder diffraction datasets, provides dependable and fast predictions, eliminating the need for specialized expertise. While limitations exist in transferability to other mineral traces, the approach offers automation without expertise and a potential for real-world applications with minimal prediction time.

Keywords: machine learning; classification; minerals; X-ray diffraction; construction materials

Citation: Wilhelm, M.; Lotter, F.; Scherdel, C.; Schmitt, J. Advancing Efficiency in Mineral Construction Materials Recycling: A Comprehensive Approach Integrating Machine Learning and X-ray Diffraction Analysis. *Buildings* **2024**, *14*, 340. https://doi.org/10.3390/buildings14020340

Academic Editor: Syed Minhaj Saleem Kazmi

Received: 21 December 2023
Revised: 15 January 2024
Accepted: 23 January 2024
Published: 25 January 2024

1. Introduction

In the context of environmental protection, resource recycling is mandatory. The construction industry assumes a pivotal role, manifesting notable CO_2 emissions in the production of mineral-based construction materials. A vital measure to alleviate this is the recycling rate of mineral-based construction materials. Low traces of minerals in the materials to be processed can lead to undesirable problems. They can have an impact on recyclability, as the properties of the end product may change due to these traces of minerals, making it necessary to change the production processes. A complicating factor in this endeavor is the potential presence of hazardous substances in older building materials, necessitating their exclusion from the material cycle. Within large-scale industrial applications, time emerges as a critical economic factor. Additionally, legislation also requires a significant increase in the recycling rate of construction and demolition waste and aims to prevent the release of hazardous substances into the environment [1]. Strict national regulations must be consistently adhered to, especially when it comes to threshold values for hazardous substances. The economic reintegration of substances into the cycle is contingent upon the near-100% reliability of predictions, achieved within the briefest feasible timeframe. A viable methodology for discerning substances within mixtures is X-ray powder diffraction (XRD). However, such analyses are proficiently executed solely by specialists, entailing significant time investment. For this reason, methods must be developed with which even small mass fractions can be quickly recognized. This work introduces an approach to predicting trace quantities ($\geq$0.06 wt%) of minerals, employing siderite in mixtures as a paradigmatic instance. The beginning provides a brief overview of current literature in the field of minerals combined with machine learning. Additionally,

a brief overview of X-ray diffraction in the context of mineralogy is given. Subsequently, data preparation for the later machine learning model is explained. The training data are obtained from an experimental calcium carbonates dataset, utilizing pure siderite measurements from XRD databases. By clustering the database data, relevant areas of the experimental data can be identified and extracted as features. Following this, based on the insights from the literature, a machine learning model is trained to detect mineral traces. The following protocol is applied:

- Obtain pure measurements of the target mineral from the database and cluster the highest peaks.
- Utilize these peaks to define a data region from which information can be extracted for training and prediction from compound measurements.
- Train and test a machine learning model to detect traces of the target mineral.

The model thus affords a dependable and expeditious prediction of an arbitrary quantity of measurements, devoid of the prerequisite of specialized expertise. In addition to the improvements in the time and human capital factors, this avoids undesirable mineral traces in recycled material in a very short time before the production process. It allows adjustments before the production process right at the start of the recycling material feed. This avoids rejects due to poor quality. But it also prevents hazardous impurities from being fed in.

2. Literature Review

In addition to the classical methods for recognition, approaches using machine learning and deep learning methods are increasingly emerging to recognize the characteristics of minerals and mineral compounds. These characteristics are important, for example, in the production of mineral materials for determining their mechanical properties. They are also essential when it comes to introducing recycled materials into the manufacturing process while retaining these attributes. In particular, in recent years, approaches using neural networks (NN) are predominantly used for the detection of phases, crystal structures, and other attributes of minerals and mineral compounds. The following section and Table 1 are a brief overview of the current state of the literature.

Park et al. [2] used a convolutional neural network (CNN) for the prediction of space-group, extinction-group, and crystal-system classifications. Ryan et al. [3] chose a deep neural network (DNN) approach to predict the crystal structure. For this purpose, chemical elements are detected using a DNN, and subsequently, the most similar crystal structure from a database is identified through another DNN. Utimula et al. [4] used a dynamic time warping (DTW) model to disregard unimportant information such as intensity and to counteract a possible peak shift, in order to identify the concentration of substituents in magnetic alloys, due to the fact that the classification is based on the absolute distance between the peaks. Vecsei et al. [5] used a DNN for inorganic powder XRD to determine space groups and crystal systems. In addition, they used the standard distribution and additional noise for data augmentation. A comparison was made between a DNN and a CNN. Oviedo et al. [6] implemented an all-convolutional neural network (a-CNN) in combination with "class activation maps", trained with data from the ICSD database and applied to experimental data to classify the space groups and crystal dimensionality of metal-alloy XRD spectra. They also augmented data to achieve better results, including the use of a Savitzky–Golay filter, but also peak scaling, peak elimination, and pattern shifting. They also compared the results of different machine learning models like naive Bayes, k-nearest neighbors (kNN), logistic regression, random forest (RF), decision trees (DT), support vector machine (SVM), gradient boosting regression trees, a fully connected DNN, and an a-CNN with a global pooling layer and normalized DTW combined with a kNN classification. The best results were achieved by the a-CNN and the DNN. Lee et al. [7] applied different CNN architectures to enable phase identification. In addition, they compared these networks with the classical machine learning methods kNN, RF, and SVM. RF was shown to be the most suitable method for prediction in comparison. Wang et al. [8] also trained a CNN

on theoretical and limited experimental data for the identification of experimental XRD patterns of metal–organic frameworks. Dong et al. [9] presented a CNN approach to predict scale-factor, lattice-parameter, and crystallite-size maps for all phases as a real-time full profile analysis. Szymanski et al. [10] also used a CNN for predicting complex multiphase mixtures. They showed a branching algorithm that iteratively predicts the contents of the sample and at the end, the branch with the highest average probability is chosen. Szymanski et al. [11] presented an approach to a CNN-guided XRD measurement that autonomously identifies the phases. For this purpose, the unpredictable range is iteratively restricted and measured with higher resolution. This is to make peaks predictable, which for example are lost due to noise in a faster measurement. In addition to the approaches using neural networks, there are also approaches using machine-learning algorithms. Bunn et al. [12] showed an approach to identify different phases using ADAboost, which was extended by Bunn et al. [13] using k-means clustering and an expert labeling data. As already mentioned, Lee et al. [7] showed an additional approach that uses, among various other methods, an RF. Yanxon et al. [14] used kNN, extra tree, gradient boosting, and RF for the detection of single-crystal diffraction spots in XRD images so as to enable precise analyses of 1D powder diffraction patterns. Again, RF proved to be the most suitable method.

