Volcanic ash classification through Machine Learning

Damià Benet[1], Fidel Costa[2], and Christina Widiwijayanti[3]

[1]Institut De Physique Du Globe De Paris
[2]Institut de Physique du Globe de Paris
[3]Nanyang Technological University

September 11, 2023

Abstract

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

Volcanic ash classification through Machine Learning

Damià Benet[1,2,3,†], Fidel Costa[1], Christina Widiwijayanti[2]

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

[2] EOS, Earth Observatory of Singapore, Nanyang Technological University, Singapore.

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

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

Key Points:

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

Abstract

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

1 Introduction

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

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

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

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

2 Materials and Methods

2.1 VolcAshDB dataset

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

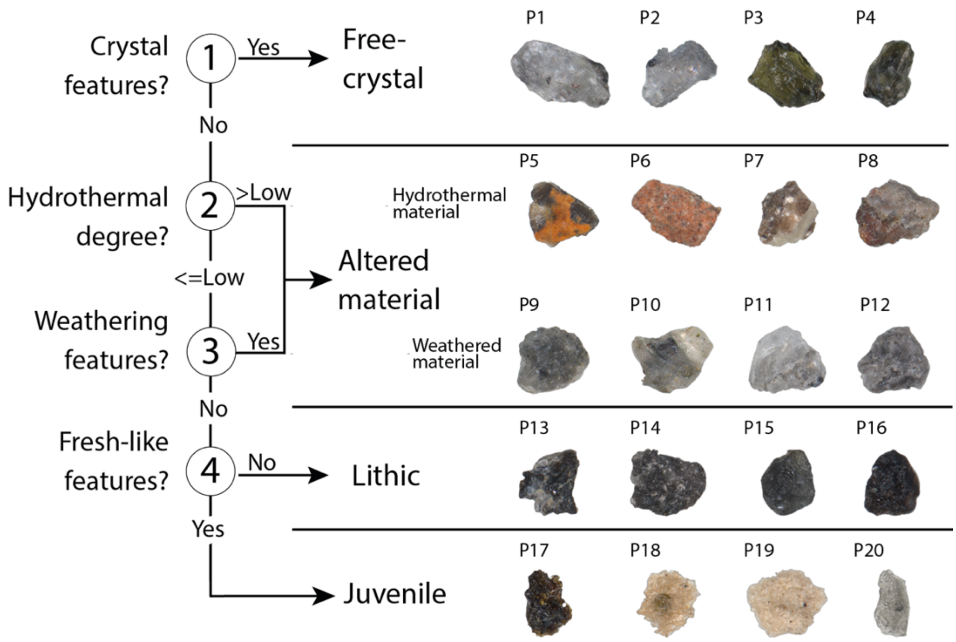

Figure 1. Example of classification process and particle images in VolcAshDB based on the steps for petrographic classification in Benet et al., (*preprint*). Note that the particle type altered material comprises both hydrothermal and weathered material.

Table 1. Main sample characteristics, and proportion of main particle types in VolcAshDB. The associated error is calculated using the equation of margin of error Benet et al., (*preprint*) 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	
Cumbre Vieja									
CV-DB1	19/10/21	Mafic	Cinder cone	Lava fountaining	3 (±0.3)	1(±0.2)	719(±2.8)	352(±1.4)	1075
Kelud									
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
Merapi									
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
Sourfière de Guadeloupe									
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
Nevados de Chillán									
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
Ontake									

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

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

2.2 Development of a particle classifier using the measured particle features

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

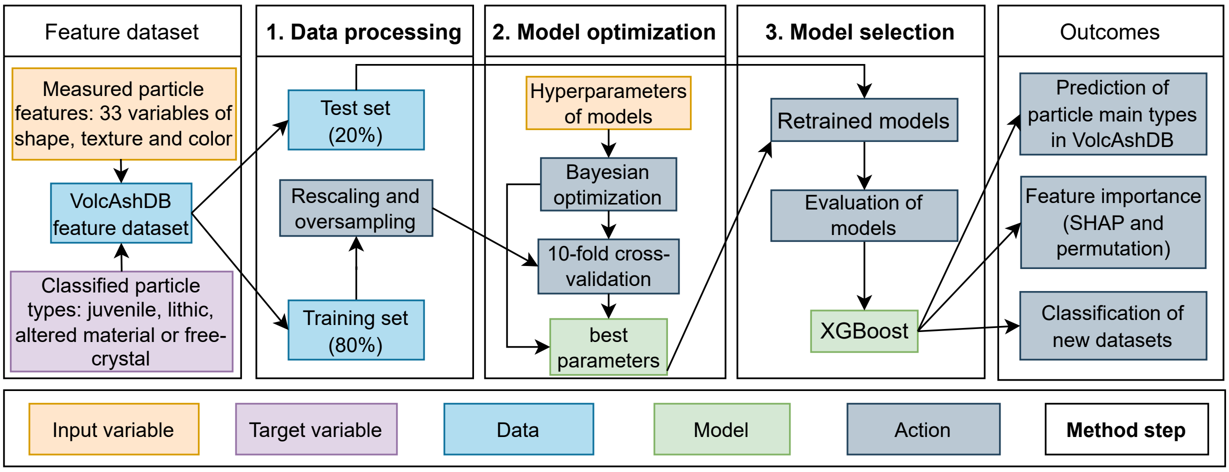

Figure 2. Illustration of the steps involved from the dataset to the outcomes, including those to obtain the best optimized model, XGBoost. (1) Data processing of the full dataset (features and particle types), including the oversampling of the training set. (2) hyperparameter optimization and cross-validation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

2.2.1 Data processing

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

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

2.2.2 Hyperparameter optimization

Hyperparameters control the model learning process and are explicitly defined by the 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

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

2.2.3 Model evaluation and selection

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

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\text{-}score = \frac{2*TP}{2*TP+FP+FN} \tag{4}$$

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

2.2.4 Explaining the model's predictions

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

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

2.2.5 Classification strategies

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

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

As noted above, we first split the dataset into a training (80% of all particle features in VolcAshDB) and a test set (20%) and used the latter to evaluate the XGBoost's performance. However, as splitting process is random it may affect the precision and accuracy of the measured *F1-scores*. To estimate this error, we re-trained and evaluated the XGBoost at 1,000 different values of random state, i.e., the hyperparameter that controls randomness. We obtained an average accuracy (*macro F1-score* of *0.76*; Table S4) that is like the accuracy from the test set (*macro F1-score* of *0.75*). The free-crystal type shows the widest variability (standard deviation of 0.04) and is the most inaccurate (*F1*-score of 0.57; Figure 3) amongst the particle types. This is likely because it is the least abundant type, and its classification is challenging given the different types of minerals and lack of a discriminant feature as explained below (Section 3.1). Accuracies of the three other types are higher (*F1*-score of 0.73–0.88) and with better precision (standard deviation is < 0.02; Table S4).

