

1 **Volcanic ash classification through Machine Learning**

2 **Damià Benet^{1,2,3,†}, Fidel Costa¹, Christina Widiwijayanti²**

3 ¹Institut de Physique du Globe de Paris, Université Paris Cité, CNRS, Paris, France.

4 ²EOS, Earth Observatory of Singapore, Nanyang Technological University, Singapore.

5 ³ Asian School of the Environment, Nanyang Technological University, Singapore.

6 †Corresponding author: Damià Benet (dbenet@ipgp.fr)

7 **Key Points:**

- 8
- Volcanic ash particles are classified through machine learning algorithms into juvenile, lithic, free-crystal and altered material types
- 9
- Discriminant features per each particle type are revealed by the Shapley values of XGBoost's predictions
- 10
- Classification by a Vision Transformer model is very accurate and could be used by
- 11
- volcano observatories
- 12
- 13
- 14

15 Abstract

16 Volcanic ash provides information that can help understanding the evolution of volcanic
17 activity during the early stages of a crisis, and possible transitions towards different eruptive
18 styles. Ash consists of particles from a range of origins in the volcanic system and its analysis
19 can be indicative of the processes driving activity. However, classifying ash particles into
20 different types is not straightforward. Diagnostic observations for particle classification are not
21 standardized and vary across samples. Here we explore the use of machine learning (ML) to
22 improve the classification accuracy and reproducibility. We use a curated database of ash
23 particles (VolcAshDB) to optimize and train two ML-based models: an Extreme Gradient
24 Boosting (XGBoost) that uses the measured physical attributes of the particles, from which
25 predictions are interpreted by the SHAP method, and a Vision Transformer (ViT) that classifies
26 binocular, multi-focused, particle images. We find that the XGBoost has an overall
27 classification accuracy of 0.77 (*macro F1-score*), and specific features of color (*hue_mean*)
28 and texture (*correlation*) are the most discriminant between particle types. Classification using
29 the particle images and the ViT is more accurate (*macro F1-score* of 0.93), with performances
30 across eruptive styles from 0.85 in dome explosion, to 0.95 for phreatic and subplinian events.
31 Notwithstanding the success of the classification algorithms, the used training dataset is limited
32 in number of particles, ranges of eruptive styles, and volcanoes. Thus, the algorithms should be
33 tested further with additional samples, and it is likely that classification for a given volcano is
34 more accurate than between volcanoes.

35 1 Introduction

36 A central challenge in volcanology is to anticipate the likely evolution of a restless
37 volcano at a given point in time (Bebbington & Jenkins, 2019). During a period of unrest, small
38 explosions or phreatic events may precede larger ones, or the volcano may remain at low
39 activity levels and go back to dormancy (Marzocchi et al., 2012; Moran et al., 2011; Tilling,
40 2008). Moreover, many eruptions consist of various phases, changing or alternating between
41 explosive to effusive eruptive styles over time. To evaluate whether a volcano will progress
42 towards one type of activity or another, an array of geophysical and geochemical tools is used
43 to monitor and interpret the processes happening underneath the volcano (Newhall &
44 Punongbayan, 1996). However, interpretation may not be straightforward and available data
45 limited, and thus diagnosis is typically quite uncertain (Tilling, 2008).

46 An additional tool that can provide critical insights on the state of a volcano is studying
47 the volcanic ash. Ash can be classified into particle types that are indicative of processes
48 driving the activity (Alvarado et al., 2016; D’Oriano et al., 2022; Gaunt et al., 2016; Pardo et
49 al., 2014). For instance, the so-called juvenile particles are associated with the ascent of magma
50 at shallow depth, and their identification, together with other monitoring signals, may warn of
51 an ensuing magmatic eruption. For example, a-posteriori studies of ash from early and small
52 phreatic eruptions of Mount St. Helens (USA, 1980) and Mount Unzen (Japan, 1991),
53 identified minor amounts of juvenile particles in these pre-climactic deposits (Cashman &
54 Hoblitt, 2004; Watanabe et al., 1999). Thus, had these been found in a timely manner, it could
55 have altered the perception for explosive potential that followed (Cashman & Hoblitt, 2004). In
56 other cases, the ambiguity of classification of the juvenile component in early explosions has
57 led to very complex management of the volcanic crises such as the 1975–1977 Soufrière
58 Guadeloupe crisis (Feuillard et al., 1983; Hincks et al., 2014; Le Guern et al., 1980).
59 Furthermore, tracking the proportions of the different components in ash, their shape, and
60 crystallinity, can give clues on possible transitions of eruption styles to better mitigate the
61 associated hazards (e.g., Benet et al., 2021; Suzuki et al., 2013).

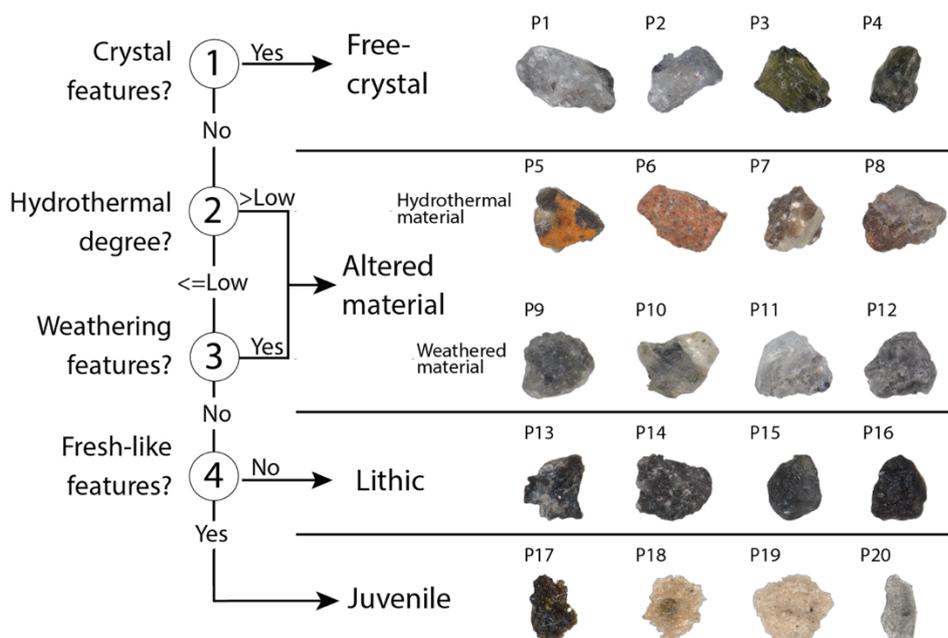
62 The classification of particles into types is typically done by collecting qualitative or
63 quantitative data on a single particle level using a variety of techniques. This includes using
64 binocular microscope (e.g., D’Oriano et al., 2014; Miwa et al., 2009; Pardo et al., 2014) to
65 observe the gloss, color and shape, as well as the particles’ surface and shape (Dellino & La
66 Volpe, 1996; Dürig et al., 2021; E. J. Liu et al., 2015; Ross et al., 2022). More detailed
67 observations including the internal microstructures are typically done using the Scanning
68 Electron Microscope (e.g., Miwa et al., 2013; Pardo et al., 2020), whereas the chemical
69 analyses are made with the electron microprobe (Pardo et al., 2014), mass spectrometers (Rowe
70 et al., 2008), and measurement of refractive indices (e.g., by the thermal immersion method;
71 Watanabe et al., 1999). However, systematic and reproducible particle classification is
72 problematic because there are few agreed diagnostic features, and these may vary from sample
73 to sample depending on the eruptive style and the volcano (e.g., Pardo et al., 2014). Whilst a
74 standardized analytical procedure of juvenile particles has been proposed (Ross et al., 2022),
75 the step of particle classification relies on observer’s experience, making it subject to varying
76 interpretations, and hindering comparison of datasets produced by different labs.

77 An approach commonly employed to address such classification challenges in various
78 domains is through the utilization of Machine Learning (ML). ML-based models can classify
79 complex images in a wide range of situations (He et al., 2015). ML-based models are capable
80 of learning patterns to classify objects, and use them for classification of future datasets, such
81 as mushrooms (Lee et al., 2022) or leaf diseases (Sujatha et al., 2021), and have already been
82 used for classification of ash particle shapes (Shoji et al., 2018). In this study, we trained two
83 models using the VolcAshDB curated dataset (Benet et al. *preprint*) with the objectives of: (i)
84 identification of the most important features for discrimination of particle types, and (ii)
85 obtaining a particle classifier as accurate as possible. The results of our study should be a step
86 forward towards a universal and unbiased classification of ash particles as more data becomes
87 available and better algorithms are developed.

88 2 Materials and Methods

89 2.1 VolcAshDB dataset

90 We used the data from the open-access database VolcAshDB, which comprises images
91 and measurements (here referred as features) of more than 6,300 volcanic ash particles
92 (<https://volcash.wovodat.org/>). These were obtained with the binocular microscope and
93 processed to obtain multi-focused, high-resolution images (Benet et al., *preprint*). The images
94 have been classified with a dichotomous key (Figure 1), using some key particle features as
95 reported in Benet et al., (*preprint*). The database contains ash particles from 12 samples from 8
96 volcanoes and 11 eruptions from a range of magma compositions and eruptive styles (Table 1).
97 These include (1) phreatic eruptions of Soufrière de Guadeloupe (Lesser Antilles) in 1976 and
98 1977 (Feuillard et al., 1983), the early activity of April 1991 of Mt. Pinatubo (Philippines;
99 Paladio-Melasantos et al., 1996), and Ontake (Japan) in 2014 (Miyagi et al., 2020), (2) dome
100 explosions of Nevados de Chillán volcanic complex (Chile) from the beginning of the eruptive
101 period in December 2016 and after the extrusion of a dome in April 2018 (Benet et al., 2021),
102 explosions from Merapi volcano (Indonesia) in July and November 2013 (Nurfiani & Bouvet
103 de Maisonneuve, 2018), (3) the basaltic lava fountaining of Cumbre Vieja (Canary Islands) in
104 October 2021 (Romero et al., 2022), and (4) two samples from different locations (KE-DB2
105 and KE-DB3) of the plinian/sub-plinian eruptions of Kelud (Indonesia) in 2014 (Maeno et al.,
106 2019; Utami et al., 2022), and a sample from the climactic plinian eruption of Mount St.
107 Helens (USA) in 1980 (Scheidegger et al., 1982).



108

109 **Figure 1.** Example of classification process and particle images in VolcAshDB based on the
 110 steps for petrographic classification in Benet et al., (*preprint*). Note that the particle type
 111 altered material comprises both hydrothermal and weathered material.

112 **Table 1.** Main sample characteristics, and proportion of main particle types in VolcAshDB.
 113 The associated error is calculated using the equation of margin of error Benet et al., (*preprint*)
 114 at a confidence interval of 95% and expressed in absolute values.

Samples	Eruption date	Magma composition	Volcano type	Eruptive style	Number of particles per component and associated error				Total
					Altered material	Free-crystal	Juvenile	Lithic	
<i>Cumbre Vieja</i>									
CV-DB1	19/10/21	Mafic	Cinder cone	Lava fountaining	3 (± 0.3)	1 (± 0.2)	719 (± 2.8)	352 (± 1.4)	1075
<i>Kelud</i>									
KE-DB2	14/2/14	Intermediate	Stratovolcano	Subplinian	50 (± 3.9)	4 (± 1.2)	268 (± 4.1)	3 (± 1.0)	325
KE-DB3	14/2/14	Intermediate	Stratovolcano	Subplinian	162 (± 5.3)	59 (± 4.0)	54 (± 3.9)	65 (± 4.2)	340
<i>Merapi</i>									
ME-DB1	22/7/13	Intermediate	Stratovolcano	Dome explosion	232 (± 4.9)	13 (± 2.2)	0	78 (± 4.7)	323
ME-DB2	22/11/13	Intermediate	Stratovolcano	Dome explosion	595 (± 2.9)	76 (± 2.1)	4 (± 0.5)	100 (± 2.4)	775
<i>Sourfière de Guadeloupe</i>									
SG-DB1	8/7/76	Intermediate	Stratovolcano	Phreatic	222 (± 5.1)	54 (± 3.9)	0	66 (± 4.2)	342
SG-DB2	1/3/77	Intermediate	Stratovolcano	Phreatic	134 (± 3.8)	8 (± 3.8)	0	0	142
<i>Nevados de Chillán</i>									
NC-DB15	3/4/18	Intermediate	Dome complex	Dome explosion	224 (± 2.3)	77 (± 1.5)	92 (± 1.6)	749 (± 2.8)	1142
NC-DB2	29/12/16	Intermediate	Dome complex	Dome explosion	99 (± 5.4)	12 (± 2.3)	14 (± 2.4)	171 (± 5.6)	296
<i>Ontake</i>									

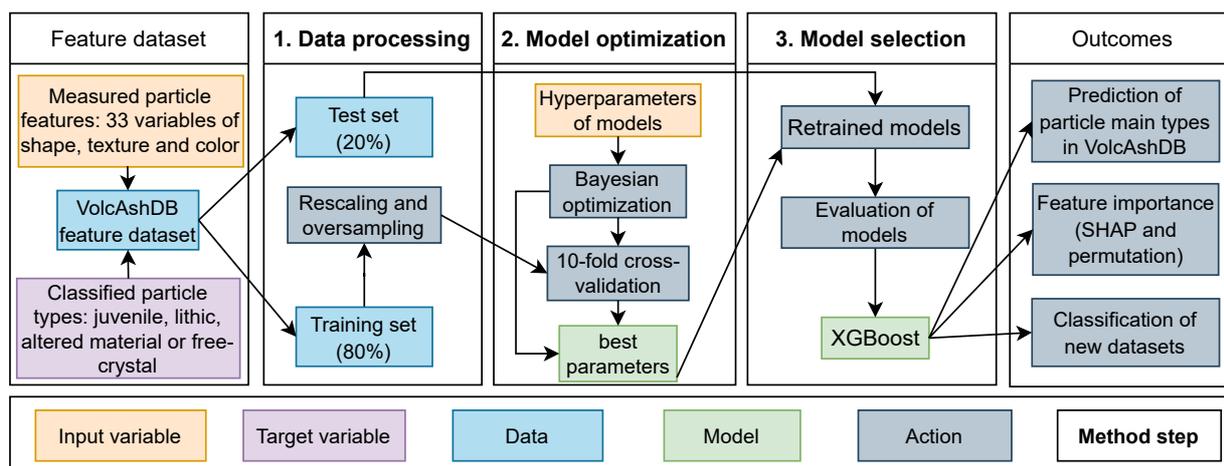
ON-DB1	27/9/14	Intermediate	Stratovolcano	Phreatic	777(\pm 0)	0	0	0	777
<i>Pinatubo</i>									
PI-DB1	2/4/91	Silicic	Caldera	Phreatic	386(\pm 3.7)	104(\pm 3.5)	0	16(\pm 1.5)	506
<i>Mount St Helens</i>									
MS-DB1	18/5/80	Silicic	Stratovolcano	Plinian	4(\pm 1.5)	0	255(\pm 1.8)	2(\pm 1.1)	261
Total					2888(\pm 1.2)	408(\pm 0.6)	1406(\pm 1.0)	1602(\pm 1.0)	6304

115 In addition to ash images, VolcAshDB also includes: (i) the value of 33 features of each
 116 ash particle related its shape, texture, and color, (ii) a label with the identification of the types
 117 of particle (free-crystal, altered material, juvenile, and lithic; Figure 1), and (iii) metadata for
 118 each particle, such as the sample grain-size fraction, the number of magnifications used for
 119 image acquisition, amongst others. The shape features in the database have been used in
 120 previous studies (Cioni et al., 2014; Dellino & La Volpe, 1996; Dürig et al., 2018; Leibrandt &
 121 Le Pennec, 2015; E. J. Liu et al., 2015), and include those sensitive to particle-scale cavities,
 122 (e.g., solidity), perimeter-based irregularities (e.g., convexity), and form (e.g., elongation; Liu
 123 et al., 2015). The textural features in VolcAshDB were obtained from calculations of the
 124 distribution of pixel intensities in grayscale across several particle regions based on the so-
 125 called Gray Level Cooccurrence Matrix (GLCM, Haralick et al., 1973). From the GLCMs we
 126 obtained features that indicate a more uniform texture (e.g., *Homogeneity*), and those that
 127 indicate a more complex or heterogeneous texture (e.g., *Dissimilarity*; Hall-Beyer, 2017). The
 128 color features of each particle were taken from the measurement of the mean, mode and
 129 standard deviation of the histogram distribution for each of the six channels in the Red-Green-
 130 Blue (RGB), and Hue-Saturation-Value (HSV) color spaces. For more details on the
 131 calculation and references of each feature, the reader is referred to Benet et al., (*preprint*), and
 132 they are summarized with the abbreviation in Table S1.