Table 1. Literature overview.

Authors	Year	Algorithm	Target
Bunn et al. [12]	2015	ADA boost	Phase identification
Bunn et al. [13]	2016	k-means, experts	Phase identification
Park et al. [2]	2017	CNN	Space-group, extinction-group, crystal-system
Ryan et al. [3]	2018	DNN	Crystal structure
Vecsei et al. [5]	2019	DNN, CNN	Space-group, crystal-system
Oviedo et al. [6]	2019	Naive Bayes, kNN, logistic regression, RF, DT, SVM, gradient boosting regression trees, fully connected DNN, a-CNN, DTW + kNN classification.	Space-group, crystal dimensionality
Utimula et al. [4]	2020	DTW	Concentration of substituents
Lee et al. [7]	2020	CNN, kNN, RF, SVM	Phase identification
Wang et al. [8]	2020	CNN	Identification of XRD patterns
Dong et al. [9]	2021	CNN	Scale factor, lattice parameter, crystallite size maps
Szymanski et al. [10]	2021	CNN, branching algorithm	Phase identification
Szymanski et al. [11]	2023	CNN-guided measurement, RF	Phase identification
Yanxon et al. [14]	2023	kNN, extra tree, gradient boosting, RF	Single-crystal diffraction spots

Even if the majority of current works use a neural network, it must be viewed critically, especially with regard to an industrial approach. Here, regulations apply from the legislator, who wants proof on demand. The traceability of the prediction process is not explicitly addressed in the listed works. There is also a lack of explicit information on the detection limits, which are a prerequisite for industrial use. For example, Dong et al. [9] stated a limit value of less than <1 wt% PdO in very specific sample configurations. Bunn et al. [13] even utilized a threshold of 5% of the maximum intensity value for feature extraction. In Yanxon et al. [14], the weight proportions of the minor components were between 2% and 20% by mass for two-phase mixtures. Classical methods such as RF have an advantage, as they offer traceability and explainability in the decisions. As an example, it is possible to extract the specific probabilities for the presence of a substance, which is important when this substance has low concentration limits. Further analytical methods can also be used if higher probabilities are shown for a substance. Lastly, it should be mentioned that only some of the work enables the detection of individual components in mixtures. However, this is a very important point for commercial use, as it would not be realistic to assume absolute purity.

3. X-ray Diffraction in Mineralogy

X-ray diffraction (XRD) is based on the interaction of X-rays with crystalline materials. The X-rays are diffracted due to the periodic arrangement of the atoms within the crystal lattice. The diffraction results in a distinct pattern (diffractogram), which can be translated

into the crystallographic structure of the material. Bragg's law (Equation (1)) describes the fundamental theory of the method:

$$n\lambda = 2d \sin\theta \quad (1)$$

with λ being the X-ray wavelength, d the lattice constant, and θ the diffraction angle.

Especially in mineralogy, utilizing monochromatic X-rays for XRD is one of the most common methods. It allows extracting the exact crystal structure (via the lattice parameter d) by analyzing the positions (the angle of diffraction) and intensities of the diffraction peaks. This knowledge is fundamental to understanding the physical and chemical properties of the investigated minerals.

The work concentrates on identifying siderite traces in different carbonate compounds. Apart from siderite ($FeCO_3$), the compounds consist of calcite ($CaCO_3$), high-Mg calcite ($Ca,MgCO_3$), vaterite ($CaCO_3$), smithsonite ($ZnCO_3$), siderite ($FeCO_3$), rhodochrosite ($MnCO_3$), dolomite ($MgCO_3$), monohydrocalcite ($CaCO_3 \cdot H_2O$), and otavite ($CdCO_3$). The crystal structures of siderite and the other carbonate minerals are shown in Table 2.

Table 2. Lattice parameters of carbonate minerals [15].

Mineral	Crystal System	Unit Cell
Siderite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $R\bar{3}c$	a, b = 4.67 Å; c = 15.34 Å
Calcite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $R\bar{3}c$	a, b = 4.99 Å; c = 17.07 Å
High-Mg calcite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $R\bar{3}c$	a, b = 4.94 Å; c = 16.85 Å
Vaterite	hexagonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $P6_3/mmc$	a, b = 4.13 Å; c = 8.49 Å
Smithonite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $R\bar{3}c$	a, b = 4.65 Å; c = 14.50 Å
Rhodochrosite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $R\bar{3}c$	a, b = 4.77 Å; c = 15.63 Å
Dolomite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $R\bar{3}$	a, b = 4.81 Å; c = 16.01 Å
Monohydrocalcite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $P3_121$	a, b = 10.55 Å; c = 7.54 Å
Otavite	trigonal	$\alpha, \beta = 90°; \gamma = 120°$
	space group $R\bar{3}$	a, b = 4.93 Å; c = 16.27 Å

To differentiate between phases in a material compound, typically the diffractogram is visually analyzed. The intensity of the corresponding peaks is proportional to the concentration of the respective phase. This visual phase identification is typically possible down to phase concentrations of 1–0.5%, depending on the experimental setup (instrument resolution, integration time, background noise, etc.). Here, the ML-assisted analysis comes into play, where even the slightest anomalies in the data signature can be recognized, which leads to a lower detection limit.

4. Methods and Data

This section aims to delineate the methodological approach for detecting traces of minerals within a mixture. For this approach, pure substance measurements are selected from databases, and their highest peaks are determined, from which search ranges are derived. Subsequently, these search ranges are utilized to delimit the regions containing the requisite information for prediction in mixture measurements and for model training. Following this, the results are verified, and the detection limit is established using this model. This procedure is structured into three segments: data preparation, implementing the machine learning model, and validating the outcomes.

4.1. Data Preparation

The following section describes the data preparation process. It can be divided into three main steps. The first step is to extract the three highest peaks of pure substance measurements of the searched substance from databases. In the next step, the peaks are

clustered to obtain characteristic starting values, which should indicate the position of the substance in the compound measurements. The final third step is the feature extraction from an experimental dataset of compounds, which can be found in the literature.