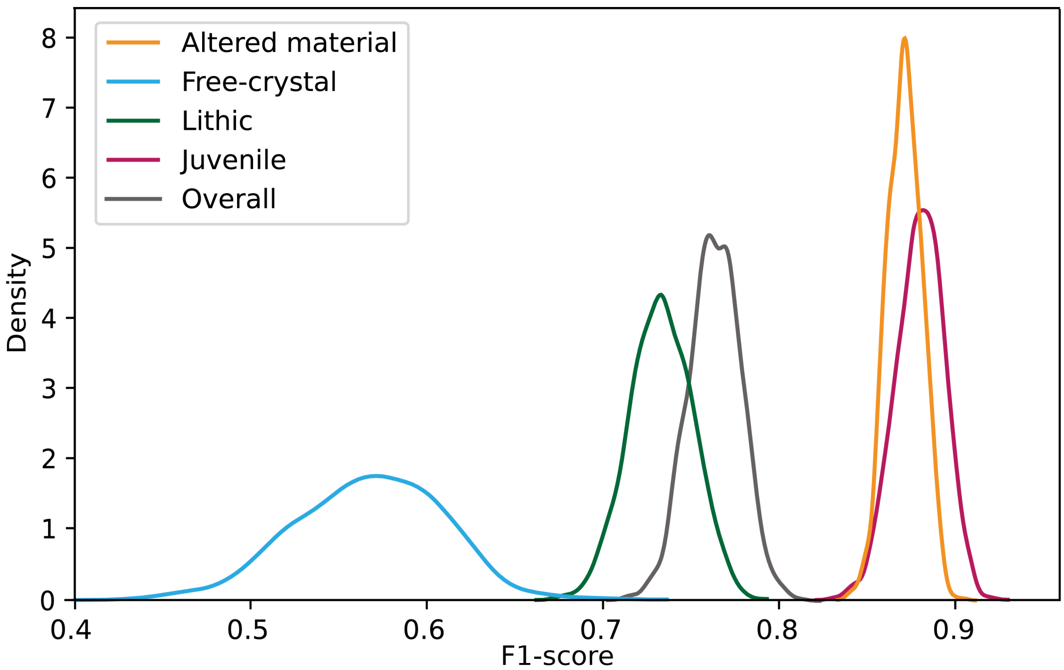

Figure 3. Density plots of the *F1-scores* obtained from 1,000 runs of the XGBoost at different random state across particle types and aggregated as *macro F1-score* (Overall).

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

$$\alpha = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \text{ x2}} \qquad (5)$$

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

2.3 Development of a particle classifier using VolcAshDB image dataset

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

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

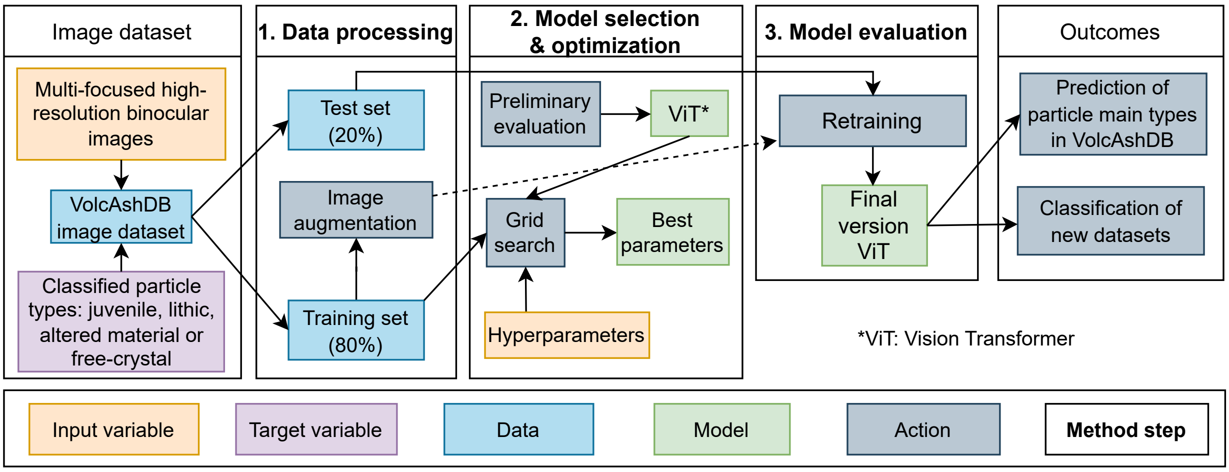

Figure 4. Illustration of the steps involved from the dataset to the outcomes, including those to fine-tune the Vision Transformer (ViT). (1) Data processing of the full dataset (images and particle types). (2) preliminary evaluation of the models using the base hyperparameters, selection of ViT and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

2.3.1 Image augmentation and processing

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

2.3.2 Fine-tuning, preliminary evaluation, and model selection

We fine-tuned the classifiers and did a preliminary round of evaluations to choose the best-performing model. Fine-tuning consists in making small adjustments to an already trained classifier, as opposed to training, where the data drives the model's learning process without any prior exposure. We selected the model before hyperparameter optimization because each run is time consuming (lasting about 14–18 hours) and because the authors of each model already provide the base hyperparameters (Table S5). The fine-tuned model that yielded the highest accuracy is ViT (0.88), followed by ConvNext and ResNet, both with an accuracy of 0.86.

2.3.3 ViT Hyperparameter optimization

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

2.3.4 Model evaluation

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

3 Results

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

Table 2. *F1-score* values for the whole database (unweighted average or *macro*) and particle types obtained from various models, including XGBoost multilabel, One-vs-One (OVO), One-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

3.1 XGBoost quantitative evaluation

Overall, the XGBoost shows rather accurate *F1-scores* across classification strategies: 0.76 for multilabel, 0.75 for OVO, and 0.76 for OVR (Table 2). Computation of the confusion matrix (Figure 5) shows that the model classifies best the altered material type (*F1-score* of 0.9), closely followed by the juvenile type (*F1-score* of 0.88), and less accurately the lithic type (*F1-score* of 0.74), and significantly less the free-crystal type (*F1-score* of 0.57).