133 2.2 Development of a particle classifier using the measured particle features

134 The steps needed to develop a volcanic ash particle classifier vary if the input data are
 135 the measured features, or the particle images directly. Because the particle types are already
 136 classified, the models are trained by supervised learning (Verdhan, 2020). We used three steps
 137 to identify the best-performing classifier for the feature data (Figure 2): data processing, model
 138 optimization, and selection. We also compared the ability to classify unseen (test set) data
 139 using non-parametric, tree- and ensemble-based ML models. We found that the XGBoost
 140 model had the best scores, as is the case in studies in other fields (Chen & Guestrin, 2016;
 141 Dhaliwal et al., 2018). The XGBoost model was used to gain insights on the most important
 142 features by calculating the Shapley values and with feature permutation (Molnar, 2021).

143



144

145 **Figure 2.** Illustration of the steps involved from the dataset to the outcomes, including those to
 146 obtain the best optimized model, XGBoost. (1) Data processing of the full dataset (features and
 147 particle types), including the oversampling of the training set. (2) hyperparameter optimization
 148 and cross-validation to obtain the models with the highest cross-validation scores. (3)
 149 evaluation of the models with the test set (unseen by the model) and selection of XGBoost with
 150 the highest classification scores. The XGBoost classifier was applied for prediction of particle
 151 types and feature importance. See more details in main text and subsequent figures.

152

2.2.1 Data processing

153

154 The dataset consists of 33 features measured from each particle (variables; Table S1)
 155 and the particle types (target variable; Figure 2). The dataset is made of 6,300 particles and was
 156 divided into a training set (80% of the total particles) to optimize and fit the models, and a test
 157 set (20%), not used during the model's learning process. The original feature distributions are
 158 heterogenous and were standardized using the Scikit-learn's function *StandardScaler*, as it is
 159 commonly done to ease convergence of ML models (Géron, 2017). The standard scaler
 160 redistributes the values of each feature with the mean at 0, and the first standard deviation at 1
 161 and -1. The features from the test set were also standardized according to the scaler that was fit
 162 into the training set to avoid data leakage. Any outliers, defined as values higher and smaller
 163 than two standard deviations (Verdhan, 2020), were kept after visually confirming that the
 164 source images had no errors. Highly correlated variables were kept for estimating their
 165 importance for classification in the step of feature permutation (more details are reported in
 166 'Explaining the model's predictions' in Section 2.3.4). Highly correlated variables may cause
 167 multi-collinearity issues in regression models, but these haven't been reported in tree-based
 168 models (Kotsiantis, 2013).

168

169 The VolcAshDB dataset contains more altered material than juvenile and lithic particle
 170 types, and free crystals are relatively scarce (Table 1). Such uneven distribution of particle
 171 types may cause an imbalanced dataset problem. We addressed this issue by oversampling the
 172 less abundant particle types, using the SMOTE package, which uses a K-Nearest Neighbor
 173 algorithm (KNN) to generate synthetic data (Brownlee, 2020). This technique is strongly
 174 recommended to prevent the model from not learning to classify the less abundant class
 (Brownlee, 2020).

175

2.2.2 Hyperparameter optimization

176

177 Hyperparameters control the model learning process and are explicitly defined by the
 178 user. Hyperparameters are defined by ranges of values intrinsic to each model. We considered
 Decision Trees (DT), K-Nearest Neighbor (KNN), Random Forest (RF), Gradient Boost

179 Classifier (GBC), and the Extreme Gradient Boosting (XGBoost), and compiled their best
 180 hyperparameters values using Bayesian optimization, from the Scikit-optimize's function
 181 *BayesSearchCV*. This function searches for the optimal hyperparameters depending on the
 182 previous iterations, making computation faster and less intensive than iterating through the
 183 entire search space (Owen, 2022). The scores to evaluate the effect of the hyperparameters
 184 were obtained from 10-fold cross-validation of the training set. In the K-fold Cross-validation
 185 (where K is an integer), the data are iteratively divided into K training and testing folds for K
 186 times, as recommended to avoid overfitting (Verdhan, 2020). The highest cross-validation
 187 scores, using the optimal hyperparameters (Table S2), were obtained with the XGBoost with
 188 0.9 *F1-score* (as defined and calculated below in Section 2.2.3) closely followed by KNN and
 189 GBC with 0.88 *F1-score* (obtained scores of each model are shown in Figure S1).

190 2.2.3 Model evaluation and selection

191 The cross-validation scores indicate how well a model fits the training set. To evaluate
 192 the models' ability to generalize we also computed the predictions on the test set. Each
 193 prediction contains a confidence score per class which represents the likelihood of the
 194 prediction belonging to the class, and the score is given as a percentage (Mandal et al., 2021).
 195 The class, that is, the particle type in our case, with the highest confidence score is considered
 196 the predicted type by the model. Comparison between the predicted and the true types from
 197 VolcAshDB allows to categorise each prediction in one of the four following groups: True
 198 Positive (TP), where the prediction correctly identifies the class; True Negative (TN), where
 199 the prediction correctly identifies the absence of a class; False Positive (FP), where the
 200 prediction wrongly identifies the presence of a class, and False Negatives (FN), where the
 201 prediction wrongly identifies the absence of a class. The classification matrix (Figure S2) is
 202 typically used in ML to show the proportions of TP, TN, FP and FN for each class. Based on
 203 these proportions, we can calculate four well-known metrics to evaluate the models'
 204 performance (e.g., Verdhan, 2020):

$$205 \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$206 \text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$207 \text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$208 \text{F1-score} = \frac{2*TP}{2*TP+FP+FN} \quad (4)$$

209
 210 Classification scores in this study are reported based on the *F1-score*, as it combines the
 211 precision, dependent on the *FP*, and recall, dependent on the *FN*, into a single metric (Verdhan,
 212 2020), and is recommended for imbalanced datasets when *FN* and *FP* are equally important
 213 (Brownlee, 2020). We use the unweighted average of the *F1-scores* (the so-called *macro* from
 214 macro-averaging) of the four particle types to evaluate the overall model performance, as
 215 opposed to the weighted averaging, where the average is multiplied to a coefficient based on
 216 the number of particles per class (Verdhan, 2020). We found that XGBoost has the best
 217 classification performance with 0.76 *macro F1-score* amongst the optimized models and
 218 therefore is our selected model (classification score for each model are reported in Table S3
 219 and shown in Figure S3).

220 2.2.4 Explaining the model's predictions

221 Explainable AI (xAI) is a set of methods that provide explanations on the variables that
222 drive the model's predictions (Gianfagna & Di Cecco, 2021; Mishra, 2022; Molnar, 2021). We
223 used the method called "permutation feature importance" to assess the contribution of the 33
224 features to the model's prediction across all instances (i.e., the feature values from all
225 particles), and the SHapley Additive exPlanations (SHAP; Lundberg and Lee, 2017) method to
226 estimate the contribution of the features for each particle and, by aggregation, their global
227 importance (Molnar, 2021). In the permutation feature importance, the values of each feature
228 from the dataset are shuffled to measure the increase in prediction error. We used Scikit-learn's
229 function *permutation* on the test set from which we obtained a ranking of the features'
230 contribution between two end-members: "important" features, which cause an increase in
231 prediction error when shuffled, and "unimportant" features, where the error remains unchanged
232 or decreases (Molnar, 2021). We estimated the feature importance on each class by permuting
233 the features between each class and the rest (e.g., One-vs-Rest strategy).

234 The SHAP library can be used to explain individual model's predictions in regression
235 (e.g., Biass et al., 2022; Kondylatos et al., 2022), and classification problems (e.g., Panati et al.,
236 2022; Tang et al., 2021). The methods from the SHAP library are based on the Shapley values
237 (Shapley, 1953), which measure the contribution of the feature values to predict a certain value
238 with respect to the average prediction for all instances (Molnar, 2021). Shapley values were
239 calculated using TreeSHAP estimation method with raw output. Because Shapley values are
240 additive, TreeSHAP method adds and averages the contribution of each node in the ensembled
241 trees to obtain the Shapley value of each feature value per instance (Lundberg et al., 2018)—in
242 our study, an instance are the feature values per particle. The highest Shapley positive values
243 per instance are those which contribute the most to predict a given class. Averaging of the
244 Shapley values by particle type, or across the four particle types (free-crystal, altered material,
245 juvenile, and lithic), informs about the global feature importance (Lundberg et al., 2018),
246 which can be used for comparison with the permutation feature importance.

247 2.2.5 Classification strategies

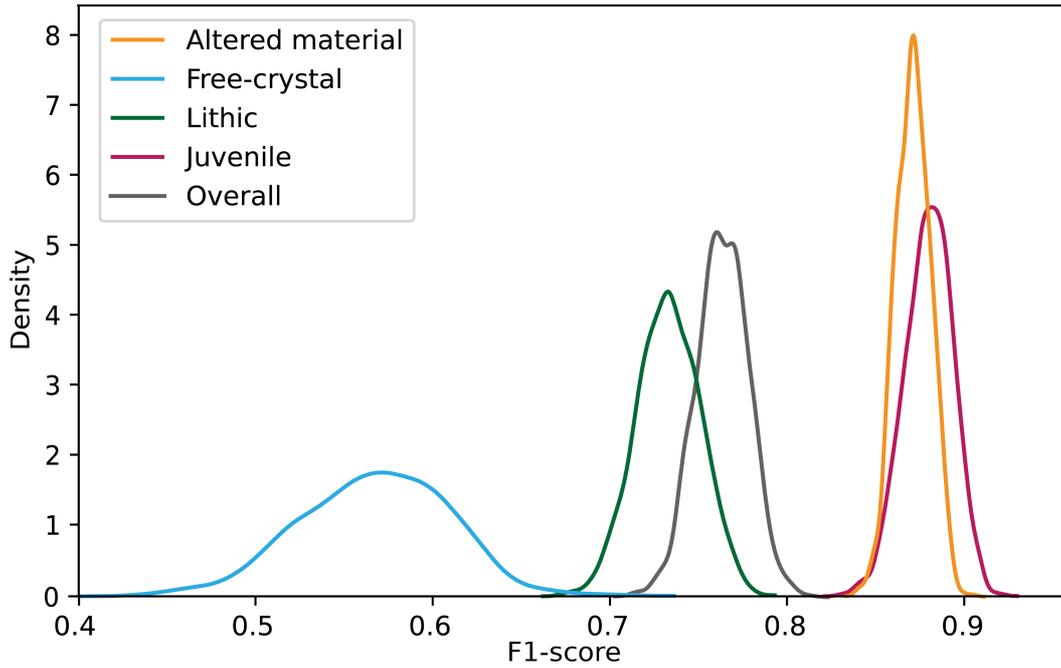
248 We applied three classification strategies to evaluate which model performs best: (i) the
249 multilabel, where the four classes are used to train the model at once and one prediction
250 probability is given for each class, with the highest value being the predicted class, (ii) the
251 One-vs-One (OVO), where each possible pair of classes trains a binary classifier (i.e., a total of
252 six classifiers, as there are six possible pairs for four classes), and their outputs are aggregated
253 to yield the predicted class (Herrera et al., 2016), and (iii) the One-vs-Rest (OVR), where each
254 class and its complementary (e.g., lithic vs non-lithic) are used to train a binary classifier (i.e., a
255 total of four), and their outputs are aggregated to yield the predicted class (Herrera et al., 2016).
256 For the OVO and OVR strategies, the outputs from the binary classifiers were aggregated with
257 the same weight to obtain the predicted class. There are more sophisticated aggregation
258 methods, such as the calibrated label ranking method (Fürnkranz et al., 2008), which adjust the
259 weights of each binary classifier aiming to mitigate class dependencies, and making the global
260 classification more robust (Herrera et al., 2016). However, we don't know of any
261 implementation of these methods in Python for the XGBoost model, and developing them from
262 scratch is out the scope of this study.

263 2.2.6 Effect of the training and test data split on the XGBoost scores

264 As noted above, we first split the dataset into a training (80% of all particle features in
265 VolcAshDB) and a test set (20%) and used the latter to evaluate the XGBoost's performance.
266 However, as splitting process is random it may affect the precision and accuracy of the

267 measured *F1-scores*. To estimate this error, we re-trained and evaluated the XGBoost at 1,000
 268 different values of random state, i.e., the hyperparameter that controls randomness. We
 269 obtained an average accuracy (*macro F1-score* of 0.76; Table S4) that is like the accuracy from
 270 the test set (*macro F1-score* of 0.75). The free-crystal type shows the widest variability
 271 (standard deviation of 0.04) and is the most inaccurate (*F1-score* of 0.57; Figure 3) amongst the
 272 particle types. This is likely because it is the least abundant type, and its classification is
 273 challenging given the different types of minerals and lack of a discriminant feature as
 274 explained below (Section 3.1). Accuracies of the three other types are higher (*F1-score* of
 275 0.73–0.88) and with better precision (standard deviation is < 0.02; Table S4).

276



277

278 **Figure 3.** Density plots of the *F1-scores* obtained from 1,000 runs of the XGBoost at different
 279 random state across particle types and aggregated as *macro F1-score* (Overall).

280 By averaging the *F1-scores* of each particle type, we obtain the *macro F1-score*
 281 distribution (Figure 3) and its variability (standard deviation; Table S4). To quantify the
 282 associated error (α), we use the second standard deviation (Hughes and Hase 2010):

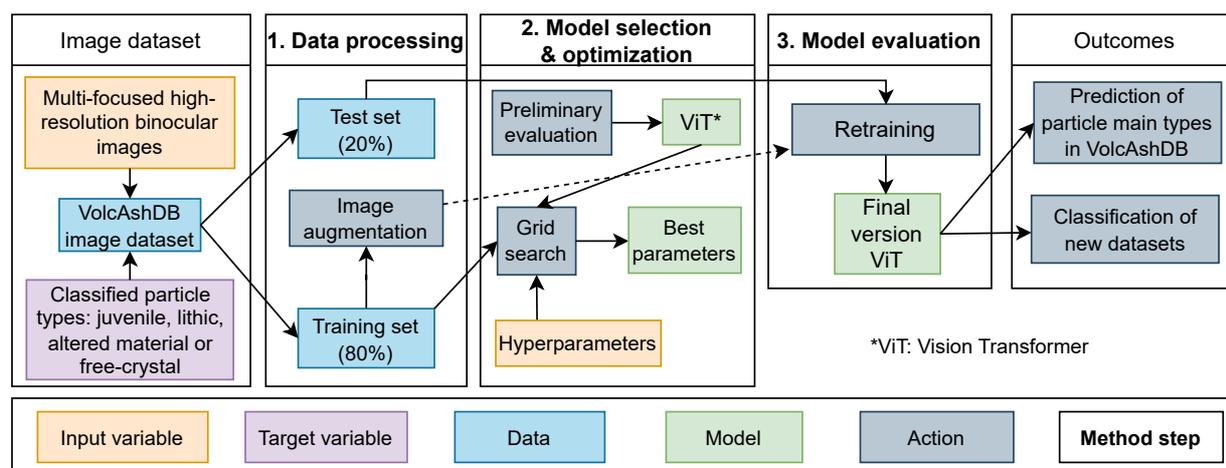
$$\alpha = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (5)$$

283 where N is the number of experiments, x is each measured value (i.e., *macro F1-score*)
 284 and \bar{x} is the mean. With the values noted above we obtain an error (α) of 0.03 for *macro F1-*
 285 *score* distribution and, since we used the second standard deviation, it is for a 95% confidence
 286 level. Therefore, the *F1-score* values can be reported as: 0.76 ± 0.03 *macro F1-score*, which is a
 287 small relative error of <5 %.

288 2.3 Development of a particle classifier using VolcAshDB image dataset

289 We used four steps to develop an optimized classifier for the image dataset (Figure 4):
 290 data augmentation, fine-tuning, selection, and evaluation. We compared the performance
 291 between three state-of-the-art models that have top accuracies in the reference dataset

292 ImageNet (Jia Deng et al., 2009): ResNet (He et al., 2016), which is the prevalent model of the
 293 so-called convolutional neural networks (CNN), Vision transformer (ViT; Dosovitskiy et al.,
 294 2020), which superseded ResNet in image classification, and ConvNeXT (Z. Liu et al., 2022),
 295 which is an optimized convolutional neural network that has surpassed performances of vision
 296 transformers. The models are available in the *Hugging Face* library (<https://huggingface.co/>),
 297 which also provides application programming interfaces (API) for their deployment. The
 298 model that yielded highest classification score was the ViT. We augmented the training dataset
 299 with an array of variations from the original images (see below), and the ViT reached a *macro*
 300 *F1-score* of 0.93, outperforming the XGBoost classifier. The images of the ash particle in
 301 VolcAshDB were obtained from processed multi-focused binocular images, but this is not the
 302 standard practice, and thus we also tested the ViT's ability to classify standard single-focus
 303 binocular images used in most studies of ash particles.