For this purpose, the 2-theta value of the centroids is taken and extended to a surrounding area from which the intensity data of the curves are taken. These data are later used as training features for the model. As initially mentioned, in the first step, pure substance measurements of siderite are obtained from XRD databases—in this case, from the freely accessible RRUFF database [15]. There are 34 measurements available for determining the highest peaks of pure siderite. The data consist of dimensionless intensities and their corresponding 2-theta values in degrees. In reality, slight deviations from the ideal crystal structure can occur due to, e.g., moisture content, higher pressure, lattice parameters, impurities, instrumental errors, or sample-displacement errors. Therefore, the siderite peaks are clustered due to these possible peak shifts. The resulting centroids serve as a center-point to define the areas of the peaks that are characteristic of siderite. The elbow method is commonly used for determining the necessary number of clusters [16,17]. The clustering method used for this is k-means. Two inflection points can be identified at positions 2 and 3 of the centroids (Figure 1), which are then usually assumed to be a suitable number of clusters.

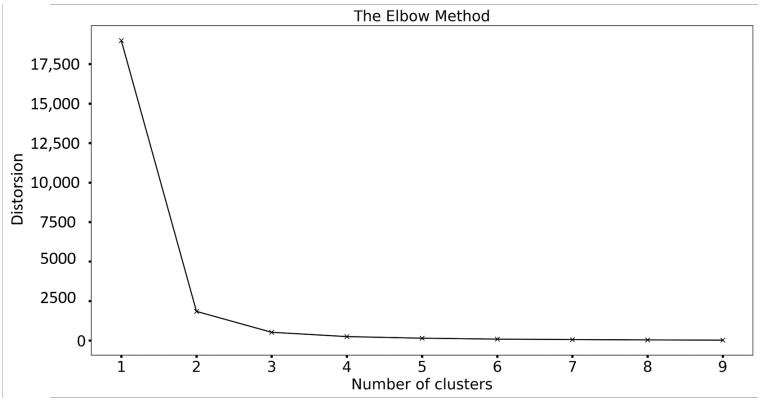

Figure 1. The elbow method, which shows a suitable number of clusters based on the inflection points.

For the subsequent procedure, three centroids of the pure siderite-peak clusters are chosen as initial values. Due to potential overlaps of peaks in composite measurements, multiple characteristic positions indicating the substance are advantageous in the measurement. Two scenarios are shown below. In the first scenario, all database data related to siderite are used. This approach is realistic if the data are retrieved via an API, for example, without the measurement conditions being subsequently checked by an expert. This also corresponds to one of the objectives of the work. In the second scenario, only database data from siderite measurements under normal conditions are used. This serves in particular to validate the results and to rule out the possibility of errors in connection with a possible shift. The following values define the 2θ-interval of each cluster: [24.76°, 27.68°], [31.98°, 37.44°], and [52.81°, 61.74°]. The significant peak shifts within the mentioned ranges can be explained by the presence of data in the database, such as measurements taken under high-pressure conditions [18]. This is a positive side effect when all database data are taken into account, as such anomalous data are also included in cases where, for example, an automated database query is performed or no experts in the field are involved in the process. This approach increases the resilience of the system when it comes to dealing with problems of this kind.

The centroids of the siderite-peak clusters are represented by the values 26.38°, 35.08°, and 57.84°. To determine the subsequent range from which data will be extracted for training and prediction, $\pm 1°$ around the centroids is considered (Figure 2). This results in ranges of [25.38°, 27.38°], [34.08°, 36.08°], and [56.84°, 58.84°]. For training and validation, the dataset from Amao et al. [19] is utilized. It comprises 1,680 measurements with precisely

defined percentages of various calcium carbonate compounds. These composites can contain up to nine components, which are calcite ($CaCO_3$), high-Mg calcite ($CaCO_3$), vaterite ($CaCO_3$), smithsonite ($ZnCO_3$), siderite ($FeCO_3$), rhodochrosite ($MnCO_3$), dolomite ($MgCO_3$), monohydrocalcite ($CaCO_3 \cdot H_2O$), and otavite ($CdCO_3$).