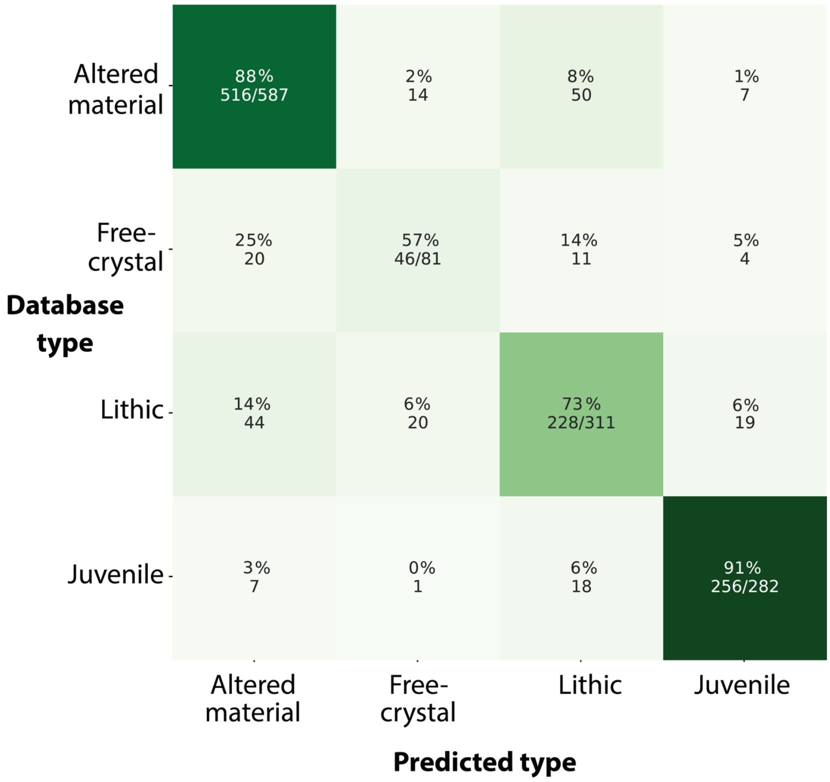

Figure 5. Confusion matrix of the predictions by the XGBoost multilabel classifier. The percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and otherwise, the False Negative rate (lighter), all percentages with the corresponding number of particles per predicted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

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

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

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

	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

[1]*F*: Free-crystal [2]*A*: Altered material [3]*L*: Lithic [4]*J*: Juvenile

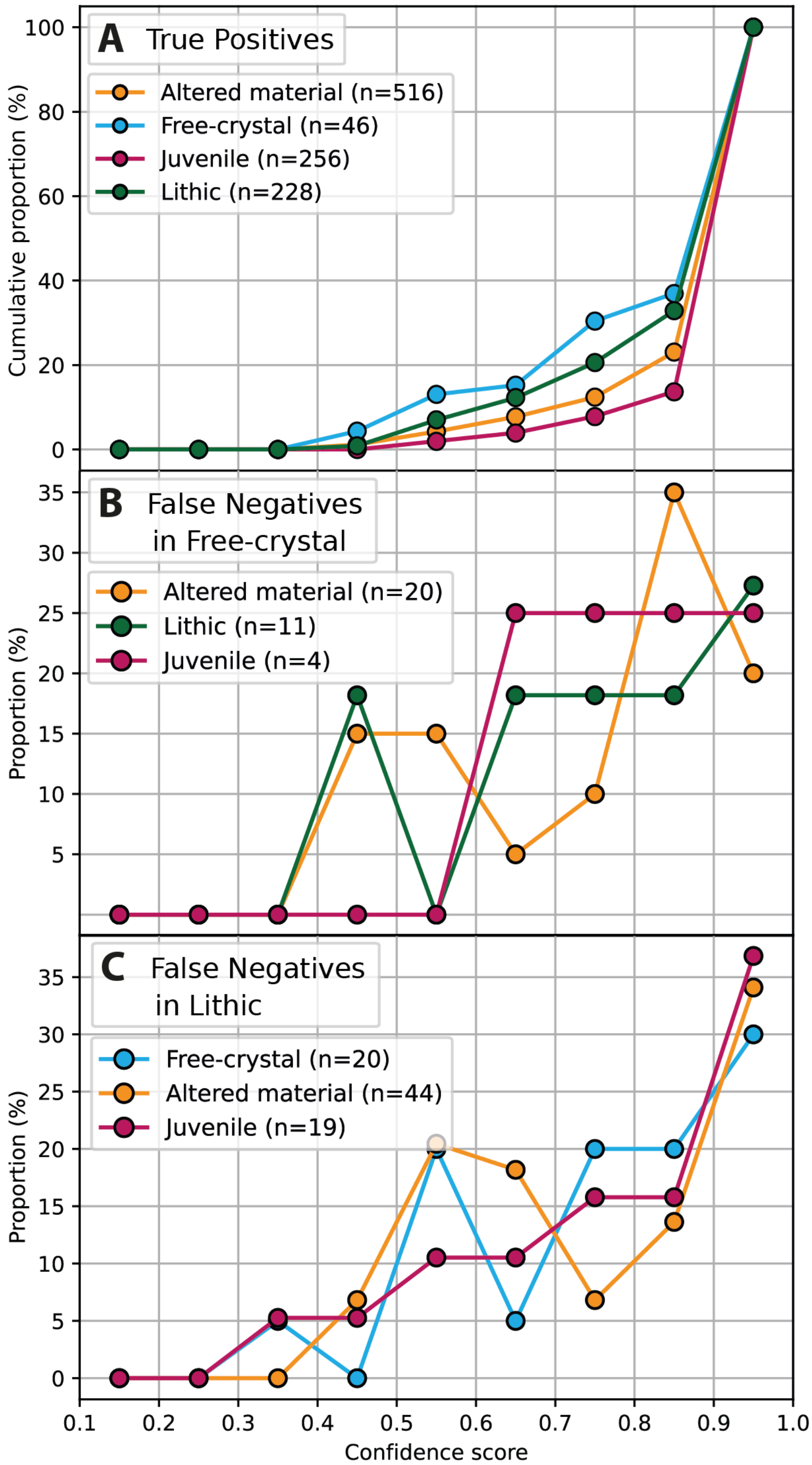

Figure 6. Line plots of the confidence score versus (A) the cumulative proportion of True Positives (TP), (B) False Negatives (FN) in free-crystal, and (C) lithic types. The distribution of the data have been plotted into 9 bins of size 0.1. We don't use cumulative proportion in (B) and (C) given the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