304

305 **Figure 4.** Illustration of the steps involved from the dataset to the outcomes, including
 306 those to fine-tune the Vision Transformer (ViT). (1) Data processing of the full dataset (images
 307 and particle types). (2) preliminary evaluation of the models using the base hyperparameters,
 308 selection of ViT and hyperparameter optimization through grid search. (3) Fine-tuning with the
 309 augmented dataset and final evaluation using the test set. The ViT classifier can be then applied
 310 for prediction of particle types. See more details in main text and subsequent figures.

311 2.3.1 Image augmentation and processing

312 The binocular images of ash particles in VolcAshDB are multi-focused, and contain
 313 four color channels: red, green, blue and alpha. The alpha channel is a binary mask that takes a
 314 value of 1 or 0 to separate between the particle pixels and those of the background (more
 315 details in the segmentation step in (Benet et al., *preprint*). We split the dataset into a train (80%
 316 of the total images in VolcAshDB) and test set. Then, we augmented the number of images in
 317 the training set based on six standard methods (Ayyadevara & Reddy, 2020): rotation (at 45°),
 318 translation (at 25 pixels), up-down and left-right flipping, and adding random noise and
 319 Gaussian blur at sigma values of 0.155 and 0.55. Increasing the amount of images allowed us
 320 to balance the distribution across particle types (~2900/class), and is generally recommended to
 321 increase model's robustness (Brownlee, 2020). Images were stored into four subdirectories,
 322 one for each class, of a root directory which is inputted to the *Hugging Face's API* for fine-
 323 tuning.

324 2.3.2 Fine-tuning, preliminary evaluation, and model selection

325 We fine-tuned the classifiers and did a preliminary round of evaluations to choose the
 326 best-performing model. Fine-tuning consists in making small adjustments to an already trained
 327 classifier, as opposed to training, where the data drives the model's learning process without

328 any prior exposure. We selected the model before hyperparameter optimization because each
 329 run is time consuming (lasting about 14–18 hours) and because the authors of each model
 330 already provide the base hyperparameters (Table S5). The fine-tuned model that yielded the
 331 highest accuracy is ViT (0.88), followed by ConvNext and ResNet, both with an accuracy of
 332 0.86.

333 2.3.3 ViT Hyperparameter optimization

334 We obtained the optimal hyperparameters following the grid search technique for two
 335 ranges of batch size and learning rate. In grid search, each hyperparameter is modified one step
 336 at a time, while the other hyperparameters remain fixed, throughout all the possible
 337 combinations (Owen, 2022). We found that the optimal batch size and learning rate are 16 and
 338 3×10^{-5} , respectively (accuracies obtained from grid search are reported in Table S6). Using
 339 these values, we tested three different optimizers, AdamW (Loshchilov & Hutter, 2019),
 340 Stochastic Gradient Descent (Sutskever et al., 2013) and Adagrad (Duchi et al., 2011) with the
 341 former performing the best (Table S7). We also tested and an increasing number of epochs
 342 (i.e., 5, 10, 15, 20), which didn't improve performance above 10, probably because the model
 343 had already converged.

344 2.3.4 Model evaluation

345 We fine-tuned again the ViT with the augmented training set and the optimal set of
 346 hyperparameters, and obtained a significant improvement, with a *macro F1-score* of 0.93. We
 347 obtained the same metrics of precision, recall, accuracy and F1-score, confusion matrix, and
 348 confidence scores as defined and calculated above (Section 2.2.3 Model evaluation and
 349 selection). In contrast with the XGBoost, the explainability of the model is very limited as
 350 further discussed below (see Section 4.1).

351 3 Results

352 We used the VolcAshDB ash particle features and images to train the XGBoost and
 353 ViT models and to evaluate their ability to classify them into altered material, free-crystal,
 354 lithic or juvenile types (Table 2). We found that overall, the ViT classifies very accurately,
 355 with a *macro F1-score* of 0.93, whereas the XGBoost is less performant with a *macro F1-score*
 356 of 0.77 (Table 2) but allows for explaining the model's predictions by interpretable AI
 357 methods. We describe below the model performance through the two datasets by particle type
 358 and some particle subgroups, such as those divided by the volcano, or one class versus another.

359 **Table 2.** *F1-score* values for the whole database (unweighted average or *macro*) and particle
 360 types obtained from various models, including XGBoost multilabel, One-vs-One (OVO), One-
 361 vs-Rest (OVR), and the multilabel image-based model ViT.

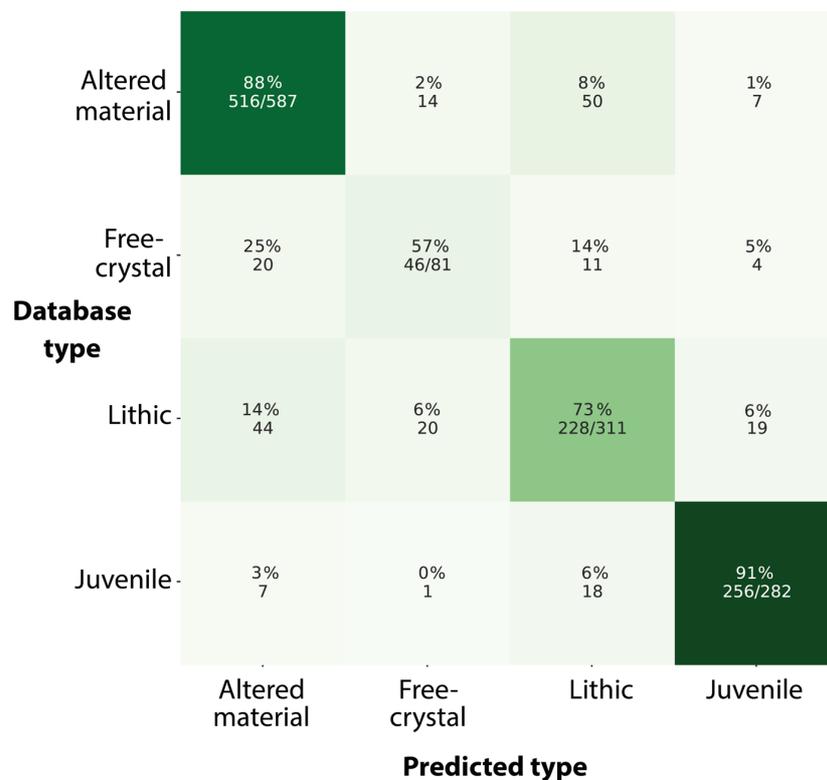
	Overall	Free-crystal	Altered material	Lithic	Juvenile
Multilabel XGBoost	0.77	0.57	0.88	0.74	0.90
OVO XGBoost	0.75	0.56	0.89	0.71	0.85
OVR XGBoost	0.76	0.55	0.90	0.73	0.88
Multilabel ViT	0.93	0.91	0.95	0.89	0.95

362

363 3.1 XGBoost quantitative evaluation

364 Overall, the XGBoost shows rather accurate *F1-scores* across classification strategies:
 365 0.76 for multilabel, 0.75 for OVO, and 0.76 for OVR (Table 2). Computation of the confusion

366 matrix (Figure 5) shows that the model classifies best the altered material type (*F1-score* of
 367 0.9), closely followed by the juvenile type (*F1-score* of 0.88), and less accurately the lithic type
 368 (*F1-score* of 0.74), and significantly less the free-crystal type (*F1-score* of 0.57).



369

370 **Figure 5.** Confusion matrix of the predictions by the XGBoost multilabel classifier. The
 371 percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and
 372 otherwise, the False Negative rate (lighter), all percentages with the corresponding number of
 373 particles per predicted type. The best classification is for altered material followed in
 374 descending order by juvenile, lithic and free-crystal types.
 375

376 Binary classifications using OVO and OVR between altered material, lithic and
 377 juvenile have accuracies > 0.80 (*macro F1-scores* of 0.82–0.97), whereas the free-crystal type
 378 is systematically lower (Table S8). A closer inspection by volcano and eruptive style reveals a
 379 wide range in XGBoost’s performances (Table 3). Predictions of juvenile particles are very
 380 accurate (*F1-score* of 0.97) at Kelud volcano but inaccurate (*F1-score* of 0.32) at Nevados de
 381 Chillán. Classification of lithics is rather accurate for samples of dome explosions (*F1-score* of
 382 0.77) but inaccurate (*F1-score* of 0.28) for those of phreatic events. Such fluctuations indicate
 383 limited robustness by the classifier and care should be taken for its application to other datasets
 384 on a case-by-case basis.

385 The likelihood that a particle belongs to a given type according to the model is reflected
 386 in the distribution of the confidence scores, and varies across particle types. Within the True
 387 Positives (*TP*), almost 90% of the juvenile *TP* have confidence scores > 0.9 , whereas ~40% of
 388 the free-crystal *TP* have confidence scores between 0.4–0.9 (Figure 6A). This means that the
 389 XGBoost is almost certain when predicting juvenile particles, but more unstable for free
 390 crystals. The confidence scores over the False Negatives (*FN*) show that the XGBoost
 391 identifies a relatively high number of lithic particles and free-crystals as altered material, with
 392 confidence scores > 0.9 (Figure 6B–C), hinting at some classification challenges that are
 393 revealed below using the Shapley values (see ‘Local feature importance’ in Section 4.3.2).

394 **Table 3.** *F1-scores* obtained from the multilabel XGBoost classifier of each particle type and their unweighted average (i.e., *macro*) for all
 395 particles in the test set (Overall), and across volcanoes and eruptive styles. These measurements also have an estimated precision of ± 0.03 .
 396

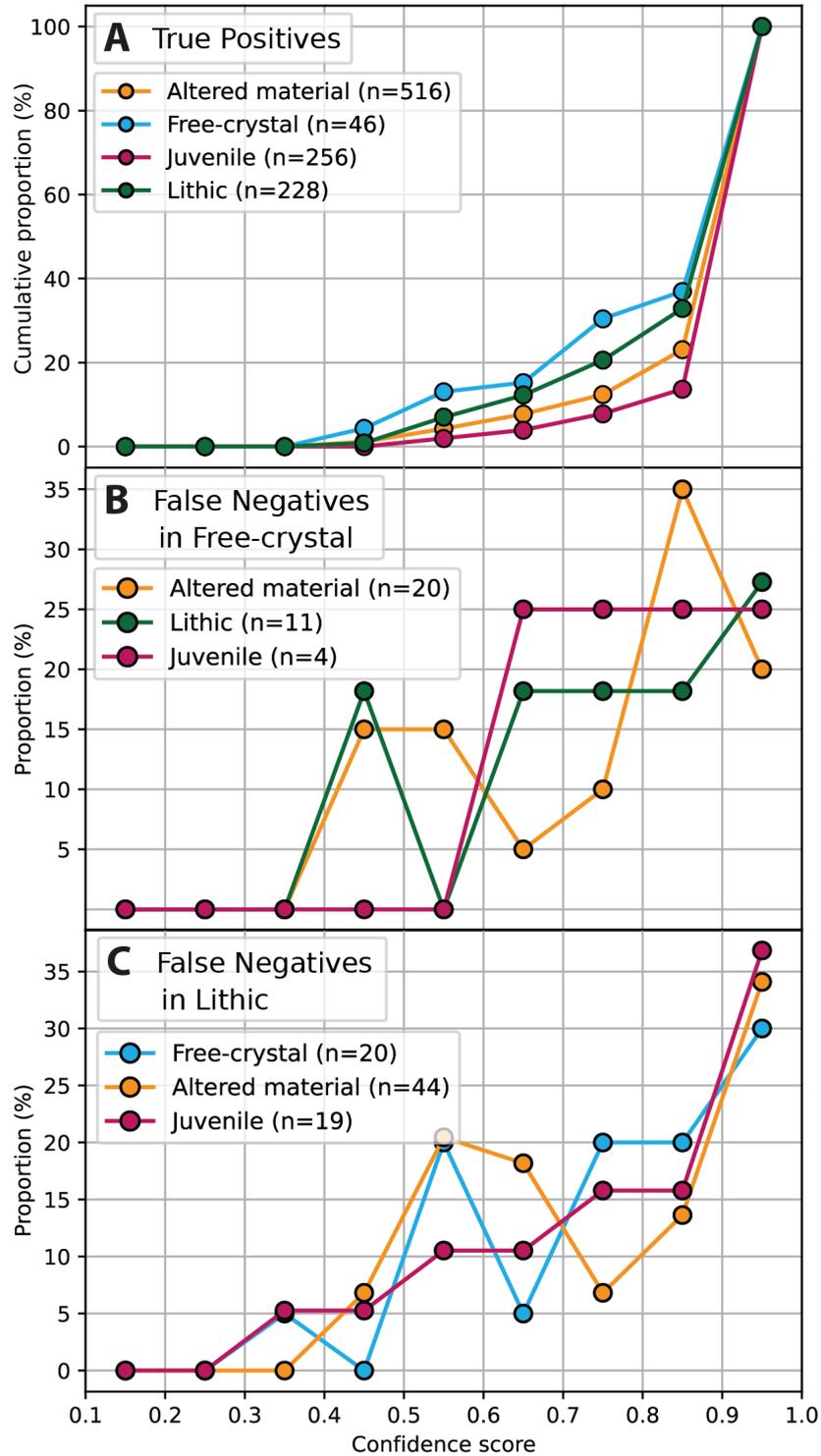
	Overall	Volcano					Eruptive style			
		Soufrière de Guadeloupe	Merapi	Nevados de Chillán	Cumbre Vieja	Kelud	Phreatic	Dome explosion	Lava fountaining	Sub- plinian/ Plinian
F1-score (macro)	0.77	0.76	0.73	0.6	0.87	0.73	0.62	0.65	0.87	0.76
F^1	0.57	0.7	0.67	0.59	–	0.6	0.64	0.51	–	0.7
A^2	0.88	0.92	0.91	0.7	–	0.81	0.95	0.82	–	0.84
L^3	0.74	0.67	0.6	0.77	0.83	0.54	0.28	0.8	0.83	0.42
J^4	0.9	–	–	0.32	0.92	0.97	–	0.46	0.92	0.99

397 1F : Free-crystal

2A : Altered material

3L : Lithic

4J : Juvenile



398

399 **Figure 6.** Line plots of the confidence score versus (A) the cumulative proportion of True
 400 True Positives (TP), (B) False Negatives (FN) in free-crystal, and (C) lithic types. The distribution
 401 of the data have been plotted into 9 bins of size 0.1. We don't use cumulative proportion in
 402 (B) and (C) given the limited number of FN. The meaning of the Plot in (A) can be
 403 understood by the following two examples: if we take the value of juvenile TP at a
 404 confidence score between 0.8–0.9, there is a low cumulative proportion of ~10%, whereas in
 405 the next bin, 0.9–1.0 of confidence score, we have the vast majority (~90%) of the juvenile
 406 TP. If we take the value of free-crystal TP at a confidence score between 0.8–0.9, there is a

407 significant cumulative proportion of almost 40%. This shows that XGBoost is more reliant
408 predicting juvenile particles than free crystals.

409 3.2 What features drive XGBoost ash particle type predictions?

410 3.2.1 Global feature importance

411 We identified the features driving the XGBoost's predictions with two approaches:
412 applying the permutation feature importance, and computing the mean of the Shapley values
413 (see Section 2.3.4). Although the calculation of the two methods is quite different, they
414 yielded overall a similar feature importance ranking, and we identified the following three as
415 the most important features (Table 4): (i) the mean of the hue channel (*hue_mean*), which is a
416 feature from the Hue-Saturation-Value color space that measures the averaged chromaticity;
417 (ii) the *correlation*, a textural feature that measures the degree of similarity between pixel
418 relationships (Hall-Beyer, 2017); and (iii) the mode of the blue channel (*blue_mode*), which
419 measures the most frequent pixel intensity of the blue channel of the particle image.