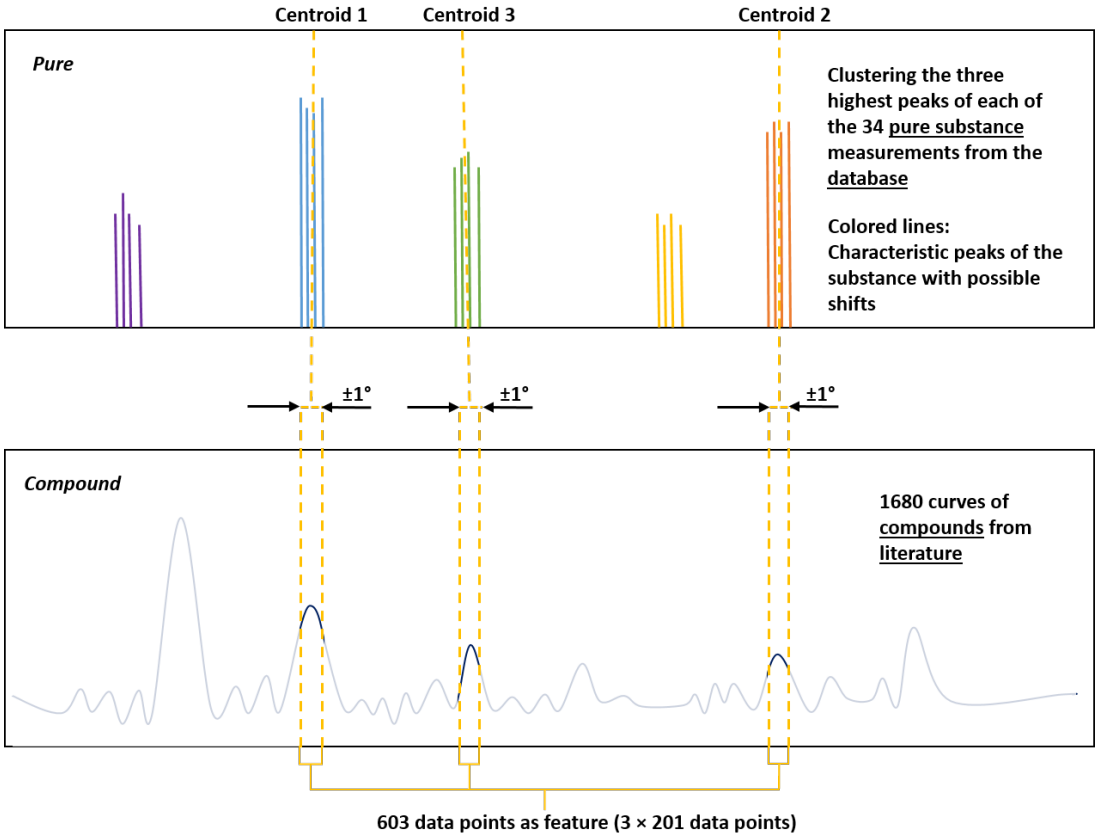

Figure 2. Visualization of the centroids that define the search area for the substance of interest; the resulting data points serve as features for subsequent training.

For the presented approach, siderite was chosen based on its maximum and minimum proportions in the mixtures, reaching a maximum of 0.79 wt%, a minimum of 0.00 wt%, and a mean value of 0.07 wt%. In total, there were 1238 samples out of 1680 that included siderite. Among them, 1672 composites had less than 0.5 wt%, which is generally accepted for human recognition by experienced individuals in a diagram. Due to a 2-theta range from 2° to 70° and a 0.01° step size in the measurements, this range contained 6603 data points. In total, 603 data points from each curve were used for training and validation, with 201 points for each range around the centroid. Only the intensity values were considered as training features, while the 2-theta values were used solely to define the range in which the information to be trained lies (Figure 3).

4.2. Machine Learning Model

The process used for training and testing the model is outlined below (Figure 4). All libraries and parameters employed are listed. Scikit [20] was utilized for machine learning. The intensities of the 603 data points of each curve were used as the features for training. Each curve consisted of a total of 6603 values. Each array containing 603 values was assigned a binary label: 1 if siderite is present and 0 if not. Subsequently, training and validation were conducted with these labeled data. A train–test split of 0.3 was used, resulting in sets of 1176 test and 504 training data pieces. Based on the results from [6,7,14], RF was chosen as the prediction method. This decision was based on the insights from the literature described above, which identified RF, alongside NN, as the most suitable

classifier. NN were not considered due to their lack of transparency, which makes them less applicable for providing evidence, for example, to regulators.

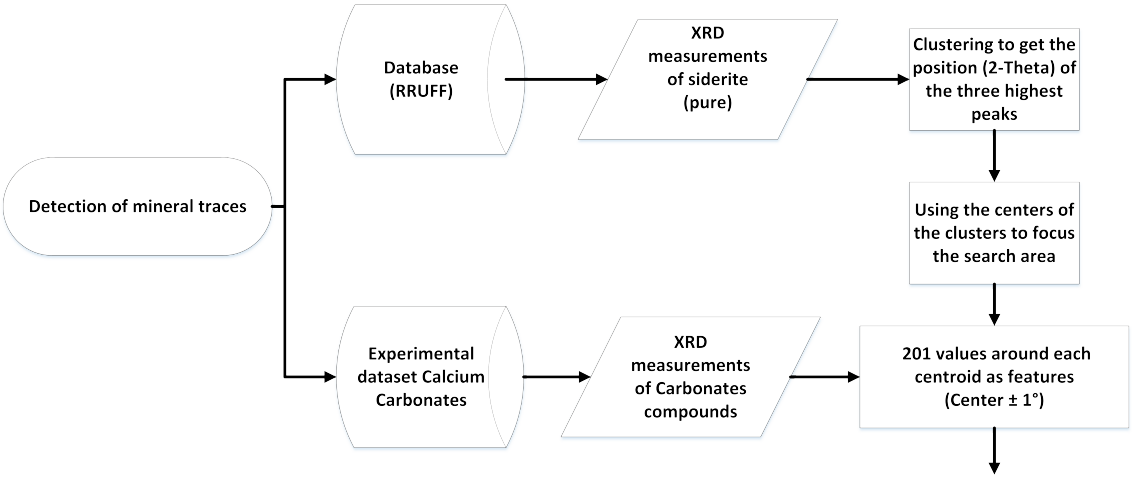

Figure 3. Process flowchart depicting data preparation.

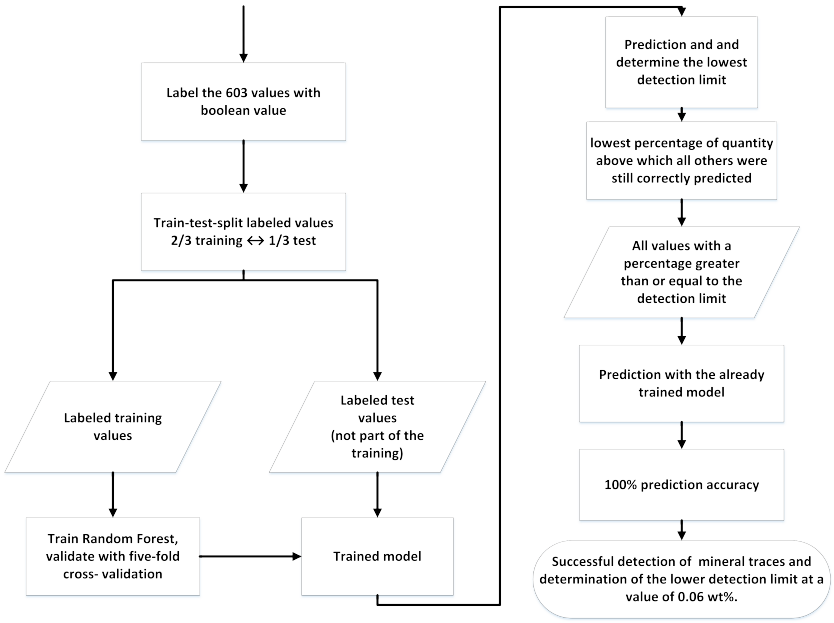

Figure 4. Continuation of the process flowchart (see Figure 3) involving the training and validation of the machine learning model using an experimental dataset.