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

3.2 What features drive XGBoost ash particle type predictions?

3.2.1 Global feature importance

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

Table 4. Feature importance identification based on mean of Shapley values and feature permutation. These two methods calculate the feature importance values differently and can't be directly compared. The relative ranking of the features importance is similar (top ten ranked features in bold) with the same top two ranked features (*hue_mean* and *correlation*). We used the Shapley mean value for feature importance per particle type (shown as a plot in Figure 7), the top three of which are underlined. For the meaning of the abbreviations of each feature please see Table S1. The permutation feature values have been multiplied by ten for better readability, as the importance lies on the relative values across 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	**0.78**	**0.86**	**0.12**	**1.15**	**2.91**	**0.91**	**0.41**	**0.15**	**0.91**	**1.22**
correlation	**0.46**	**0.33**	**0.33**	**0.55**	**1.68**	**0.34**	0.02	**0.19**	**0.04**	**0.29**
blue_mode	**0.31**	0.10	**0.48**	**0.54**	**1.43**	0.06	**0.04**	0.00	**0.01**	**0.10**
value_mode	**0.28**	**0.23**	**0.60**	0.20	**1.31**	0.05	**0.05**	**0.24**	0.00	0.00
saturation_mode	**0.10**	**0.27**	-0.01	**0.80**	**1.17**	0.02	**0.06**	**0.10**	**0.10**	**0.13**
convexity	0.02	**0.52**	0.06	**0.48**	**1.10**	0.01	**0.06**	0.00	**0.03**	0.03
red_mean	**0.16**	0.18	**0.53**	**0.21**	**1.07**	0.03	0.03	0.01	**0.01**	0.04
blue_std	-0.06	**0.81**	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	

3.2.2 Local feature importance across particle types

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

(1) Altered material has the highest classification success with a *F1-score* of 0.90 and is predicted through color (*hue_mean* and *blue_std*), texture (*correlation*) and shape (*convexity*) (Figure 8A). A group of True Positives (*TP*) with *hue_mean* values between -3 and -2 (rescaled as described in Section 2.3.1) is revealed by the Shapley dependence plot (Figure 8B), which relates feature values (*hue_mean*) and their associated Shapley values for each particle (Lundberg et al., 2018). Such *TP* have almost 100% of confidence scores and consist of white (Figure 8C), red (predicted by *red_mode*, Figure 8D), rounded, hydrothermally altered material.

(2) The juvenile particles are accurately classified with a *F1-score* of 0.88 with color (*hue_mean*, *saturation_mode)*, texture (*correlation*), and shape (*convexity*) (Figure 9A). The *saturation_mode* feature, which relates to the intensity of color, is discriminant (Shapley values > 1) with values of 0–2 (Figure 9B). The *value_mode,* which measures the amount of reflected light, or gloss, and which is considered characteristic of juvenile particles under the binocular (Miwa et al., 2009) is also very important. Low values of *convexity* are also relevant for discrimination, as could be expected by the presence of vesicles on the particles' surfaces (Figure 9C). Moreover, the XGBoost predicts instances with lower *hue_mean* and *saturation_mode* as lithic (i.e., False Negative, FN), which correspond to darker, mid to high crystallinity juvenile particles from dome explosions (Figure 9D).

(3) The lithic particles are moderately well classified with a *F1-score* of 0.74, and is mainly discriminated through color (*value_mode* and *read_mean*) and texture (*homogeneity* and *correlation*) features (Figure 10A). Low values of *value_mode*, ranging between of -1.7 to 0 (Figure 10B), discriminate lithic particles. These features together with relatively high values of *correlation* reflect dark lithic particles with uniform texture (Figure 10C). In contrast, instances with higher pixel intensity-based features (*hue_mean* and *green_mean*) are a source of *FN*, as suggested by negative Shapley values, and are classified as altered material (Figure 10D).

(4) Free-crystals are the least accurately classified with *F1-score* of 0.54, and is mainly discriminated by color (*blue_std, hue_mean),* shape (*convexity)* and textural (*correlation*; Figure 11A*)*. Unlike the other types, the most discriminant feature doesn't cluster particles as shown by the *blue_std* values as a function of the Shapley values doesn't yield any cluster of *TP* (Figure 11B), and those with Shapley values > 1.5 overlap with altered material (Figure 11C). Thus, the XGBoost has limited predictability of free crystals, which is consistent with low a *F1-score* yielded from Free-crystals vs Rest binary classification (Table S8). Possible causes for this, besides the lack of a discriminant feature, include the presence of glass films on the crystal's surface, the wide range of aspects of different minerals (mostly plagioclase and

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).

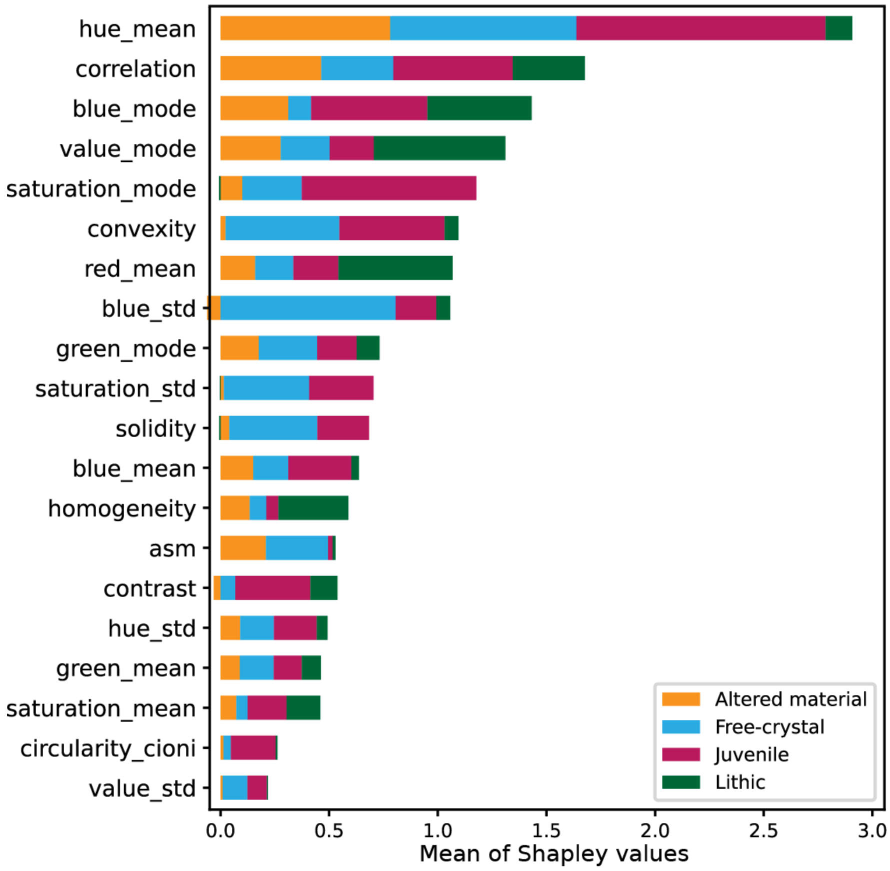

Figure 7. Aggregated mean of the Shapley values by particle type. Note that some features are important for discrimination of multiple particle types (e.g., *hue_mean*) and other features are more discriminant of a specific type (e.g., *value_mode* for lithic type). Meaning of the abbreviations can be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