420

421 **Table 4.** Feature importance identification based on mean of Shapley values and
422 feature permutation. These two methods calculate the feature importance values differently
423 and can't be directly compared. The relative ranking of the features importance is similar (top
424 ten ranked features in bold) with the same top two ranked features (*hue_mean* and
425 *correlation*). We used the Shapley mean value for feature importance per particle type
426 (shown as a plot in Figure 7), the top three of which are underlined. For the meaning of the
427 abbreviations of each feature please see Table S1. The permutation feature values have been
428 multiplied by ten for better readability, as the importance lies on the relative values across
429 features.

Feature importance method	Mean of Shapley values					Feature permutation				
	Per particle type (Multilabel)				Total	Per particle type (OVR)				Total
	A	F	L	J		A	F	L	J	
hue_mean	<u>0.78</u>	<u>0.86</u>	0.12	<u>1.15</u>	<u>2.91</u>	0.91	0.41	0.15	0.91	1.22
correlation	<u>0.46</u>	0.33	0.33	<u>0.55</u>	<u>1.68</u>	0.34	0.02	0.19	0.04	0.29
blue_mode	<u>0.31</u>	0.10	<u>0.48</u>	0.54	<u>1.43</u>	0.06	0.04	0.00	0.01	0.10
value_mode	0.28	0.23	<u>0.60</u>	0.20	1.31	0.05	0.05	0.24	0.00	0.00
saturation_mode	0.10	0.27	-0.01	<u>0.80</u>	1.17	0.02	0.06	0.10	0.10	0.13
convexity	0.02	<u>0.52</u>	0.06	0.48	1.10	0.01	0.06	0.00	0.03	0.03
red_mean	0.16	0.18	<u>0.53</u>	0.21	1.07	0.03	0.03	0.01	0.01	0.04
blue_std	-0.06	<u>0.81</u>	0.06	0.19	1.00	0.34	0.24	0.04	0.04	0.28
green_mode	0.18	0.27	0.11	0.18	0.73	0.03	0.02	0.01	0.03	0.02
saturation_std	0.02	0.39	0.00	0.30	0.70	0.07	0.00	0.00	0.08	0.11
solidity	0.04	0.40	-0.01	0.24	0.68	0.08	0.01	0.07	0.02	-0.04
blue_mean	0.15	0.16	0.03	0.29	0.64	0.06	0.05	0.01	0.01	0.05
homogeneity	0.13	0.08	0.32	0.06	0.59	0.16	0.03	0.12	0.00	0.06
asm	0.21	0.29	0.01	0.02	0.53	0.18	0.03	0.00	0.00	0.14
contrast	-0.03	0.07	0.12	0.35	0.51	0.11	0.03	0.02	0.00	0.03
hue_std	0.09	0.16	0.05	0.20	0.49	0.14	0.13	0.11	0.00	0.14
green_mean	0.09	0.16	0.09	0.13	0.46	0.16	0.02	0.13	0.00	0.13
saturation_mean	0.07	0.05	0.15	0.18	0.46	0.01	0.05	0.00	0.01	0.04
circ_cioni	0.01	0.03	0.01	0.21	0.26	0.01	0.00	0.02	-0.01	-0.02
energy	0.05	0.02	0.06	0.00	0.14	0.03	0.00	0.09	0.00	0.01
red_std	-0.01	0.00	0.03	0.09	0.11	0.03	0.13	0.00	0.00	0.03
Total	3.12	5.51	3.13	6.51		2.86	1.43	1.33	1.29	

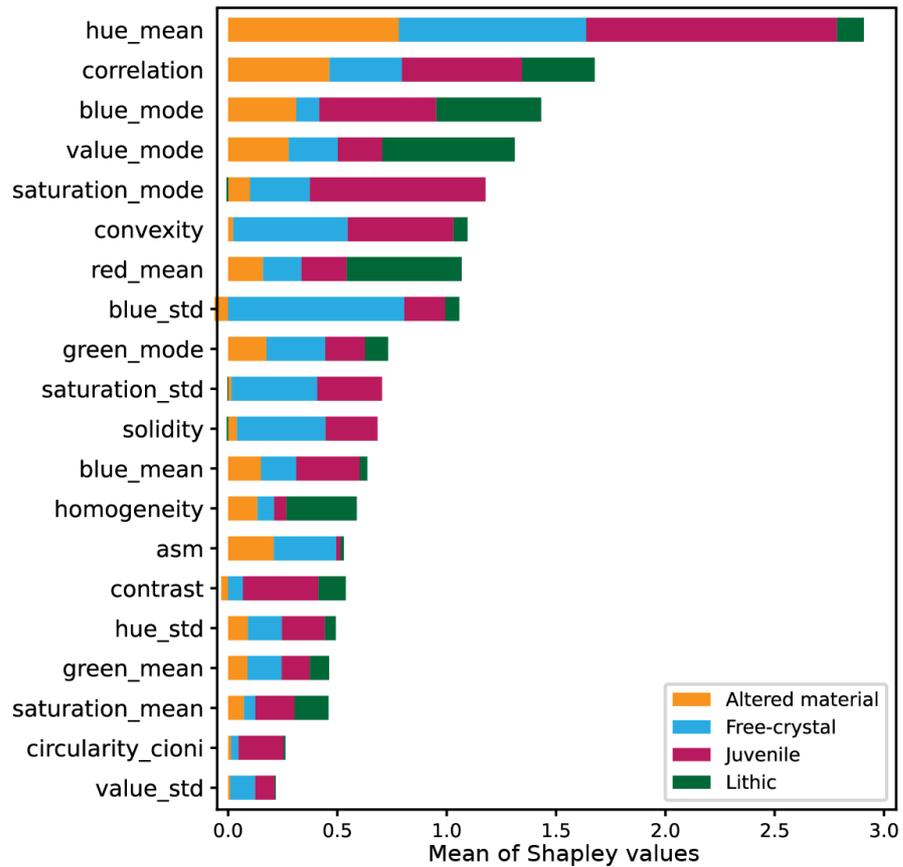
431 3.2.2 Local feature importance across particle types

432 We identified the most important features used by the XGBoost to predict each
 433 particle type based on the Shapley values, which considers the interaction between the four
 434 particle types, unlike permutation which is based on the One-vs-Rest approach. Shapley
 435 values calculate the contribution of each feature to the actual prediction with respect to the
 436 expected prediction (Gianfagna & Di Cecco, 2021; Lundberg et al., 2018; Molnar, 2021).
 437 Thus, we can use the Shapley values of an individual particle prediction to identify which
 438 features were more important or average them across particle types to identify the global
 439 discriminant features per type (Figure 7). These vary according to the particle type as
 440 follows:

- 441 (1) Altered material has the highest classification success with a *F1-score* of 0.90 and is
 442 predicted through color (*hue_mean* and *blue_std*), texture (*correlation*) and shape
 443 (*convexity*) (Figure 8A). A group of True Positives (*TP*) with *hue_mean* values
 444 between -3 and -2 (rescaled as described in Section 2.3.1) is revealed by the Shapley
 445 dependence plot (Figure 8B), which relates feature values (*hue_mean*) and their
 446 associated Shapley values for each particle (Lundberg et al., 2018). Such *TP* have
 447 almost 100% of confidence scores and consist of white (Figure 8C), red (predicted by
 448 *red_mode*, Figure 8D), rounded, hydrothermally altered material.
- 449 (2) The juvenile particles are accurately classified with a *F1-score* of 0.88 with color
 450 (*hue_mean*, *saturation_mode*), texture (*correlation*), and shape (*convexity*) (Figure
 451 9A). The *saturation_mode* feature, which relates to the intensity of color, is
 452 discriminant (Shapley values > 1) with values of 0–2 (Figure 9B). The *value_mode*,
 453 which measures the amount of reflected light, or gloss, and which is considered
 454 characteristic of juvenile particles under the binocular (Miwa et al., 2009) is also very
 455 important. Low values of *convexity* are also relevant for discrimination, as could be
 456 expected by the presence of vesicles on the particles' surfaces (Figure 9C). Moreover,
 457 the XGBoost predicts instances with lower *hue_mean* and *saturation_mode* as lithic
 458 (i.e., False Negative, FN), which correspond to darker, mid to high crystallinity
 459 juvenile particles from dome explosions (Figure 9D).
- 460 (3) The lithic particles are moderately well classified with a *F1-score* of 0.74, and is
 461 mainly discriminated through color (*value_mode* and *read_mean*) and texture
 462 (*homogeneity* and *correlation*) features (Figure 10A). Low values of *value_mode*,
 463 ranging between of -1.7 to 0 (Figure 10B), discriminate lithic particles. These features
 464 together with relatively high values of *correlation* reflect dark lithic particles with
 465 uniform texture (Figure 10C). In contrast, instances with higher pixel intensity-based
 466 features (*hue_mean* and *green_mean*) are a source of FN, as suggested by negative
 467 Shapley values, and are classified as altered material (Figure 10D).
- 468 (4) Free-crystals are the least accurately classified with *F1-score* of 0.54, and is mainly
 469 discriminated by color (*blue_std*, *hue_mean*), shape (*convexity*) and textural
 470 (*correlation*; Figure 11A). Unlike the other types, the most discriminant feature
 471 doesn't cluster particles as shown by the *blue_std* values as a function of the Shapley
 472 values doesn't yield any cluster of *TP* (Figure 11B), and those with Shapley values >
 473 1.5 overlap with altered material (Figure 11C). Thus, the XGBoost has limited
 474 predictability of free crystals, which is consistent with low a *F1-score* yielded from
 475 Free-crystals vs Rest binary classification (Table S8). Possible causes for this, besides
 476 the lack of a discriminant feature, include the presence of glass films on the crystal's
 477 surface, the wide range of aspects of different minerals (mostly plagioclase and

478
479
480

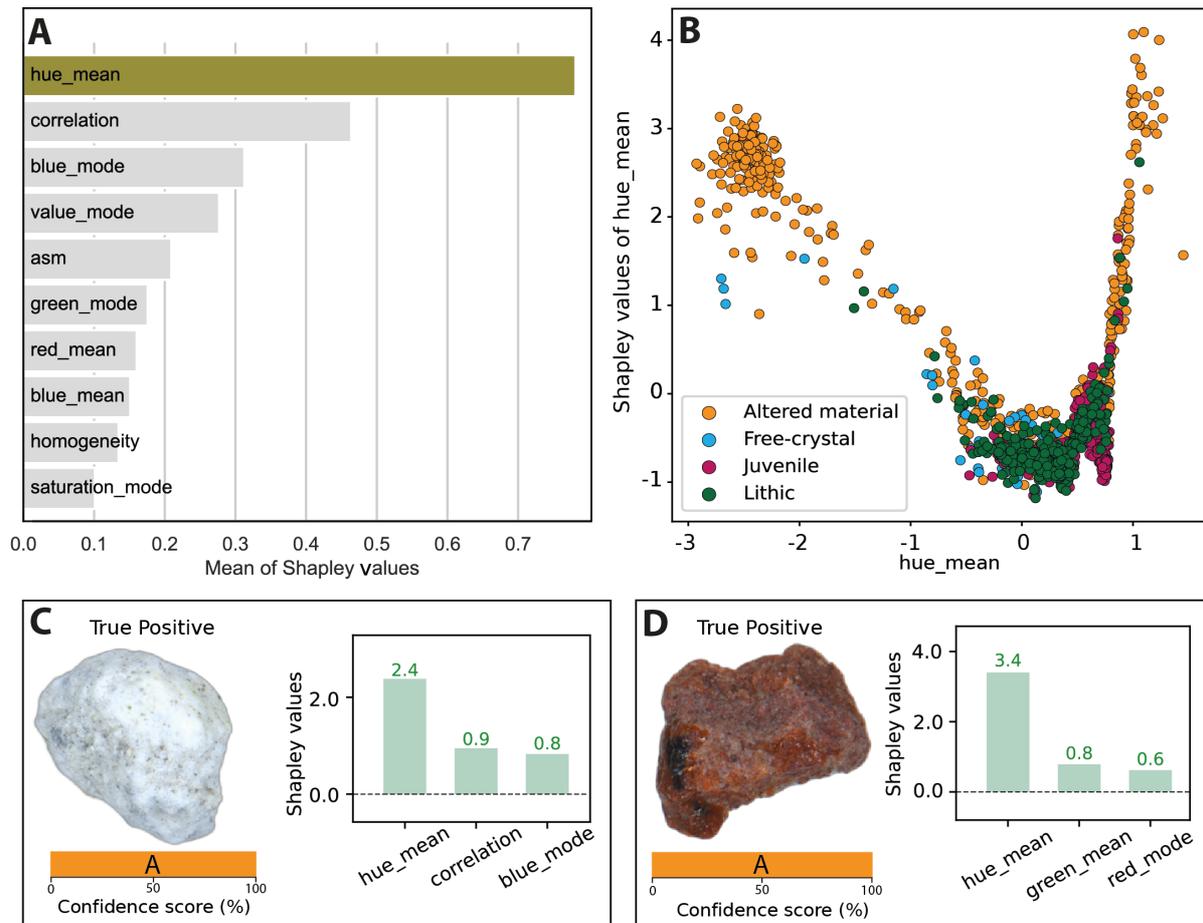
pyroxene but also amphibole and sulfur-group minerals), and the significant rate of composite particles (e.g., crystals attached to glass) that are not reflected in the label (Figure 11D).



481

482 **Figure 7.** Aggregated mean of the Shapley values by particle type. Note that some features
483 are important for discrimination of multiple particle types (e.g., *hue_mean*) and other features
484 are more discriminant of a specific type (e.g., *value_mode* for lithic type). Meaning of the
485 abbreviations can be found in Table S1.

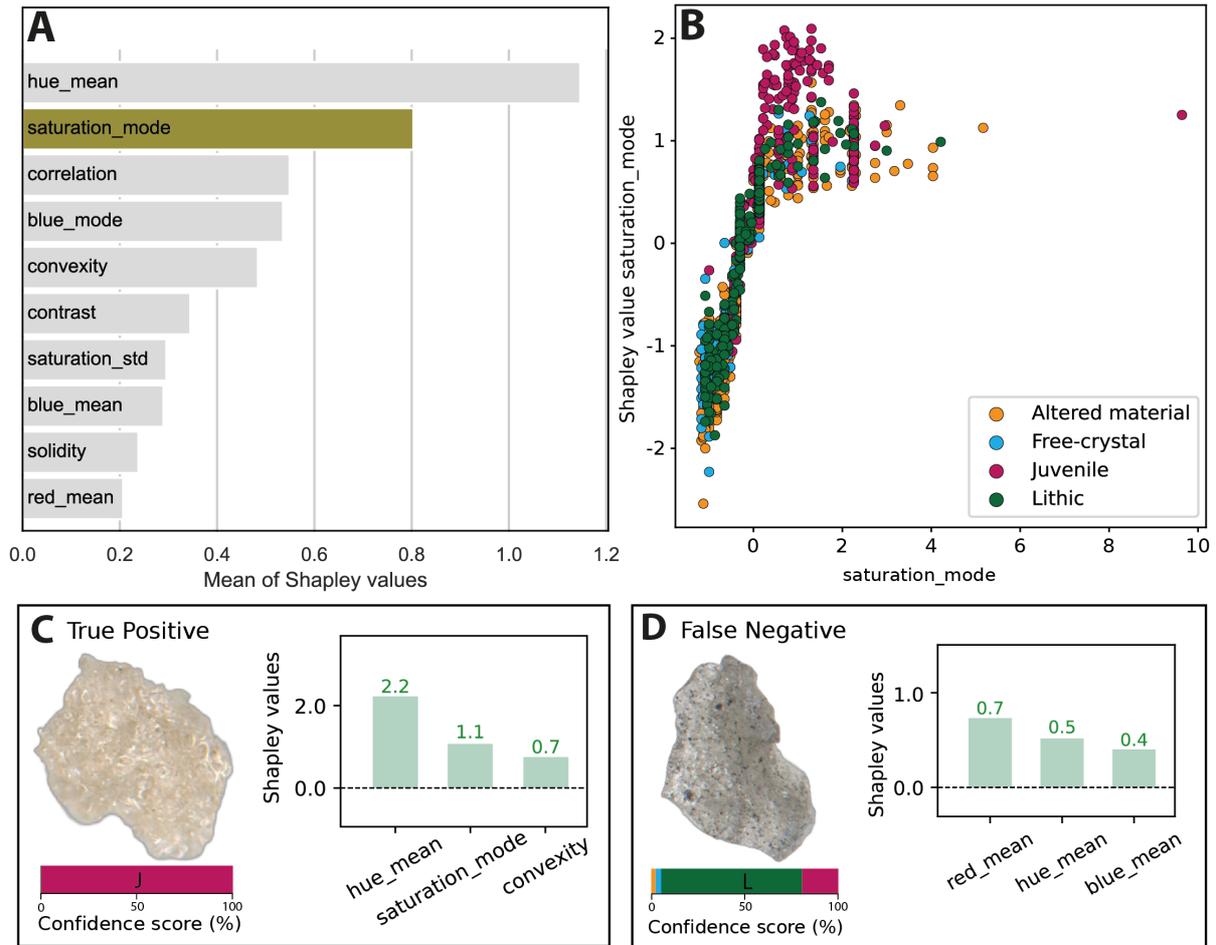
486



487

488 **Figure 8.** Summary plots to explain predictions of the altered material particle main type. (A)
 489 Feature importance according to the mean of the Shapley values, the higher the value the
 490 more the importance of the feature in the correct prediction. In (B) the Shapley dependence
 491 plot shows the relation of the Shapley value and the feature value for each particle type, and
 492 is commonly used to identify clusters of a specific class (particle main type) along the feature
 493 domain (Lundberg et al., 2018). For example, at values of -3 to -2 of *hue_mean*, there is a
 494 cluster of particles with high Shapley values and thus correctly classified as altered material.
 495 (C) and (D) are two examples of particles to show confidence score (A: Altered material),
 496 and the three features with the highest Shapley values. They are both True Positives and have
 497 been predicted at maximum confidence score with *hue_mean* (the mean of the chromaticity)
 498 being the main discriminant feature.