The five-fold cross-validation method was employed for validation. A randomized search was conducted to find fitting parameters. The following parameters were tested (Table 3):

Table 3. Hyper-parameters used for the randomized search.

Hyper-Parameter			
n_estimators	start = 10	end = 2000	num = 30
max_features	sqrt		
max_depth	start = 10	end = 200	num = 20
min_sample_leaf	1	2	4
bootstrap	True	False	
criterion	gini	entropy	

Additionally to the hyper-parameters, the following settings were utilized for the randomized search: n_iter = 10, cv = 5, verbose = 20, random_state = 42, and n_jobs = −1.

Among these, the following parameter combination performed best: bootstrap = false, criterion = gini, max_depth = 170, min_samples_leaf = 1, n_estimators = 421. With this combination, a training accuracy of 83% could be achieved. The trained model was then utilized to make predictions on the test data. This was evaluated using 504 datasets with 378 samples containing siderite and 126 samples without siderite, achieving an accuracy of 81% (Table 4).

$$\text{precision} = \frac{\text{true positives (tp)}}{(\text{tp} + \text{false positives (fp)})} \tag{2}$$

$$\text{recall} = \frac{\text{tp}}{(\text{tp} + \text{false negatives (fn)})} \tag{3}$$

$$\text{F}_1\text{-score} = 2\left(\frac{\text{recall} \cdot \text{precision}}{\text{recall} \cdot \text{precision}}\right) \tag{4}$$

Table 4. Classification report .

	Precision	Recall	F_1-Score	Support
Devoid of siderite	0.72	0.37	0.49	126
Containing siderite	0.82	0.95	0.88	378
Accuracy			0.81	504
Macro average	0.77	0.66	0.69	504
Weighted average	0.80	0.81	0.78	504

This can be further detailed in a confusion matrix. The confusion matrix reveals that out of a total of 504 measurements, 439 were classified in the class containing siderite, and 55 were classified in the class without siderite. Among these, 407 values were correctly predicted. These correct predictions can be further categorized into 360 true positives and 47 true negatives (Figure 5). Upon examining the scatter plot (Figure 6), it becomes evident that the incorrectly predicted mixtures are situated below a threshold of siderite content in the mixture. Upon closer examination of these data points, it is evident that no mixture with a siderite content above 0.06 wt% was incorrectly predicted.

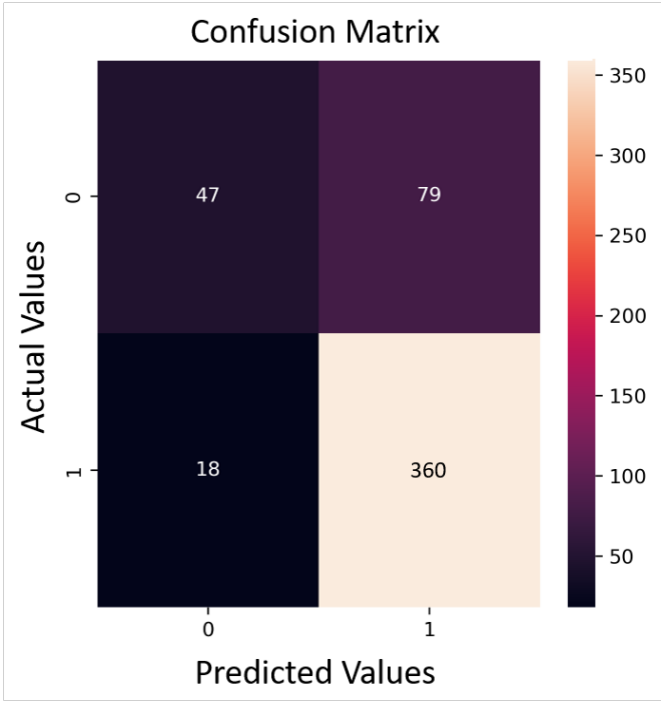

Figure 5. Confusion matrix, with the labels '1' for compounds containing siderite and '0' for compounds without siderite.

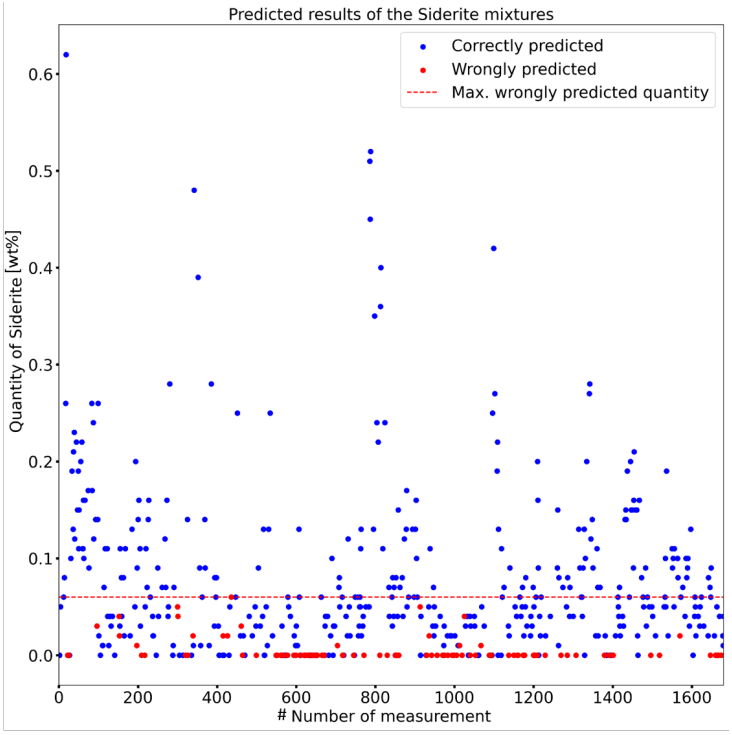

Figure 6. Scatter plot with the prediction of the test data. The ordinate shows the amount of siderite in the predicted compound, and the abscissa is the index of the predicted test data.