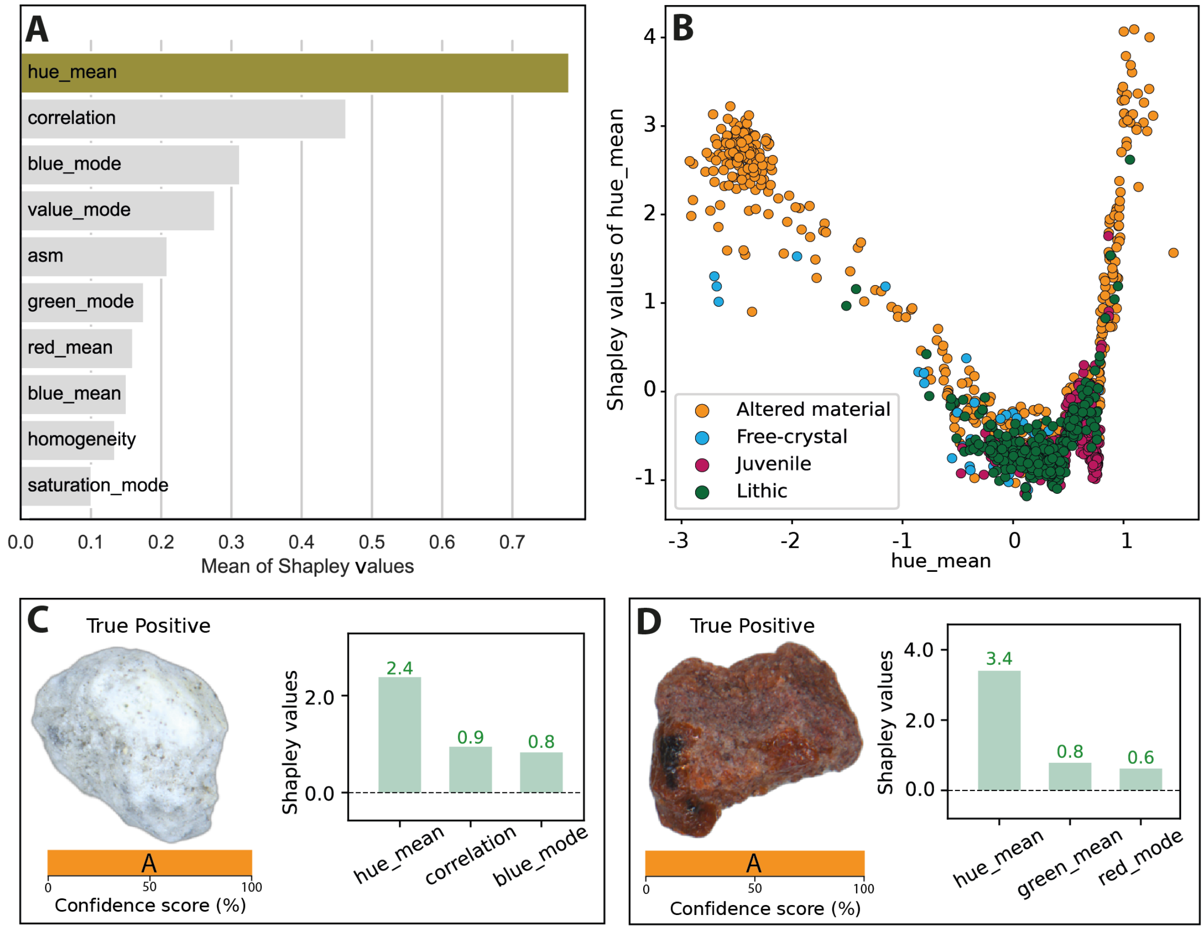

Figure 8. Summary plots to explain predictions of the altered material particle main type. (A) Feature importance according to the mean of the Shapley values, the higher the value the more the importance of the feature in the correct prediction. In (B) the Shapley dependence plot shows the relation of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