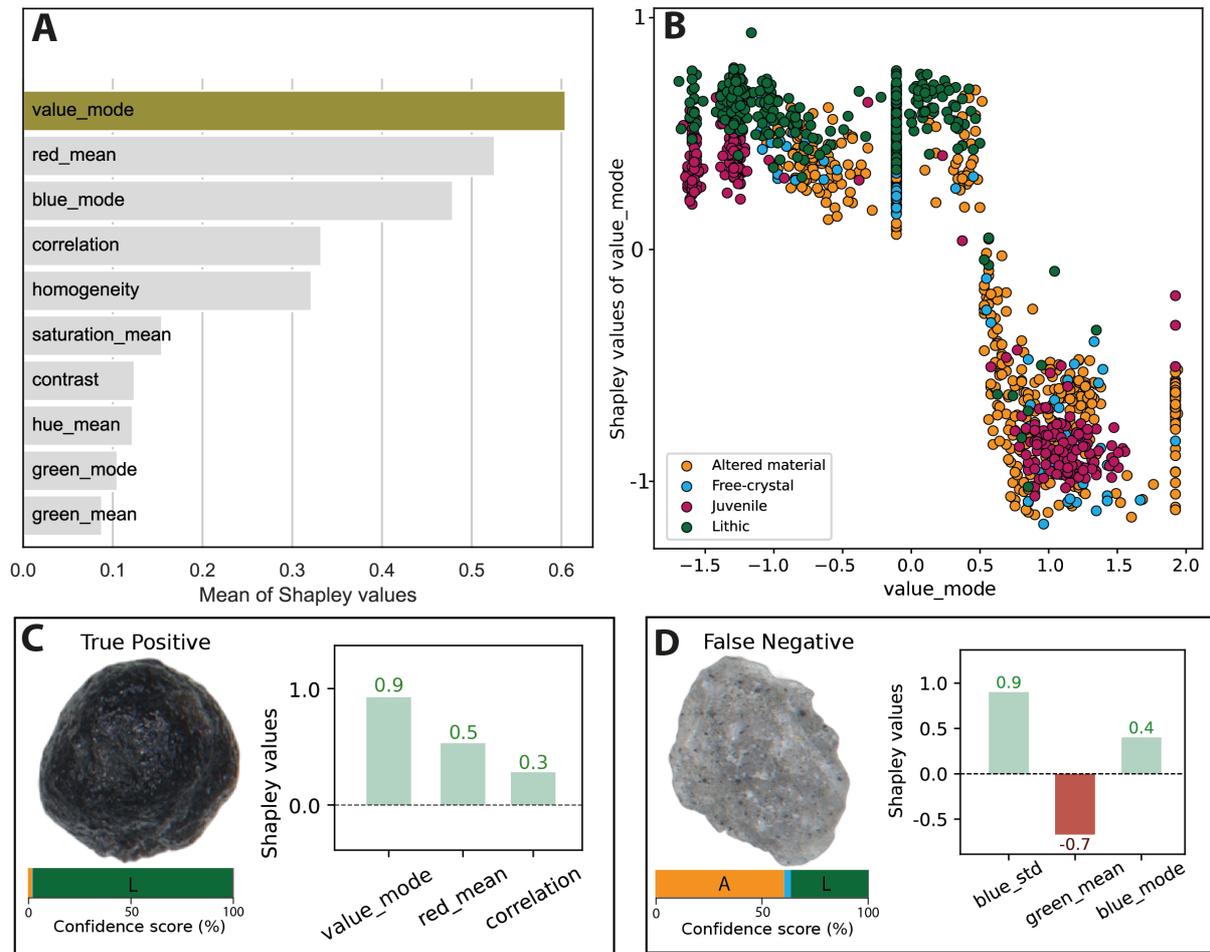
499



500

501 **Figure 9.** Summary plots to illustrate the features that contribute the most to the correct
 502 predictions of the juvenile particles. (A) Feature importance based on the mean of the
 503 Shapley values. (B) Shapley dependence plot. Note a cluster of juvenile particles around
 504 *saturation_mode* values between 1–3. (C) and (D) are examples of two predictions of the
 505 particle image, with the horizontal bar showing the confidence score across particle types,
 506 and the vertical bars the associated Shapley values. (C) shows a True Positive predicted at
 507 maximum confidence score with the *hue_mean* (chromaticity), *saturation_mode* (mode of the
 508 intensity of the color), and *convexity*. (D) is an example of a particle that was predicted by
 509 XGBoost model as lithic with a confidence of 70% (size of the green area in horizontal bar

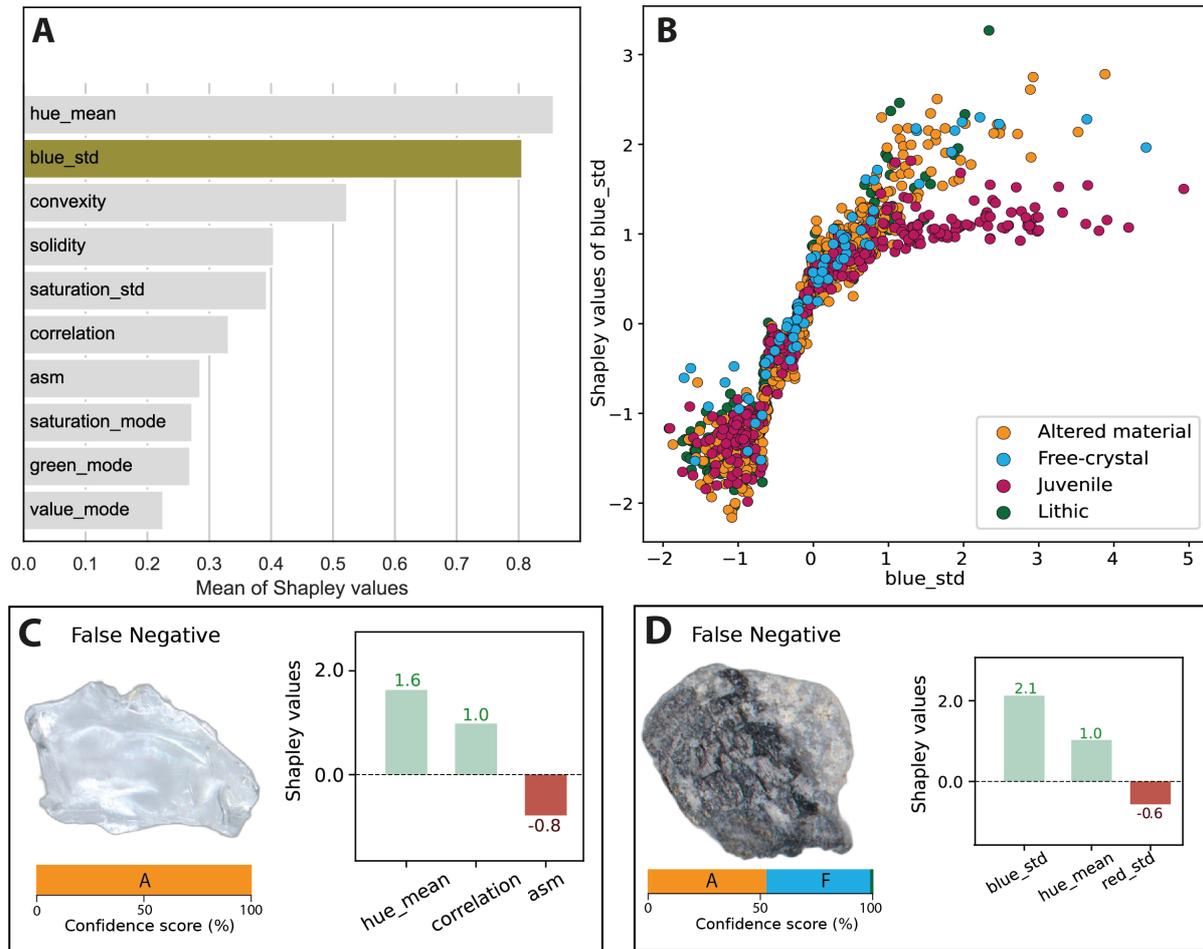
510 plot) based on the *red_mean* (mean of the red channel), which is predominantly discriminant
 511 of lithic particles (Figure 10A), but was classified as juvenile in VolcAshDB.



512

513 **Figure 10.** Summary plots to explain predictions of the lithic type. (A) Ranking of the
 514 features according to the mean of the Shapley values. (B) The Shapley dependence plot
 515 shows correct predictions of lithic particles with high Shapley values at negative values of
 516 *value_mode*. (C) and (D) show for each prediction the particle image, confidence score across
 517 particle types, and the associated Shapley values. (C) shows a dark particle that is correctly
 518 classified as lithic with low *value_mode* (luminosity), whereas (D) shows that XGBoost gives
 519 similar confidence scores to the altered material and lithic types, with the former being
 520 slightly preferred given the values of *green_mean*, which are uncharacteristic of the lithic
 521 type (shown by negative Shapley value -0.7). Discrimination of lithic and altered material

522 particles such as in (D) is often not straightforward when weathering is incipient (Benet et al.,
 523 *preprint*).



524

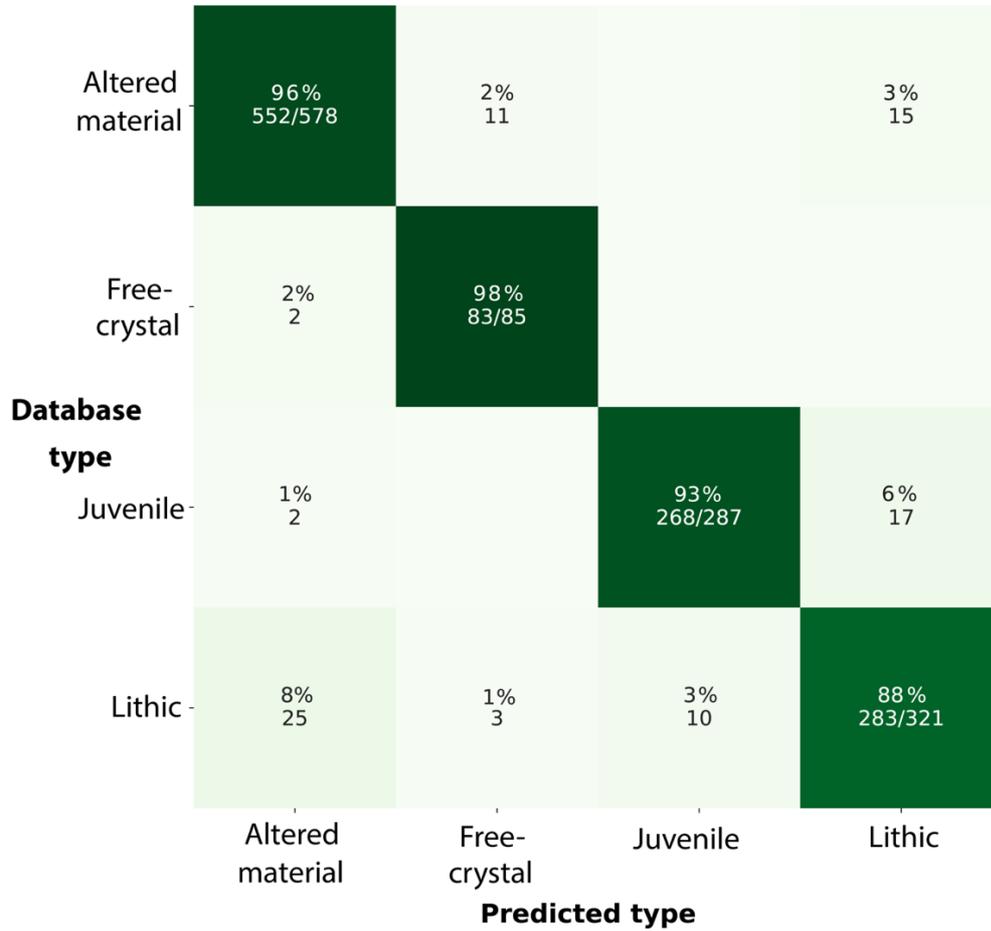
525 **Figure 11.** Summary plots to explain predictions of the models for the free-crystal type. (A)
 526 Feature importance based on the mean of the Shapley values. (B) Shapley dependence plot.
 527 Note that the feature values have been rescaled by a standard scaler. (C) and (D) show for
 528 each prediction the particle image, confidence score across particle types, and the associated
 529 Shapley values. (C) shows particle that is likely a fragment of plagioclase crystal but is
 530 misclassified as altered material, because the free-crystal type lacks discriminant features (see
 531 main text for more details). (D) an additional source of false negatives are particles consisting
 532 of more than one material, such as those made of glass attached to a crystal. In this case, the
 533 model's prediction correctly identifies two particle types, which is more accurate than using
 534 one single particle type as label.
 535

536 3.3 ViT quantitative evaluation

537 3.3.1 General evaluation

538 The ViT base model was fine-tuned using ~10,000 images from the augmented
 539 training set and evaluated with the test set (see Section 2.3 for information on each step). We
 540 obtained accurate classification for the whole test set (*macro F1-score* of 0.93), and also
 541 across particle types (Figure 12): altered material (*F1-score* of 0.95), juvenile (*F1-score* of
 542 0.95), free-crystal (*F1-score* of 0.91) and lithic (*F1-score* of 0.89). More than 85% of True
 543 Positives (*TP*) are predicted at high confidence scores (> 0.9; Figure 13A) which shows that

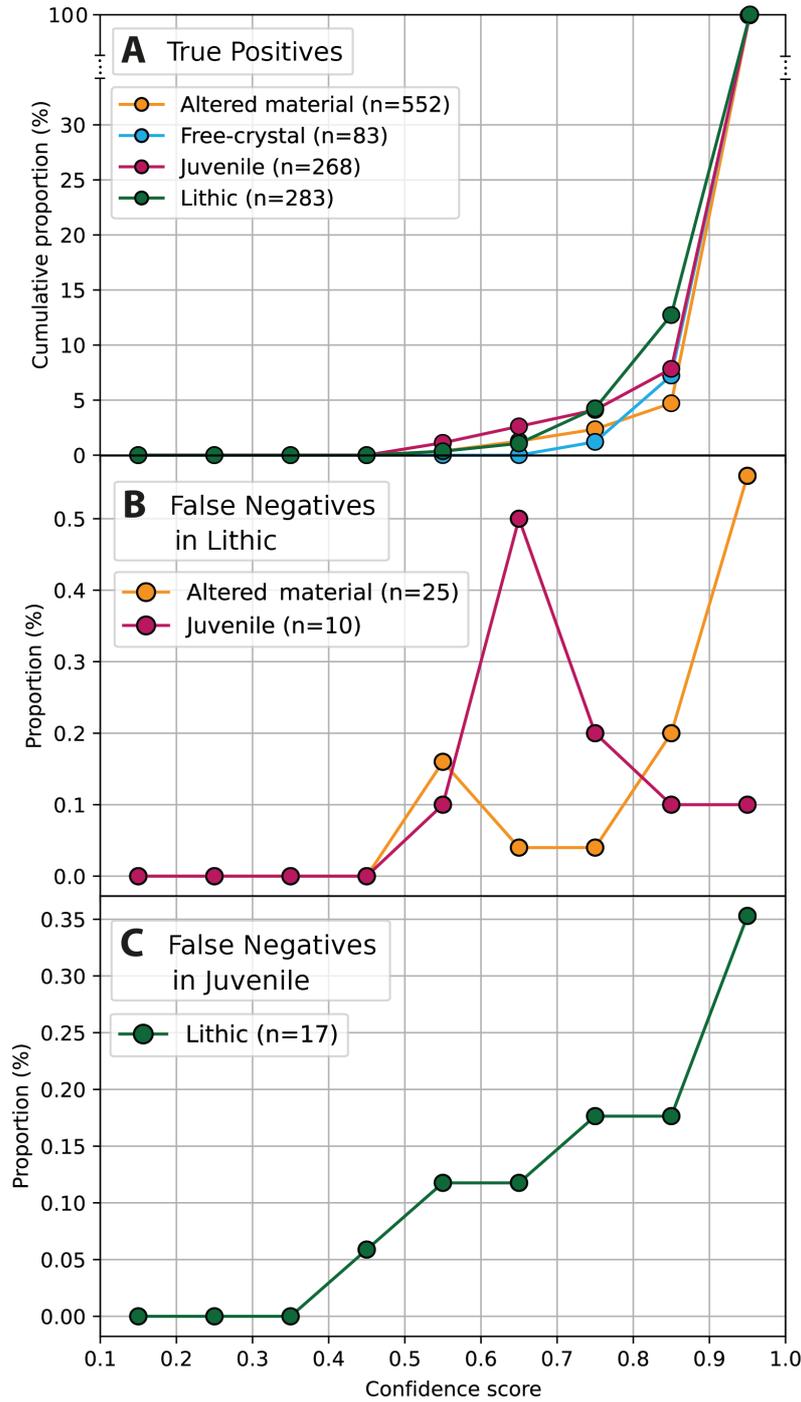
544 ViT classifies confidently and accurately. The False Negatives (*FN*) mostly consist of lithic
 545 particles classified as altered material and juvenile, a few of which at high confidence scores
 546 (Figure 13B), and also of juvenile particles classified as lithic type (Figure 13C). Below, we
 547 identify specific groups of particles that make up the *FN* and discuss the possible causes.