As already described, a second scenario was carried out with the siderite database data, which only takes into account measurements that were carried out under normal conditions. There were only three datasets remaining, but this should not have posed any problems, as there was now minimal deviation expected between the datasets. The centroids were therefore at 24.80°, 32.07°, and 52.94°. Consequently, the ranges ±1° around the centroids were determined at [23.80°, 25.80°], [31.07°, 33.07°], and [51.94°, 53.94°]. The data were extracted from the dataset in the same way as described before. For a better overview of the procedure, the intervals are shown as examples in two measurements (Figure 7). These include one measurement with 0.79 wt% (#364) and another with 0.06 wt% (#1219), which corresponds to the highest mass fraction of siderite in all compounds and the limit reached in this approach. In addition, the highest siderite peak (012) in the measurement #364 is shown enlarged. This represents the most recognizable siderite peak in all XRD patterns because of the highest siderite wt% of all compounds.

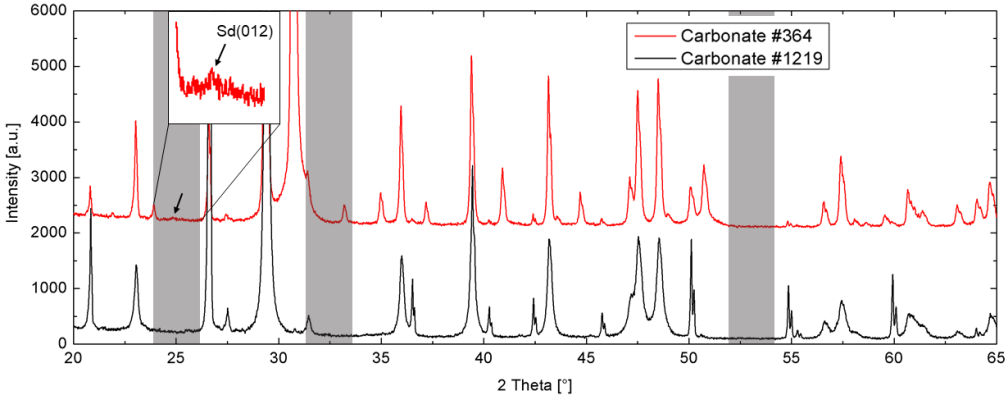

Figure 7. XRD patterns of compound #364 (0.79 wt% siderite) and #1219 (0.06 wt% siderite), with an enlarged view of the (012)-siderite peak in #364. The grey bands depict the data ranges used for the machine learning analysis.

Both the train–test split and the hyper-parameters for the randomized search remained the same. The following parameter combination delivered the best results: bootstrap = false, criterion = gini, max_depth = 52, min_samples_leaf = 1, n_estimators = 673, and a training accuracy of 80%. Even though others may have found a different parameter combination, a nearly identical result was achieved. Predictions were again made for 504 measurements, this time with 118 measurements without siderite in the mixture and 386 with siderite. There was even a slight increase in accuracy to 83%. The corresponding classification report can be found below (Table 5).

Table 5. Classification report of the second scenario.

	Precision	Recall	F_1-Score	Support
Devoid of siderite	0.92	0.29	0.44	118
Containing siderite	0.82	0.99	0.90	386
Accuracy			0.83	504
Macro average	0.87	0.64	0.67	504
Weighted average	0.84	0.83	0.79	504

As can be seen from the adjustments to the input data for the centroids and the resulting shift in the ranges, there has been a change in the predictions. The lowest misclassification is 0.03 wt% by weight (Figure 8), albeit accompanied by a deterioration in the false positive rate (Figure 9).

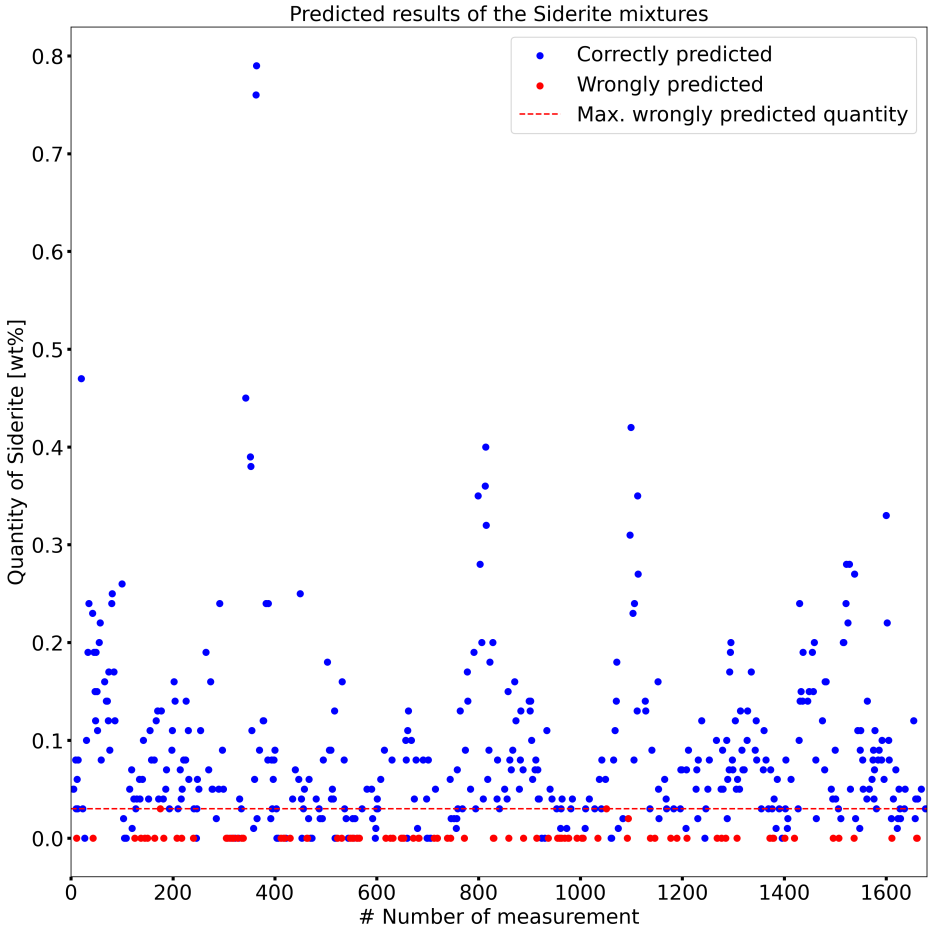

Figure 8. Scatter plot for the second run depicting predictions on test data (y-axis = siderite content in wt%, x-axis = # of the samples).