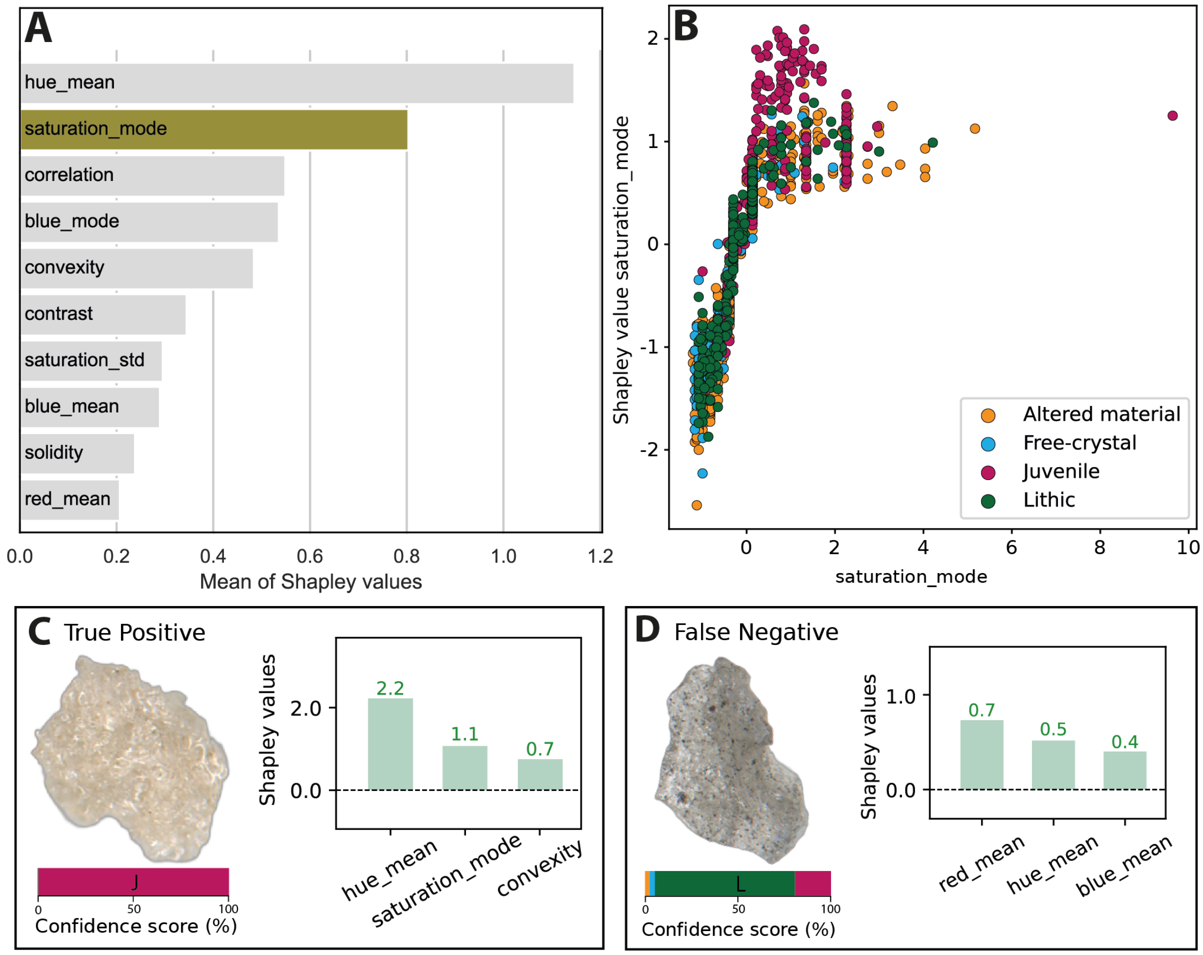

Figure 9. Summary plots to illustrate the features that contribute the most to the correct predictions of the juvenile particles. (A) Feature importance based on the mean of the Shapley values. (B) Shapley dependence plot. Note a cluster of juvenile particles around *saturation_mode* values between 1–3.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

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

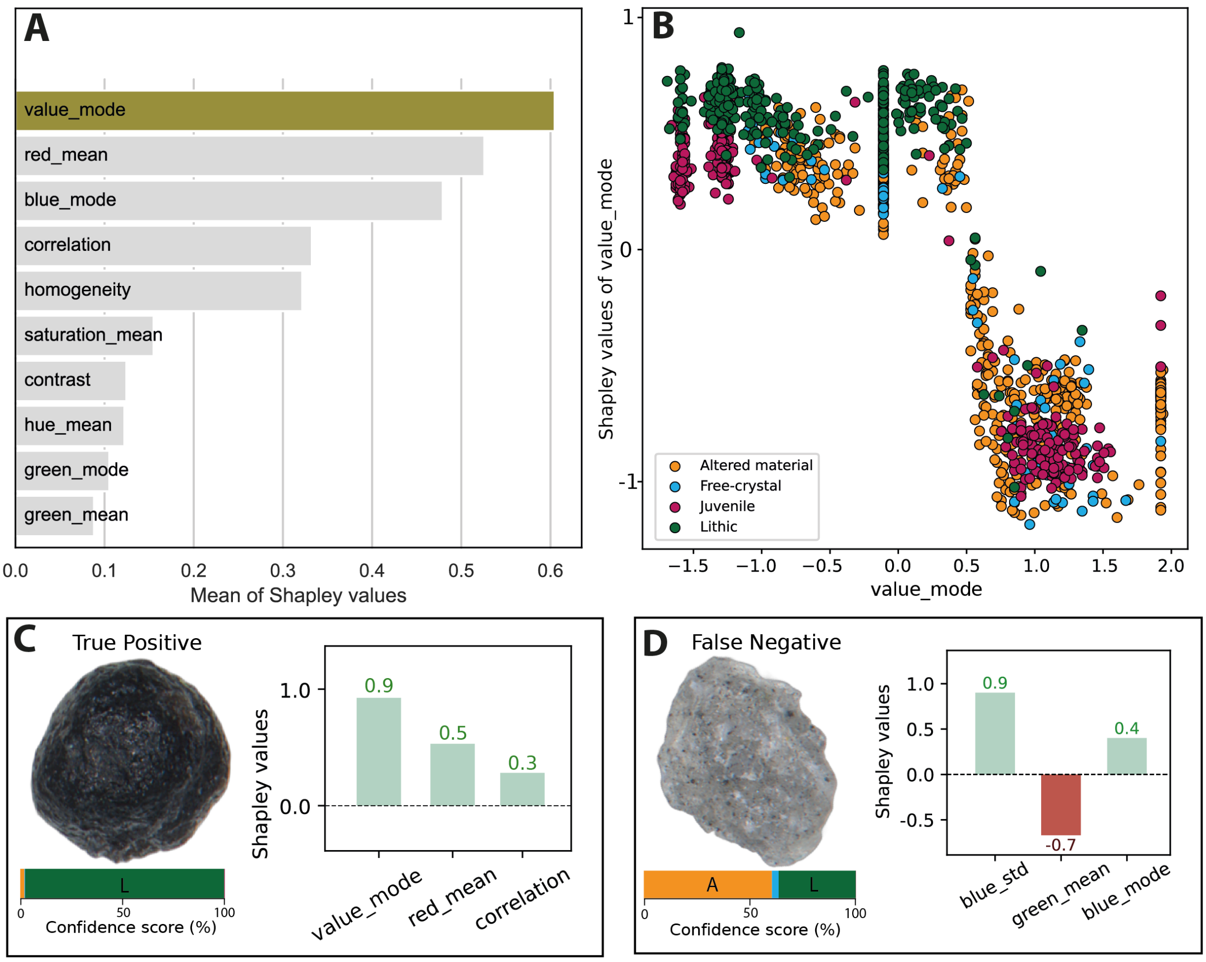

Figure 10. Summary plots to explain predictions of the lithic type. (A) Ranking of the features according to the mean of the Shapley values. (B) The Shapley dependence plot shows correct predictions of lithic particles with high Shapley values at negative values of *value_mode*. (C) and (D) show for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