548

549 **Figure 12.** Confusion matrix of the predictions by the ViT image classifier. The percentages
 550 show the True Positive rate if positioned in the diagonal matrix (darker green), and otherwise,
 551 the False Negative rate (lighter), all percentages with the corresponding number of particles

552 per predicted type. The best classification is for free-crystal followed by altered material,
 553 juvenile and lithic.



554

555 **Figure 13.** Line plots of the confidence score versus (A) the cumulative proportion of True
 556 True Positives (TP), (B) False Negatives (FN) in free-crystal and (C) lithic types. The distribution
 557 of the data have been plotted into 9 bins of size 0.1. We don't use cumulative proportion in
 558 (B) and (C) given the limited number of FN. Two examples on how to read (A) are described

559 in Figure 6. Note that the ViT predicts True Positives at high confidence score values,
560 although it is less certain about the lithic particle type.

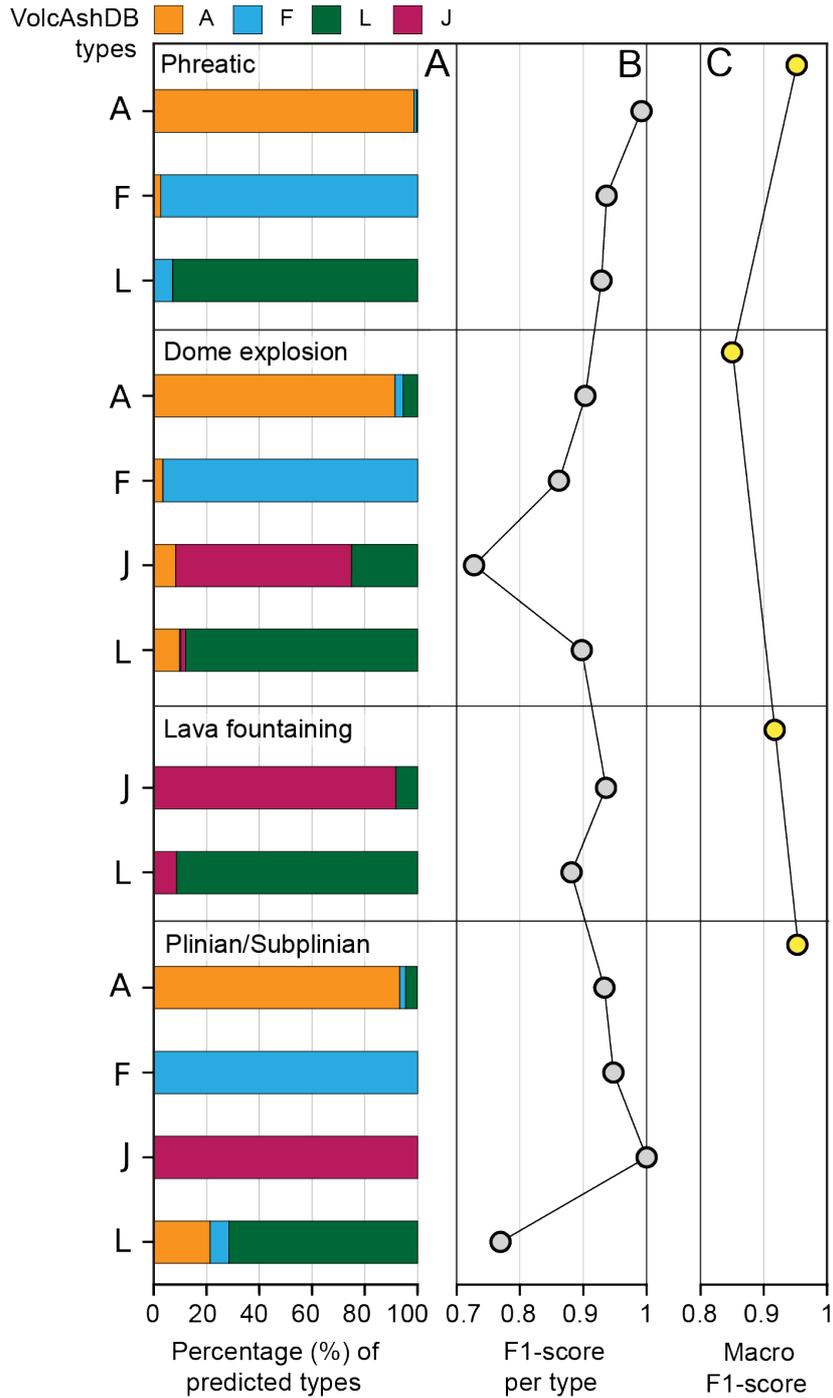
561 3.3.2 ViT's evaluation across volcanoes, eruptive styles, and individual particles

562 A closer inspection of the results across eruptive styles and volcanoes (Table S9)
563 reveals a range of classification accuracies, from moderate (*F1-score* of 0.73) up to optimal
564 classification performance with a *F1-score* of 1.0 (Figure 14):

- 565 (1) Ash particles from phreatic events are in general well classified (*macro F1-score* of
566 0.95), including the particle main types: altered material (*F1-score* of 0.99), free-
567 crystal (*F1-score* of 0.94) and lithic (*F1-score* of 0.93). The ViT successfully
568 classifies the most common groups of particles in these samples such as hydrothermal
569 aggregates (Figure 15A) and weathered material (Figure 15B).
- 570 (2) Particles from samples of dome explosions are classified with the lowest accuracy
571 (*macro F1-score* of 0.85) among the eruptive styles. The ViT accurately classifies
572 free-crystal (*F1-score* of 0.86), altered material (*F1-score* of 0.90) and lithic (*F1-*
573 *score* of 0.90) types, but is less accurate (*F1-score* of 0.73) for the juvenile type with
574 most False Negatives (*FN*) classified as lithics. However, the confidence scores of
575 some *FN* show a transition between the juvenile and lithic types that has explanatory
576 value. This means that particles may have both juvenile and lithic traits, and thus a
577 measure on the types' prevalence seems more realistic than using mutually exclusive
578 types like in VolcAshDB. Particles with combined traits are common in samples from
579 Nevados de Chillán Volcanic Complex (Figure 15C), which originated from a
580 relatively long-lived dome-forming eruption cycle. An additional challenge is that the
581 ViT confidently classifies as lithics some particles that are labelled as juvenile and,
582 since petrographic classification was not always straightforward (Benet et al.,
583 *preprint*), it is difficult to decide whether these are False Negatives, or instead,
584 petrographic classification errors (Figure 15D), especially when ML-based image
585 classifiers have surpassed human performances in other fields (He et al., 2015).
- 586 (3) Ash particles from lava fountaining are generally accurately classified (*macro F1-*
587 *score* of 0.94), between juvenile (*F1-score* of 0.94) and lithic (*F1-score* of 0.88)
588 types. Most of the lithic particles belong to recycled juvenile particles, which are
589 critical to avoid overestimating the amount of juvenile component (D'Oriano et al.,
590 2022) and their identification typically requires examination in the SEM (D'Oriano et
591 al., 2014). The high score suggests that the ViT can discriminate between them to
592 some extent (Figure 15E), but a more robust labelling by a team of experts and a
593 larger dataset containing SEM images is necessary to obtain more robust conclusions.
594 On the other hand, the juvenile particles consist of glossy, smoothed surface,
595 vesicular, elongated glass shards and are accurately classified (Figure 15F).
- 596 (4) The ViT accurately classifies ash particles from plinian and subplinian eruptive styles
597 (*macro F1-score* of 0.95), including free crystals (*F1-score* of 0.92), altered material
598 (*F1-score* of 0.93) and juvenile (1.0), but less accurate for lithics (*F1-score* of 0.77).
599 The juvenile particles consist of fragments of pumice and all particles are successfully
600 classified (Figure 15G). In contrast, the lithic particles mostly consist of dull grey
601 fragments with rounded edges, and most of the *FN* are classified as altered material,
602 which may reflect the challenge of classifying particles with incipient weathering into
603 weathered material or lithic (Figure 15H).

604

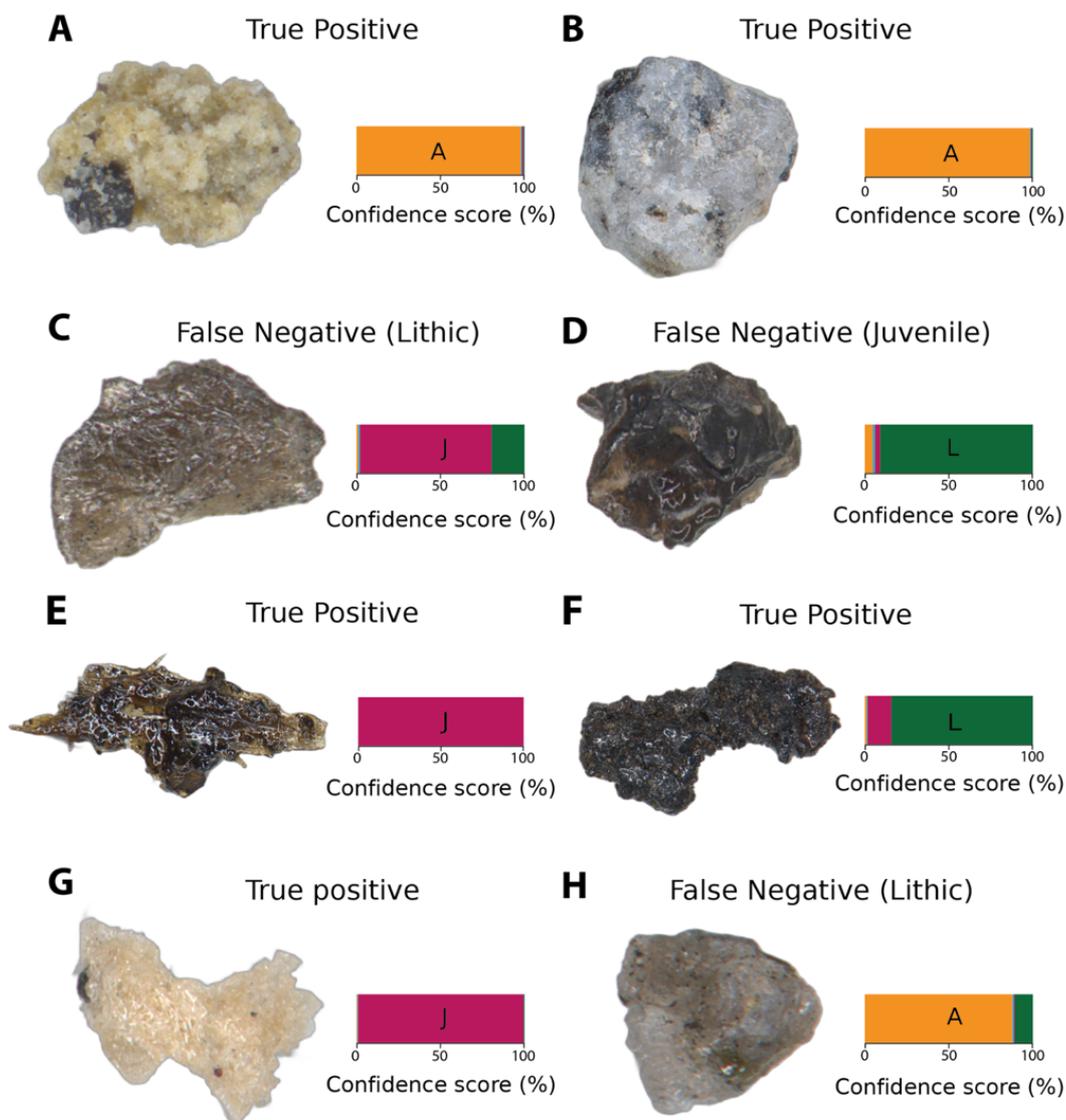
605



606

607 **Figure 14.** (A) Bar charts showing the percentage of predicted types for each particle type in
 608 VolcAshDB. If all predictions were the same as in the database, each bar would be single-
 609 colored as follows: orange for altered material (A), light blue for free-crystal (F), magenta for
 610 juvenile (J), and dark green for lithic (L). (B) shows the *F1-score* for each particle type across
 611 eruptive styles, whereas (C) shows the value of the *macro F1-score* per eruptive style. Note the
 612 range in *macro F1-score* values (C) from 0.85 for dome explosion to 0.91 for lava fountaining up

613 to 0.95 for phreatic, subplinian and plinian eruptive styles. The exact values of this figure can be
 614 found in Table S9.



615

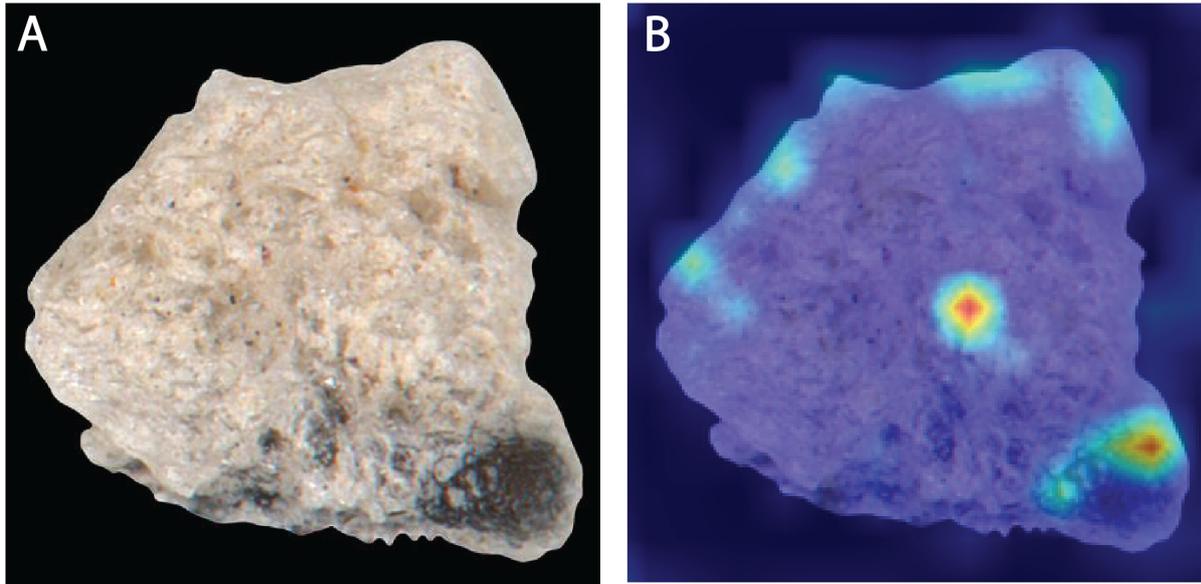
616 **Figure 15.** Representative examples of particle images and the predictions and their associated
 617 confidence score across eruptive styles, including phreatic (A,B), dome explosion (C,D), lava
 618 fountaining (E,F), and subplinian/plinian (G,H). Note that False Negatives contain in brackets
 619 the particle type according to VolcAshDB, and that color code is the same as in previous figure.