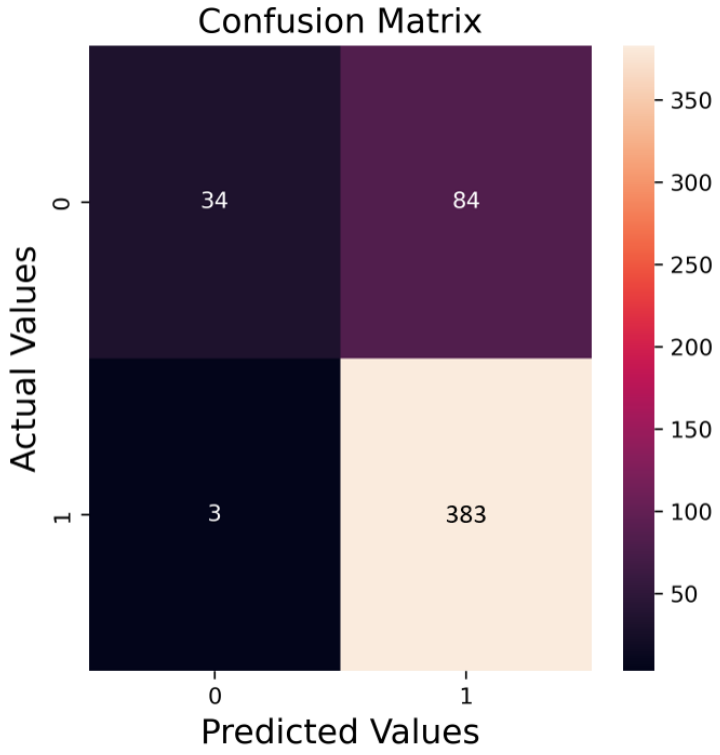

Figure 9. Confusion matrix for the second run (1 with siderite and 0 without siderite).

5. Discussion and Further Research

Looking at the results achieved with the described machine learning approach for detecting mineral traces, it becomes evident that the threshold for nearly complete detection of low traces of minerals in a mixture is plausible. Dong et al. [9] were also able to detect small phases of low-mass fractions (<1 wt% PdO) with their method under specific conditions, but this also required a CNN. Correct classifications below the threshold were possible in the used training dataset as well as when excluding siderite content. However, some decrease in accuracy must be accepted. Even though a lower value was achieved in the second iteration, the higher initial value is considered crucial. This is because in order to filter out measurements in databases, more specific knowledge is required. Since one of the objectives is to eliminate the need for a dedicated expert, it must be assumed that all database data, for instance, would be retrieved from an API. The RF approach, which has been successfully used in various scenarios in the literature for detecting properties of inorganic compounds, has proven to be suitable in this application scenario. Lee et al. [7] showed that RF achieved the best results after NN. A modification of the method could indeed offer significant potential for improving detection. It is important to note that this work presents just one approach. Current literature primarily relies on various forms of NNs, yielding good results [2,3,5,7–10]. However, there are situations where they cannot always be applied, such as when legal requirements demand evidence of how decisions were made. This proof cannot be provided with neural networks. In such cases, methods like random forests clearly have an advantage. The limitations of this work lie particularly in the transferability to other mineral traces. Further studies would be needed to explore such issues and possible solutions, such as increasing the number of considered areas or expanding their range. An important step in identifying these necessary adjustments is the identification of siderite in other mixtures, especially in minerals with a high degree of similarity or overlap of peaks. Such difficult conditions should highlight the potential to improve the model and thus lead to an increase in reliability. Another point that should be researched in the future is the detection of a variety of minerals or other substances simultaneously. All samples could then be automatically analyzed for every substance to create the individual

fingerprint of each sample. Automation, whether with one or multiple sought substances, is one of the advantages of this approach over conventional methods and could be used in the future during operation without the need for expert knowledge. Improvement through continuously gained data during usage is also conceivable. Further research should focus on requiring significantly fewer pieces of training data, simplifying real-world applications. Another advantage of this approach is that after the time-consuming training before actual use, the actual prediction time becomes negligibly small.

6. Conclusions

In this work, a machine learning approach is presented, which can be used to detect traces of minerals in mixtures analyzed using XRD. For this purpose, pure substance measurements of the target substance are used from relevant database data. The three highest points of each measurement are extracted and clustered. This results in three centroids, allowing the definition of a central point in the range where a peak shift seems possible. Expanding this point to a range of $\pm 1°$ around the centroid enables capturing the surrounding area and its characteristic curve even with a peak shift. These ranges were extracted from 1680 measurements of carbonate compounds [19], providing 603 data points as features for training and testing. This high number of data points allows for presenting the curve as a feature in the area of interest in a very granular way, making it possible to capture even small changes in the curve. This is particularly important for small admixture proportions since, naturally, with the chosen method of analysis, only minor fluctuations in the curves are to be expected. Based on relevant literature, a RF was chosen as the machine learning method. An accuracy of 83% was achieved in training. Randomized search was used for parameter tuning, and a five-fold cross-validation was employed for validation. With a train–test split of 0.3, an accuracy of 81% was achieved in predicting test data. Upon closer analysis, it was observed that no prediction errors occurred above a threshold of 0.06 wt%. This threshold is significantly below the common assumption of 0.5 wt%, where trained users typically recognize specifically sought substances in a curve. In addition, a second iteration was performed using only siderite data from measurements under normal conditions to obtain the centroids. As a result, no peak shift is expected within the data and the centers are almost identical. The result was a slight improvement to 0.03 wt%, although with an increase in the false positive rate. The work can be summarized in the following points:

- Objective: Develop a machine learning approach for detecting mineral traces in XRD-analyzed mixtures.
- Data source: Use of pure substance measurements for the searched mineral (siderite) from a relevant database and 1680 measurements of carbonate compounds for training and testing.
- Feature Extraction: Extract the highest points from each pure substance measurement and cluster them, resulting in centroids. The intensities extracted from 1680 measurements around the mean 2-theta values serve as characteristics for training and testing.
- Machine learning method: Random forest (RF) was chosen, achieving an 83% accuracy in training. A randomized search was employed for parameter tuning, and five-fold cross-validation was used for validation.
- Results: An 81% accuracy was achieved in predicting test data. No prediction errors were observed above 0.06 wt% of siderite content in the compounds.