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

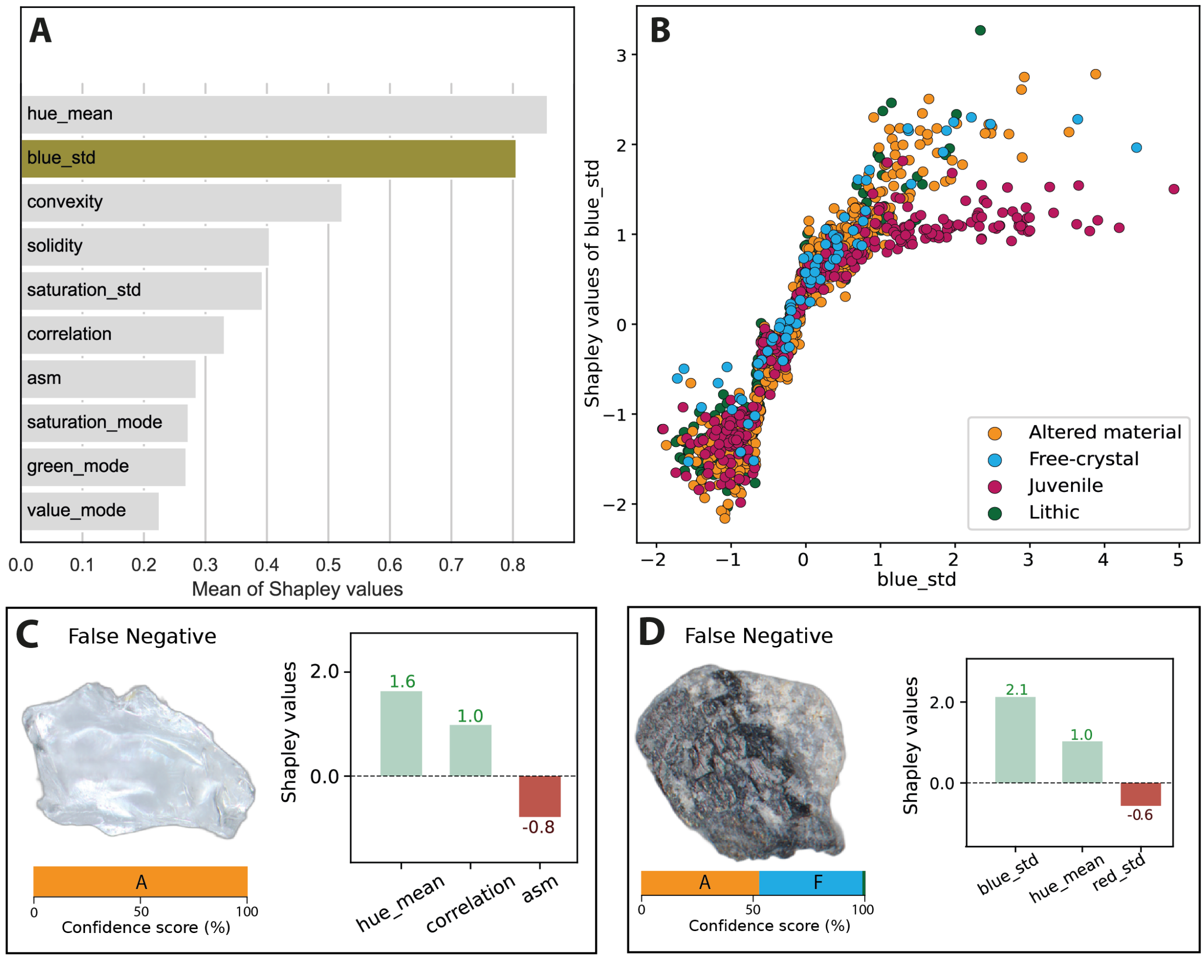

Figure 11. Summary plots to explain predictions of the models for the free-crystal type. (A) Feature importance based on the mean of the Shapley values. (B) Shapley dependence plot. Note that the feature values have been rescaled by a standard scaler. (C) and (D) show for each prediction the particle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

3.3 ViT quantitative evaluation

3.3.1 General evaluation

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

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

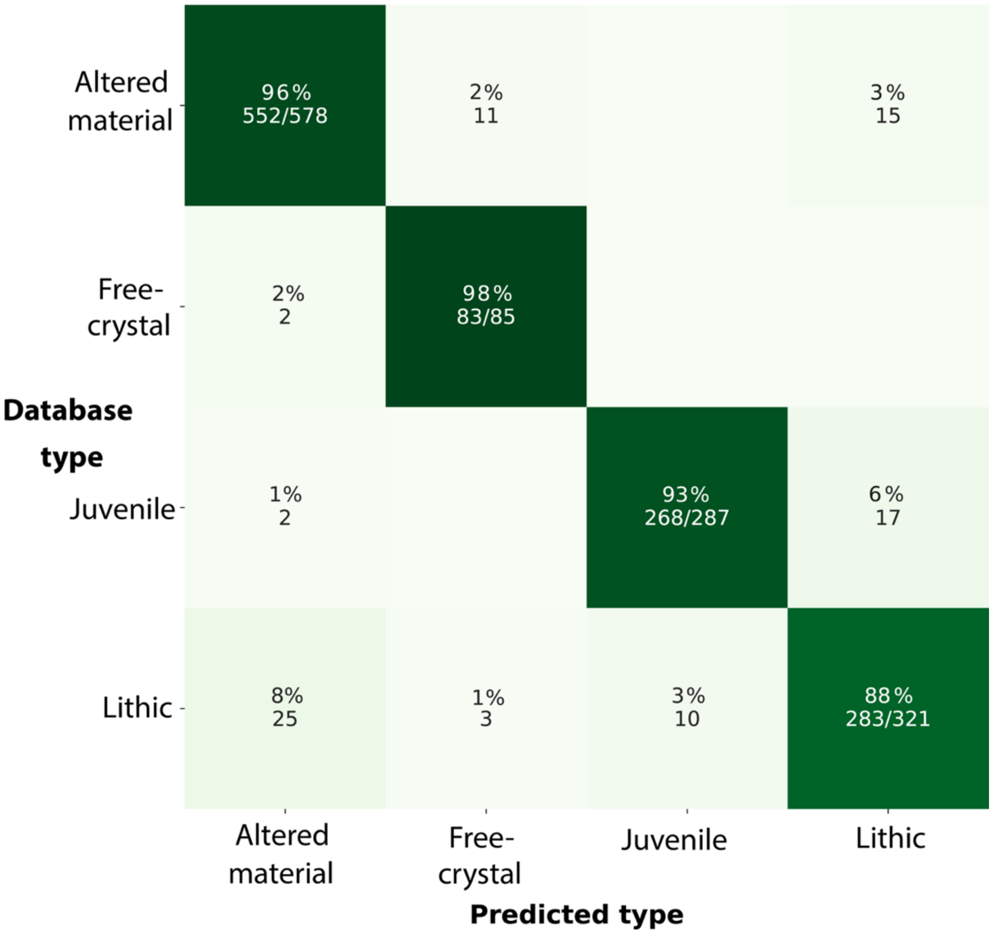

Figure 12. Confusion matrix of the predictions by the ViT image classifier. The percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and otherwise, the False Negative rate (lighter), all percentages with the corresponding number of particles