620 4 Discussion

621 4.1 Comparison between classification using particle's features versus images

622 We found that, overall, the ViT classifies more accurately with particle images (0.93 of
 623 *macro F1-score*) than the XGBoost classifies with the particle features (0.77 of *macro F1-score*).
 624 This difference is unlikely to be the XGBoost model itself, which is very popular in the literature
 625 and has had best performances amongst models for complex classification tasks (Brownlee,

626 2016; Chen & Guestrin, 2016; Dhaliwal et al., 2018). One possibility is that the extracted
 627 features don't retain certain discriminant information from the images, and as a result, the
 628 XGBoost is unable to classify particles such as free crystals (0.57 of *F1-score*). On the other
 629 hand, maintaining the physical information associated with features makes the model's outcomes
 630 more interpretable (e.g., in classification of volcano-seismic signals; Falcin et al., 2021; Malfante
 631 et al., 2018) with xAI methods. This is an important advantage over Vision Transformers, whose
 632 main xAI tool consists in a heatmap of the region(s) of attention by the model (Dosovitskiy et al.,
 633 2020) but appears insufficient to obtain well founded classification insights for ash particles
 634 (Figure 16).



635
 636 **Figure 16.** Example of (A) one multi-focused binocular image of a pumice particle from Mount
 637 St. Helens (1980), which is overlain by (B) a heatmap of the regions of attention by the base
 638 Vision Transformer (Dosovitskiy et al., 2020), typically used for interpreting image classifier's
 639 predictions. It does not appear easy to discern which aspects of the particle were relevant for
 640 classification.

641 4.2 Insights from XGBoost to better develop a classification criterion for the particles
 642 observed with the binocular

643 The XGBoost model gave a medium to high classification performance with *macro F1-*
 644 *score* of 0.77, and using the Shapley values we identified the most discriminant features of each
 645 particle type (Table 4). For instance, lithic particles can be distinguished with low values of
 646 *value_mode* which correspond to the luster of the particle according to the high Shapley values.
 647 This finding agrees with previous studies that use a dull luster (which corresponds to low values
 648 of *value_mode*) to identify lithic particles (Miwa et al., 2013). On the other hand, juvenile
 649 particles have high Shapley values for the *saturation_mode*. This feature is related to high color
 650 intensities as observed under the binocular, but it was not recognized before as a diagnostic
 651 observation of the particle type. These two examples belong to particle types that are well
 652 classified and for which the Shapley values are reliable. Shapley values obtained from particles
 653 that yielded lower accuracies, such as the free crystals, are not reliable, and thus overall
 654 performances should be improved. This could be achieved by enhancing the quality and quantity

655 of VolcAshDB dataset by (i) adding particles to balance the dataset, (ii) refining the particle
656 contour in the multi-focused images, so that shape features can measure micro-scaled cavities
657 (Benet et al., *preprint*), and (iii) the inclusion of a new feature that measures the density of lines
658 on the surface, which could be sensitive to planar structures of free crystals.

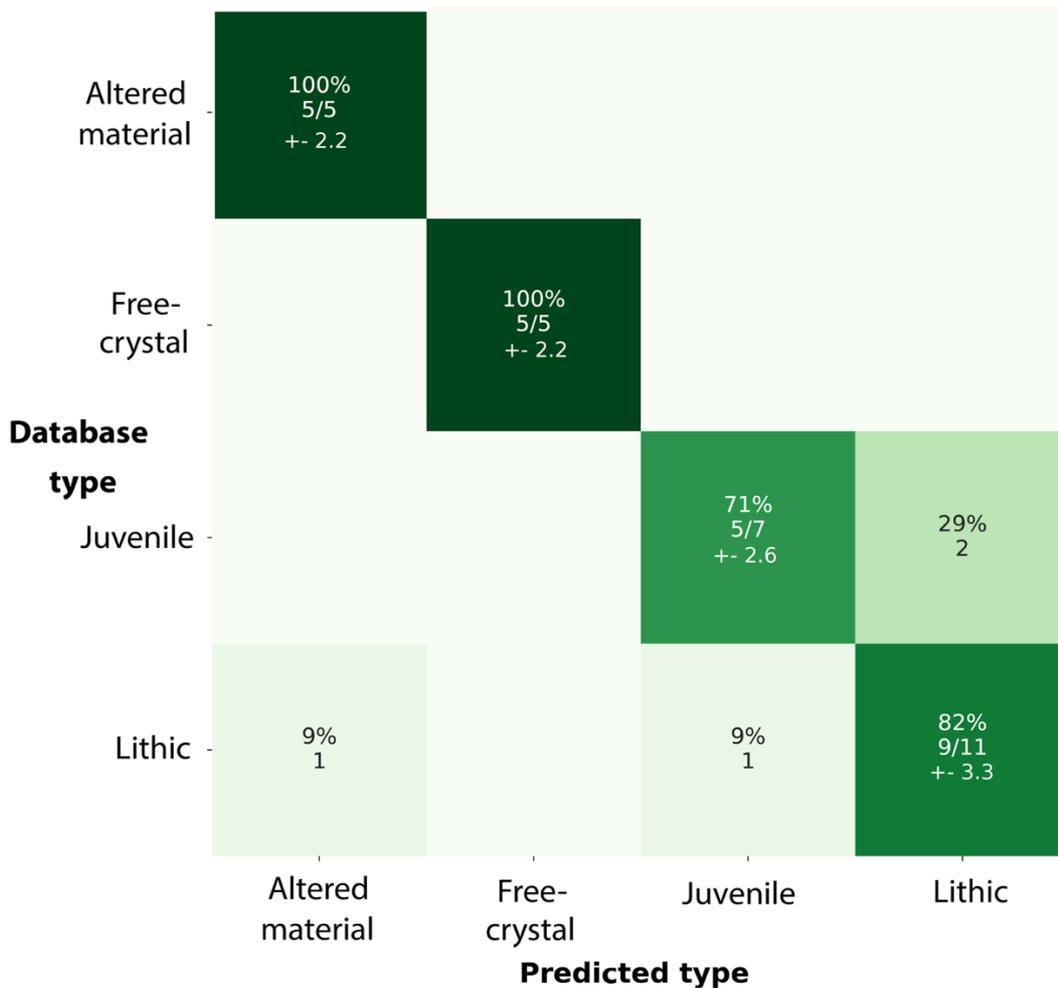
659 4.3 Deploying the ViT for automatic particle classification

660 A main goal of our research is to obtain a classifier of ash particles that is as accurate as
661 possible, and which can be applied to objectively classify new datasets in a reproducible manner.
662 The ViT model (*macro F1-score* of 0.93) currently performs very accurately for some samples
663 (e.g., Soufrière de Guadeloupe; *macro F1-score* of 0.95) but is less accurate for others (e.g.,
664 Merapi; *macro F1-score* of 0.80). This variation is also found within subgroups of particles. For
665 instance, elongated, highly-vesicular, glossy particles from basaltic lava fountaining (Cumbre
666 Vieja, 2021) or pumice fragments (Kelud, 2014) are very accurately classified, but high
667 crystallinity, blocky, dark particles from dome explosions (Nevados de Chillán, 2016–2018) are
668 less accurately classified. These changes in classification scores may be due to differences in the
669 particle-forming processes: juvenile particles from Plinian eruptions are originated from a main
670 and short fragmentation episode, whereas juvenile particles from dome explosions originate from
671 magma with a long and complex story of slow conduit ascent, degassing, crystallization,
672 fracturing, and recycling. Moreover, the variability of *F1-scores* between eruptive styles suggests
673 that to obtain a more robust model for generalization, we need more particles from such
674 problematic subgroups and labelling done by a team of experts. We will also increase our range
675 of samples, including eruptive styles like strombolian activity, submarine eruptions, phreatic
676 from water-lake interaction, and andesitic magma compositions, amongst the most important.

677

678 4.4 A ViT particle classifier for volcano monitoring

679 From an operational viewpoint, volcano observatories and laboratories are often equipped
680 with binocular microscopes that can acquire standard, single-focus binocular images, and that are
681 used to classifying ash (componentry analysis). This could be done near-real time, and it usually
682 takes from one to a few days (Re et al., 2021), or it could also be done a posteriori to obtain a
683 time series data of ash componentry that can be compared to other monitoring data to better
684 understand how the volcanic system works (Benet et al., 2021; Suzuki et al., 2013). Our dataset
685 and analysis are based on multi-focused images and therefore, we performed a preliminary test
686 of ViT's ability to classify single-focus images from a small dataset of ~1,200 images from
687 Nevados de Chillán (Benet et al., 2021). The dataset contains images of about 400 particles, with
688 3 images per particle at different focus depths. Since using the same split ratio (80:20) would
689 yield very small training set, we used all particles for training, except 28 representative particles
690 of the types of ash as described in Benet et al. (2021) as test. Fine-tuning the ViT took only 3
691 hours and we obtained decent accuracies (*macro F1-score* of 0.84) on the test set (Figure 17).
692 This suggests that volcano observatories could potentially use a ViT and obtain an objective
693 score on a particle-by-particle basis relatively rapidly.



694

695 **Figure 17.** Confusion matrix of the predictions by the ViT image classifier after being fine-tuned
 696 with a single-focused, small training set (~370 particles from Benet et al., 2021). The
 697 percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and
 698 otherwise, the False Negative rate (lighter), all percentages with the corresponding number of
 699 particles per predicted type. Note that given the limited data we used all particles for training
 700 except 28 for the test set. Since the subset is small, we report an error as the square root of the
 701 number of particles, which is known in statistics as the implicit random error (Ahmed, 2015).

702 5 Conclusions

703 Classification of the different particles that make up volcanic ash is not straightforward
 704 because diagnostic criteria are not standardized and thus reliable, and systematic identification of
 705 a given particle type is not straightforward. In this contribution, we attempt to alleviate this
 706 situation by exploring the use of state-of-the-art machine learning-based models to identify the
 707 most discriminant features of each particle type, and to evaluate their ability to classify particles.
 708 The identified features provide new insights on the recognition of juvenile and lithic particles
 709 towards a standardized classification. The image classifier performs at very high accuracies,
 710 although the variability across eruption and types shows that its capability to generalize to new
 711 samples is still unclear. Higher numbers of particles from a wider variety of eruptions and

712 volcanoes into VolcAshDB coupled to ML models should allow for unbiased comparison of ash
 713 samples, and reproducible classification of their particles as a tool for volcano monitoring
 714 studies.

715 Acknowledgments

716 I am grateful to Caroline Bouvet de Maisonneuve, John Pallister, Jacopo Taddeucci and
 717 Alison Rust for insightful discussions, to Do Xuan Long for help on the use of the Hugging Face
 718 API for image classification, to Sébastien Biass for advice and help on the use of the SHAP
 719 method for this study, and to Edwin Tan for support on the Gekko cluster. This research was
 720 supported by the Earth Observatory of Singapore via its funding from the National Research
 721 Foundation Singapore and the Singapore Ministry of Education under the Research Centres of
 722 Excellence initiative.

723 724 Open Research

725 Particle images and features can be downloaded through the publicly available
 726 VolcAshDB web database at <https://volcash.wovodat.org/>. Details on the feature measurement
 727 and image acquisition are described in Benet et al., *preprint*. The GitHub repository
 728 https://github.com/dbenet-max/volcashdb_classification contains two relevant codes: the Python
 729 code for hyperparameter optimization, development, and interpretation via xAI of the XGBoost,
 730 and the code for deployment via the API Hugging Face of the ViT.

731 732 References

- 733 Ahmed, S. N. (2015). Essential statistics for data analysis. In *Physics and Engineering of*
 734 *Radiation Detection*. <https://doi.org/10.1016/b978-0-12-801363-2.00009-7>
- 735 Alvarado, G. E., Mele, D., Dellino, P., de Moor, J. M., & Avaró, G. (2016). Are the ashes from
 736 the latest eruptions (2010–2016) at Turrialba volcano (Costa Rica) related to phreatic or
 737 phreatomagmatic events? *Journal of Volcanology and Geothermal Research*, 327, 407–415.
 738 <https://doi.org/10.1016/j.jvolgeores.2016.09.003>
- 739 Ayyadevara, V. K., & Reddy, Y. (2020). *Modern Computer Vision with PyTorch: Explore deep*
 740 *learning concepts and implement over 50 real-world image applications*. Packt Publishing
 741 Ltd.
- 742 Bebbington, M. S., & Jenkins, S. F. (2019). *Intra-eruption forecasting*.
- 743 Benet, D., Costa, F., Pedreros, G., & Cardona, C. (2021). The volcanic ash record of shallow
 744 magma intrusion and dome emplacement at Nevados de Chillán Volcanic complex, Chile.
 745 *Journal of Volcanology and Geothermal Research*, 417.
 746 <https://doi.org/10.1016/j.jvolgeores.2021.107308>
- 747 Benet, D., Costa, F., Widiwijayanti, C., Pallister, J., Pedreros, G., Allard, P., Humaida, H., Aoki,
 748 Y., & Maeno, F. (2023). VolcAshDB: Volcanic ash particle image and classification
 749 database. *Preprint in EarthArxiv*. <https://doi.org/10.31223/X53659>

- 750 Biass, S., Jenkins, S. F., Aeberhard, W. H., Delmelle, P., & Wilson, T. (2022). Insights into the
751 vulnerability of vegetation to tephra fallouts from interpretable machine learning and big
752 Earth observation data. *Natural Hazards and Earth System Sciences*, 22(9), 2829–2855.
753 <https://doi.org/10.5194/nhess-22-2829-2022>
- 754 Brownlee, J. (2016). XGBoost With python: Gradient boosted trees with XGBoost and scikit-
755 learn. *Machine Learning Mastery*.
- 756 Brownlee, J. (2020). Imbalanced Classification with Python. *Machine Learning Mastery*, 463.
- 757 Cashman, K. V., & Hoblitt, R. P. (2004). Magmatic precursors to the 18 May 1980 eruption of
758 Mount St. Helens, USA. *Geology*, 32(2), 141–144. <https://doi.org/10.1130/G20078.1>
- 759 Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the*
760 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-
761 17-Aug, 785–794. <https://doi.org/10.1145/2939672.2939785>
- 762 Cioni, R., Pistolesi, M., Bertagnini, A., Bonadonna, C., Hoskuldsson, A., & Scatani, B. (2014).
763 Insights into the dynamics and evolution of the 2010 Eyjafjallajökull summit eruption
764 (Iceland) provided by volcanic ash textures. *Earth and Planetary Science Letters*, 394(May
765 2010), 111–123. <https://doi.org/10.1016/j.epsl.2014.02.051>
- 766 D’Oriano, C., Bertagnini, A., Cioni, R., & Pompilio, M. (2014). Identifying recycled ash in
767 basaltic eruptions. *Scientific Reports*, 4. <https://doi.org/10.1038/srep05851>
- 768 D’Oriano, C., Del Carlo, P., Andronico, D., Cioni, R., Gabellini, P., Cristaldi, A., & Pompilio,
769 M. (2022). Syn-Eruptive Processes During the January–February 2019 Ash-Rich Emissions
770 Cycle at Mt. Etna (Italy): Implications for Petrological Monitoring of Volcanic Ash.
771 *Frontiers in Earth Science*, 10(February 2019). <https://doi.org/10.3389/feart.2022.824872>
- 772 Dellino, P., & La Volpe, L. (1996). Image processing analysis in reconstructing fragmentation
773 and transportation mechanisms of pyroclastic deposits. The case of Monte Pilato-Rocche
774 Rosse eruptions, Lipari (Aeolian islands, Italy). *Journal of Volcanology and Geothermal*
775 *Research*, 71(1), 13–29. [https://doi.org/10.1016/0377-0273\(95\)00062-3](https://doi.org/10.1016/0377-0273(95)00062-3)
- 776 Dhaliwal, S. S., Nahid, A. Al, & Abbas, R. (2018). Effective intrusion detection system using
777 XGBoost. *Information (Switzerland)*, 9(7). <https://doi.org/10.3390/info9070149>
- 778 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.,
779 Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020).
780 *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.
- 781 Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning
782 and Stochastic Optimization. *Journal of Machine Learning Research*, 12, 2121–2159.
- 783 Dürig, T., Bowman, M. H., White, J. D. L., Murch, A., Mele, D., Verolino, A., & Dellino, P.
784 (2018). Particle shape analyzer Partisan - An open source tool for multi-standard two-
785 dimensional particle morphometry analysis. *Annals of Geophysics*, 61(6).
786 <https://doi.org/10.4401/ag-7865>
- 787 Dürig, T., Ross, P. S., Dellino, P., White, J. D. L., Mele, D., & Comida, P. P. (2021). A review of
788 statistical tools for morphometric analysis of juvenile pyroclasts. *Bulletin of Volcanology*,
789 83(11). <https://doi.org/10.1007/s00445-021-01500-0>