This approach provides a method for identifying mineral impurities in mineral mixtures. The ability to do this quickly and without extensive expertise would be a crucial aspect in the production of mineral-based building materials. It enables rapid testing of incoming batches of recycled material and facilitates quick decisions to either reject them due to hazardous substances, for example, or to react appropriately to specific impurities if they have the potential to affect the properties of final material.

Author Contributions: Conceptualization, M.W. and J.S.; methodology, M.W. and J.S.; software, M.W.; validation, M.W.; formal analysis, M.W. and F.L.; investigation, M.W.; resources, F.L. and C.S.; data curation, M.W. and F.L.; writing—original draft preparation, M.W., F.L., C.S. and J.S.; writing—review and editing, M.W., F.L., C.S. and J.S.; visualization, M.W.; supervision, J.S.; project administration, J.S.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Bavarian State Ministry of Environment and Consumer Protection (StMUV) BAF01SoFo-80566 and coordinated by the Center for Basic Materials Efficiency (REZ) at the Bavarian Environment Agency (LfU). Supported by the publication fund of the Technical University of Applied Sciences Wuerzburg-Schweinfurt.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.1016/j.dib.2022.108204.

Acknowledgments: We are grateful to Hartwig Frimmel from the Institute of Geography and Geology at the University of Wuerzburg for providing the mineral samples and Knauf Gips KG for providing the Gypsum samples.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

a-CNN	all-Convolutional Neural Network
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DT	Decision Tree
DTW	Dynamic Time Warping
fn	false negatives
fp	false positives
kNN	k-Nearest Neighbor
NN	Neural Network
RF	Random Forest
SVM	Support Vector Machine
tp	true positives
wt%	mass fraction in %
XRD	X-ray powder diffraction

References

1. Papamichael, I.; Voukkali, I.; Loizia, P.; Zorpas, A.A. Construction and demolition waste framework of circular economy: A mini review. *Waste Manag. Res.* **2023**, *41*, 1728–1740. [CrossRef] [PubMed]
2. Park, W.B.; Chung, J.; Jung, J.; Sohn, K.; Singh, S.P.; Pyo, M.; Shin, N.; Sohn, K.S. Classification of crystal structure using a convolutional neural network. *IUCrJ* **2017**, *4*, 486–494. [CrossRef] [PubMed]
3. Ryan, K.; Lengyel, J.; Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158–10168. [CrossRef] [PubMed]
4. Utimula, K.; Hunkao, R.; Yano, M.; Kimoto, H.; Hongo, K.; Kawaguchi, S.; Suwanna, S.; Maezono, R. Machine-Learning Clustering Technique Applied to Powder X-Ray Diffraction Patterns to Distinguish Compositions of ThMn12-Type Alloys. *Adv. Theory Simul.* **2020**, *3*, 2000039. [CrossRef]
5. Vecsei, P.M.; Choo, K.; Chang, J.; Neupert, T. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Phys. Rev. B* **2019**, *99*, 245120. [CrossRef]
6. Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N.T.P.; Ramasamy, S.; DeCost, B.L.; Tian, S.I.; Romano, G.; et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *NPJ Comput. Mater.* **2019**, *5*, 60. [CrossRef]
7. Lee, J.W.; Park, W.B.; Lee, J.H.; Singh, S.P.; Sohn, K.S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **2020**, *11*, 86. [CrossRef] [PubMed]
8. Wang, H.; Xie, Y.; Li, D.; Deng, H.; Zhao, Y.; Xin, M.; Lin, J. Rapid identification of X-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *J. Chem. Inf. Model.* **2020**, *60*, 2004–2011. [CrossRef]
9. Dong, H.; Butler, K.T.; Matras, D.; Price, S.W.; Odarchenko, Y.; Khatry, R.; Thompson, A.; Middelkoop, V.; Jacques, S.D.; Beale, A.M.; et al. A deep convolutional neural network for real-time full profile analysis of big powder diffraction data. *NPJ Comput. Mater.* **2021**, *7*, 74. [CrossRef]

10. Szymanski, N.J.; Bartel, C.J.; Zeng, Y.; Tu, Q.; Ceder, G. Probabilistic deep learning approach to automate the interpretation of multi-phase diffraction spectra. *Chem. Mater.* **2021**, *33*, 4204–4215. [CrossRef]
11. Szymanski, N.J.; Bartel, C.J.; Zeng, Y.; Diallo, M.; Kim, H.; Ceder, G. Adaptively driven X-ray diffraction guided by machine learning for autonomous phase identification. *NPJ Comput. Mater.* **2023**, *9*, 31. [CrossRef]
12. Bunn, J.K.; Han, S.; Zhang, Y.; Tong, Y.; Hu, J.; Hattrick-Simpers, J.R. Generalized machine learning technique for automatic phase attribution in time variant high-throughput experimental studies. *J. Mater. Res.* **2015**, *30*, 879–889. [CrossRef]
13. Bunn, J.K.; Hu, J.; Hattrick-Simpers, J.R. Semi-supervised approach to phase identification from combinatorial sample diffraction patterns. *Jom* **2016**, *68*, 2116–2125. [CrossRef]
14. Yanxon, H.; Weng, J.; Parraga, H.; Xu, W.; Ruett, U.; Schwarz, N. Artifact identification in X-ray diffraction data using machine learning methods. *J. Synchrotron Radiat.* **2023**, *30*, 137–146 . [CrossRef] [PubMed]
15. Lafuente, B.; Downs, R.T.; Yang, H.; Stone, N. The power of databases: The RRUFF project. *Highlights Mineral. Crystallogr.* **2015**, *1*, 25. [CrossRef]
16. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
17. Yuan, C.; Yang, H. Research on K-value selection method of K-means clustering algorithm. *J* **2019**, *2*, 226–235. [CrossRef]
18. Lavina, B.; Dera, P.; Downs, R.T.; Yang, W.; Sinogeikin, S.; Meng, Y.; Shen, G.; Schiferl, D. Structure of siderite $FeCO_3$ to 56 GPa and hysteresis of its spin-pairing transition. *Phys. Rev. B* **2010**, *82*, 064110. [CrossRef]
19. Amao, A.O.; Al-Otaibi, B.; Al-Ramadan, K. High-resolution X–ray diffraction datasets: Carbonates. *Data Brief* **2022**, *42*, 108204. [CrossRef] [PubMed]
20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.