- 790 Falcin, A., Métaxian, J. P., Mars, J., Stutzmann, É., Komorowski, J. C., Moretti, R., Malfante,
791 M., Beauducel, F., Saurel, J. M., Dessert, C., Burtin, A., Ucciani, G., de Chabalière, J. B., &
792 Lemarchand, A. (2021). A machine-learning approach for automatic classification of
793 volcanic seismicity at La Soufrière Volcano, Guadeloupe. *Journal of Volcanology and*
794 *Geothermal Research*, 411. <https://doi.org/10.1016/j.jvolgeores.2020.107151>
- 795 Feuillard, M., Allegre, C. J., Brandeis, G., Gaulon, R., Le Mouel, J. ., Mercier, J. C., Pozzi, J. P.,
796 & Semet, M. . (1983). The 1975–1977 crisis of la Soufriere de Guadeloupe (F.W.I): A still-
797 born magmatic eruption. *Journal of Volcanology and Geothermal Research*, 16, 317–334.
- 798 Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., & Brinker, K. (2008). Multilabel classification
799 via calibrated label ranking. *Machine Learning*, 73(2), 133–153.
800 <https://doi.org/10.1007/s10994-008-5064-8>
- 801 Gaunt, H. E., Bernard, B., Hidalgo, S., Proaño, A., Wright, H., Mothes, P., Criollo, E., &
802 Kueppers, U. (2016). Juvenile magma recognition and eruptive dynamics inferred from the
803 analysis of ash time series: The 2015 reawakening of Cotopaxi volcano. *Journal of*
804 *Volcanology and Geothermal Research*, 328, 134–146.
805 <https://doi.org/10.1016/j.jvolgeores.2016.10.013>
- 806 Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts,
807 tools, and techniques to build intelligent systems. In *Hands-on machine learning with*
808 *Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*.
- 809 Gianfagna, L., & Di Cecco, A. (2021). Explainable AI. In *Berlin/Heidelberg, Germany:*
810 *Springer*. <https://doi.org/10.3233/FAIA190100>
- 811 Hall-Beyer, M. (2017). GLCM Texture: A Tutorial. *17th International Symposium on Ballistics*,
812 2(March), 18–19.
- 813 Haralick, R. M., Dinstein, I., & Shanmugam, K. (1973). Textural Features for Image
814 Classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6), 610–621.
815 <https://doi.org/10.1109/TSMC.1973.4309314>
- 816 He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-
817 level performance on imagenet classification. *Proceedings of the IEEE International*
818 *Conference on Computer Vision*, 1026–1034.
- 819 He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition.
820 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern*
821 *Recognition, 2016-Decem*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- 822 Herrera, F., Charte, F., Rivera, A. J., & del Jesus, M. J. (2016). *Multi-Label Classification*
823 (Springer,). <https://doi.org/10.4018/jdwm.2007070101>
- 824 Hincks, T. K., Komorowski, J. C., Sparks, S. R., & Aspinall, W. P. (2014). Retrospective
825 analysis of uncertain eruption precursors at La Soufrière volcano, Guadeloupe, 1975-77:
826 Volcanic hazard assessment using a Bayesian Belief Network approach. *Journal of Applied*
827 *Volcanology*, 3(1). <https://doi.org/10.1186/2191-5040-3-3>
- 828 Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). *ImageNet: A large-scale*
829 *hierarchical image database*. 248–255. <https://doi.org/10.1109/cvprw.2009.5206848>
- 830 Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-

- 831 Torres, M. Á., & Carvalhais, N. (2022). Wildfire Danger Prediction and Understanding
 832 With Deep Learning. *Geophysical Research Letters*, *49*(17), 1–11.
 833 <https://doi.org/10.1029/2022GL099368>
- 834 Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4),
 835 261–283. <https://doi.org/10.1007/s10462-011-9272-4>
- 836 Le Guern, F., Bernard, A., & Chevrier, R. M. (1980). Soufrière of guadeloupe 1976–1977
 837 eruption — mass and energy transfer and volcanic health hazards. *Bulletin Volcanologique*,
 838 *43*(3), 577–593. <https://doi.org/10.1007/BF02597694>
- 839 Lee, J. J., Aime, M. C., Rajwa, B., & Bae, E. (2022). Machine Learning-Based Classification of
 840 Mushrooms Using a Smartphone Application. *Applied Sciences (Switzerland)*, *12*(22).
 841 <https://doi.org/10.3390/app122211685>
- 842 Leibrandt, S., & Le Penec, J. L. (2015). Towards fast and routine analyses of volcanic ash
 843 morphometry for eruption surveillance applications. *Journal of Volcanology and*
 844 *Geothermal Research*, *297*, 11–27. <https://doi.org/10.1016/j.jvolgeores.2015.03.014>
- 845 Liu, E. J., Cashman, K. V., Miller, E., Moore, H., Edmonds, M., Kunz, B. E., Jenner, F., &
 846 Chigna, G. (2020). Petrologic monitoring at Volcán de Fuego, Guatemala. *Journal of*
 847 *Volcanology and Geothermal Research*, *405*(August 2019), 107044.
 848 <https://doi.org/10.1016/j.jvolgeores.2020.107044>
- 849 Liu, E. J., Cashman, K. V., & Rust, A. C. (2015). Optimising shape analysis to quantify volcanic
 850 ash morphology. *GeoResJ*, *8*, 14–30. <https://doi.org/10.1016/j.grj.2015.09.001>
- 851 Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the*
 852 *2020s*. 11966–11976. <https://doi.org/10.1109/cvpr52688.2022.01167>
- 853 Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *7th International*
 854 *Conference on Learning Representations, ICLR 2019*.
- 855 Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). *Consistent Individualized Feature Attribution*
 856 *for Tree Ensembles*. 2. <http://arxiv.org/abs/1802.03888>
- 857 Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions.
 858 *Advances in Neural Information Processing Systems*, *30*.
- 859 Maeno, F., Nakada, S., Yoshimoto, M., Shimano, T., Hokanishi, N., Zaennudin, A., & Iguchi, M.
 860 (2019). A sequence of a plinian eruption preceded by dome destruction at Kelud volcano,
 861 Indonesia, on February 13, 2014, revealed from tephra fallout and pyroclastic density
 862 current deposits. *Journal of Volcanology and Geothermal Research*, *382*, 24–41.
 863 <https://doi.org/10.1016/j.jvolgeores.2017.03.002>
- 864 Malfante, M., Mura, M. D., Métaxian, J., & Mars, J. I. (2018). *Machine Learning for Volcano-*
 865 *Seismic Signals*. *March*, 20–30.
- 866 Mandal, S., Mones, S. M. B., Das, A., Balas, V. E., Shaw, R. N., & Ghosh, A. (2021). Single
 867 shot detection for detecting real-time flying objects for unmanned aerial vehicle. In
 868 *Artificial Intelligence for Future Generation Robotics*. INC. [https://doi.org/10.1016/B978-](https://doi.org/10.1016/B978-0-323-85498-6.00005-8)
 869 [0-323-85498-6.00005-8](https://doi.org/10.1016/B978-0-323-85498-6.00005-8)
- 870 Marzocchi, W., Newhall, C., & Woo, G. (2012). The scientific management of volcanic crises.

- 871 *Journal of Volcanology and Geothermal Research*, 247–248, 181–189.
872 <https://doi.org/10.1016/j.jvolgeores.2012.08.016>
- 873 Mishra, P. (2022). Practical Explainable AI Using Python. In *Practical Explainable AI Using*
874 *Python*. <https://doi.org/10.1007/978-1-4842-7158-2>
- 875 Miwa, T., Geshi, N., & Shinohara, H. (2013). Temporal variation in volcanic ash texture during a
876 vulcanian eruption at the sakurajima volcano, Japan. *Journal of Volcanology and*
877 *Geothermal Research*, 260, 80–89. <https://doi.org/10.1016/j.jvolgeores.2013.05.010>
- 878 Miwa, T., Toramaru, A., & Iguchi, M. (2009). Correlations of volcanic ash texture with
879 explosion earthquakes from vulcanian eruptions at Sakurajima volcano, Japan. *Journal of*
880 *Volcanology and Geothermal Research*, 184(3–4), 473–486.
881 <https://doi.org/10.1016/j.jvolgeores.2009.05.012>
- 882 Miyagi, I., Geshi, N., Hamasaki, S., Oikawa, T., & Tomiya, A. (2020). Heat source of the 2014
883 phreatic eruption of Mount Ontake, Japan. *Bulletin of Volcanology*, 82(4).
884 <https://doi.org/10.1007/s00445-020-1358-x>
- 885 Molnar, C. (2021). Interpretable Machine Learning. *Queue*, 19(6), 28–56.
886 <https://doi.org/10.1145/3511299>
- 887 Moran, S. C., Newhall, C., & Roman, D. C. (2011). Failed magmatic eruptions: Late-stage
888 cessation of magma ascent. *Bulletin of Volcanology*, 73(2), 115–122.
889 <https://doi.org/10.1007/s00445-010-0444-x>
- 890 Newhall, C. G., & Punongbayan, R. S. (1996). The narrow margin of successful volcanic-risk
891 mitigation. In *Monitoring and mitigation of volcano hazards* (pp. 807–838). Springer
892 Science & Business Media.
- 893 Nurfiani, D., & Bouvet de Maisonneuve, C. (2018). Furthering the investigation of eruption
894 styles through quantitative shape analyses of volcanic ash particles. *Journal of Volcanology*
895 *and Geothermal Research*, 354, 102–114. <https://doi.org/10.1016/j.jvolgeores.2017.12.001>
- 896 Owen, L. (2022). *Hyperparameter Tuning with Python*.
- 897 Paladio-Melasantos, M. L., Solidum, R. U., Scott, W. E., Quiambao, R. B., Umbal, J. V,
898 Rodolfo, K. S., Tubianosa, B. S., Delos Reyes, P. J., Alonso, R. A., & Ruerlo, H. B. (1996).
899 Tephra falls of the 1991 eruptions of Mount Pinatubo. In: *Newhall, C.G. (Editor) & Others,*
900 *Fire and Mud; Eruptions and Lahars of Mount Pinatubo, Philippines, Philippine Institute of*
901 *Volcanology and Seismology, Quezon City, layer D*, 413–535.
902 <https://doi.org/10.1159/000153100>
- 903 Panati, C., Wagner, S., & Bruggenwirth, S. (2022). Feature Relevance Evaluation using Grad-
904 CAM, LIME and SHAP for Deep Learning SAR Data Classification. *Proceedings*
905 *International Radar Symposium, 2022-Septe*, 457–462.
- 906 Pardo, N., Avellaneda, J. D., Rausch, J., Jaramillo-Vogel, D., Gutiérrez, M., & Foubert, A.
907 (2020). Decrypting silicic magma/plug fragmentation at Azufral crater lake, Northern
908 Andes: insights from fine to extremely fine ash morpho-chemistry. *Bulletin of Volcanology*,
909 82(12). <https://doi.org/10.1007/s00445-020-01418-z>
- 910 Pardo, N., Cronin, S. J., Németh, K., Brenna, M., Schipper, C. I., Breard, E., White, J. D. L.,
911 Procter, J., Stewart, B., Agustín-Flores, J., Moebis, A., Zernack, A., Kereszturi, G., Lube,

- 912 G., Auer, A., Neall, V., & Wallace, C. (2014). Perils in distinguishing phreatic from
 913 phreatomagmatic ash; insights into the eruption mechanisms of the 6 August 2012 Mt.
 914 Tongariro eruption, New Zealand. *Journal of Volcanology and Geothermal Research*, 286,
 915 397–414. <https://doi.org/10.1016/j.jvolgeores.2014.05.001>
- 916 Re, G., Corsaro, R. A., D’Oriano, C., & Pompilio, M. (2021). Petrological monitoring of active
 917 volcanoes: A review of existing procedures to achieve best practices and operative protocols
 918 during eruptions. *Journal of Volcanology and Geothermal Research*, 419, 107365.
 919 <https://doi.org/10.1016/j.jvolgeores.2021.107365>
- 920 Romero, J. E., Burton, M., Cáceres, F., Taddeucci, J., Civico, R., Ricci, T., Pankhurst, M. J.,
 921 Hernández, P. A., Bonadonna, C., Llewellyn, E. W., & Pistolesi, M. (2022). *The initial*
 922 *phase of the 2021 Cumbre Vieja ridge eruption (Canary Islands): Products and dynamics*
 923 *controlling edifice growth and collapse*. 431(July).
 924 <https://doi.org/10.1016/j.jvolgeores.2022.107642>
- 925 Ross, P. S., Dürig, T., Comida, P. P., Lefebvre, N., White, J. D. L., Andronico, D., Thivet, S.,
 926 Eychenne, J., & Gurioli, L. (2022). Standardized analysis of juvenile pyroclasts in
 927 comparative studies of primary magma fragmentation; 1. Overview and workflow. *Bulletin*
 928 *of Volcanology*, 84(1), 1–29. <https://doi.org/10.1007/s00445-021-01516-6>
- 929 Rowe, M. C., Thornber, C. R., & Kent, A. J. R. (2008). Identification and Evolution of the
 930 Juvenile Component in. *A Volcano Rekindled: The Renewed Eruption of Mount St. Helens,*
 931 *2004–2006, 2004–2006.*
- 932 Shapley, L. S. (1953). A Value for n-Person Games, in: Contributions to the Theory of Games II.
 933 *Contributions to the Theory of Games*, 307–318.
 934 <https://doi.org/https://doi.org/10.1515/9781400881970-018>
- 935 Shoji, D., Noguchi, R., Otsuki, S., & Hino, H. (2018). *Classification of volcanic ash particles*
 936 *using a convolutional neural network and probability*. 1–12.
 937 <https://doi.org/10.1038/s41598-018-26200-2>
- 938 Sujatha, R., Chatterjee, J. M., Jhanjhi, N. Z., & Brohi, S. N. (2021). Performance of deep
 939 learning vs machine learning in plant leaf disease detection. *Microprocessors and*
 940 *Microsystems*, 80(November 2020). <https://doi.org/10.1016/j.micpro.2020.103615>
- 941 Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and
 942 momentum in deep learning. *30th International Conference on Machine Learning, ICML*
 943 *2013, PART 3*, 2176–2184.
- 944 Suzuki, Y., Nagai, M., Maeno, F., Yasuda, A., Hokanishi, N., Shimano, T., Ichihara, M.,
 945 Kaneko, T., & Nakada, S. (2013). Precursory activity and evolution of the 2011 eruption of
 946 Shinmoe-dake in Kirishima volcano-insights from ash samples. *Earth, Planets and Space*,
 947 65(6), 591–607. <https://doi.org/10.5047/eps.2013.02.004>
- 948 Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J. A., Zou, J., & Rubin, D. L.
 949 (2021). Data valuation for medical imaging using Shapley value and application to a large-
 950 scale chest X-ray dataset. *Scientific Reports*, 11(1), 1–9. <https://doi.org/10.1038/s41598-021-87762-2>
- 951
- 952 Tilling, R. ~I. (2008). The critical role of volcano monitoring in risk reduction. *Advances in*
 953 *Geosciences*, 14, 3–11.

- 954 Verdhan, V. (2020). Supervised Learning with Python. In *Supervised Learning with Python*.
955 <https://doi.org/10.1007/978-1-4842-6156-9>
- 956 Watanabe, K., Danhara, T., Watanabe, K., Terai, K., & Yamashita, T. (1999). Juvenile volcanic
957 glass erupted before the appearance of the 1991 lava dome, Unzen volcano, Kyushu, Japan.
958 *Journal of Volcanology and Geothermal Research*, 89(1–4), 113–121.
959 [https://doi.org/10.1016/S0377-0273\(98\)00127-9](https://doi.org/10.1016/S0377-0273(98)00127-9)
- 960