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

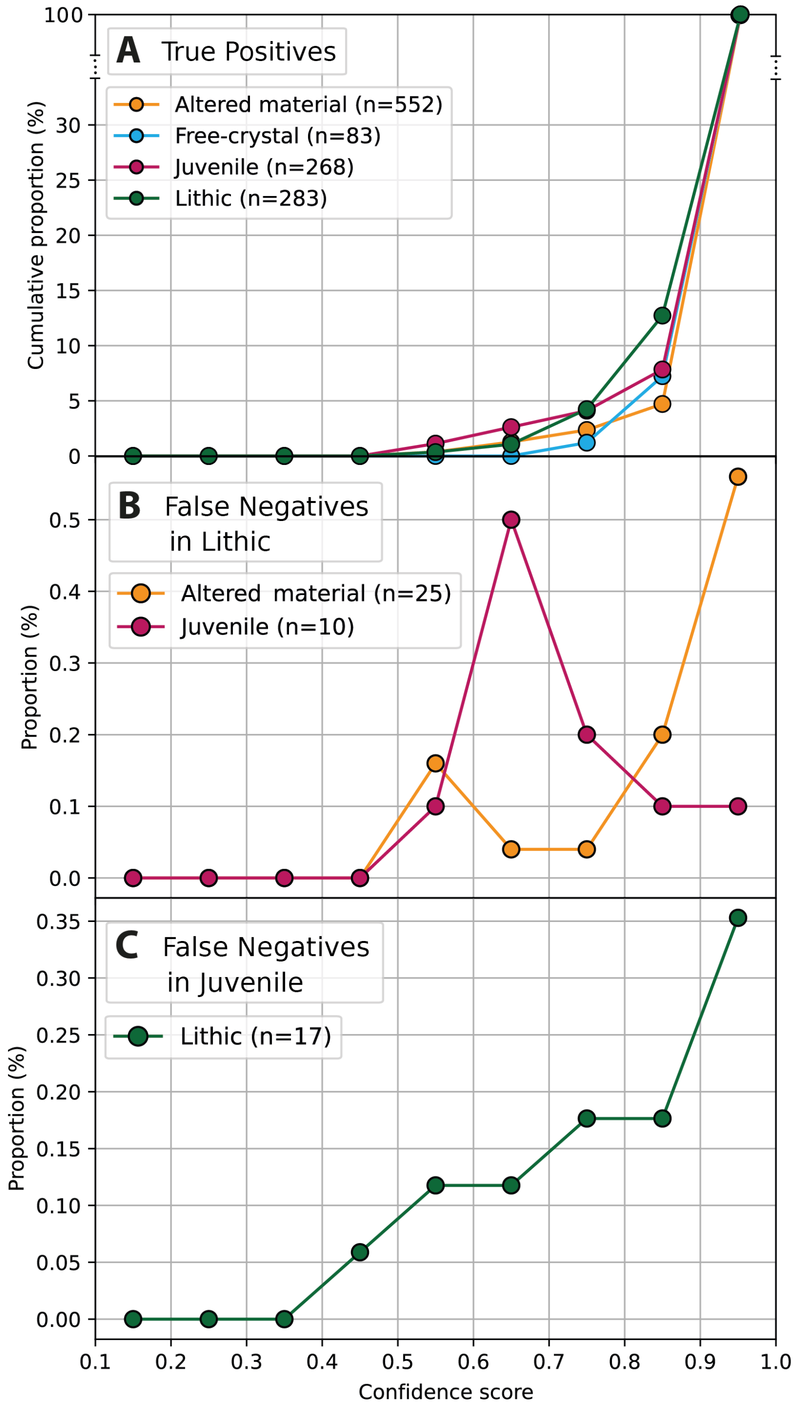

Figure 13. Line plots of the confidence score versus (A) the cumulative proportion of True Positives (TP), (B) False Negatives (FN) in free-crystal and (C) lithic types. The distribution of the data have been plotted into 9 bins of size 0.1. We don't use cumulative proportion in (B) and (C) given the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

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

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

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

(1) Ash particles from phreatic events are in general well classified (*macro F1-score* of 0.95), including the particle main types: altered material (*F1-score* of 0.99), free-crystal (*F1-score* of 0.94) and lithic (*F1-score* of 0.93). The ViT successfully classifies the most common groups of particles in these samples such as hydrothermal aggregates (Figure 15A) and weathered material (Figure 15B).

(2) Particles from samples of dome explosions are classified with the lowest accuracy (*macro F1-score* of 0.85) among the eruptive styles. The ViT accurately classifies free-crystal (*F1-score* of 0.86), altered material (*F1-score* of 0.90) and lithic (*F1-score* of 0.90) types, but is less accurate (*F1-score* of 0.73) for the juvenile type with most False Negatives (*FN*) classified as lithics. However, the confidence scores of some *FN* show a transition between the juvenile and lithic types that has explanatory value. This means that particles may have both juvenile and lithic traits, and thus a measure on the types' prevalence seems more realistic than using mutually exclusive types like in VolcAshDB. Particles with combined traits are common in samples from Nevados de Chillán Volcanic Complex (Figure 15C), which originated from a relatively long-lived dome-forming eruption cycle. An additional challenge is that the ViT confidently classifies as lithics some particles that are labelled as juvenile and, since petrographic classification was not always straightforward (Benet et al., *preprint*), it is difficult to decide whether these are False Negatives, or instead, petrographic classification errors (Figure 15D), especially when ML-based image classifiers have surpassed human performances in other fields (He et al., 2015).

(3) Ash particles from lava fountaining are generally accurately classified (*macro F1-score* of 0.94), between juvenile (*F1-score* of 0.94) and lithic (*F1-score* of 0.88) types. Most of the lithic particles belong to recycled juvenile particles, which are critical to avoid overestimating the amount of juvenile component (D'Oriano et al., 2022) and their identification typically requires examination in the SEM (D'Oriano et al., 2014). The high score suggests that the ViT can discriminate between them to some extent (Figure 15E), but a more robust labelling by a team of experts and a larger dataset containing SEM images is necessary to obtain more robust conclusions. On the other hand, the juvenile particles consist of glossy, smoothed surface, vesicular, elongated glass shards and are accurately classified (Figure 15F).

(4) The ViT accurately classifies ash particles from plinian and subplinian eruptive styles (*macro F1-score* of 0.95), including free crystals (*F1-score* of 0.92), altered material (*F1-score* of 0.93) and juvenile (1.0), but less accurate for lithics (*F1-score* of 0.77). The juvenile particles consist of fragments of pumice and all particles are successfully classified (Figure 15G). In contrast, the lithic particles mostly consist of dull grey fragments with rounded edges, and most of the *FN* are classified as altered material, which may reflect the challenge of classifying particles with incipient weathering into weathered material or lithic (Figure 15H).

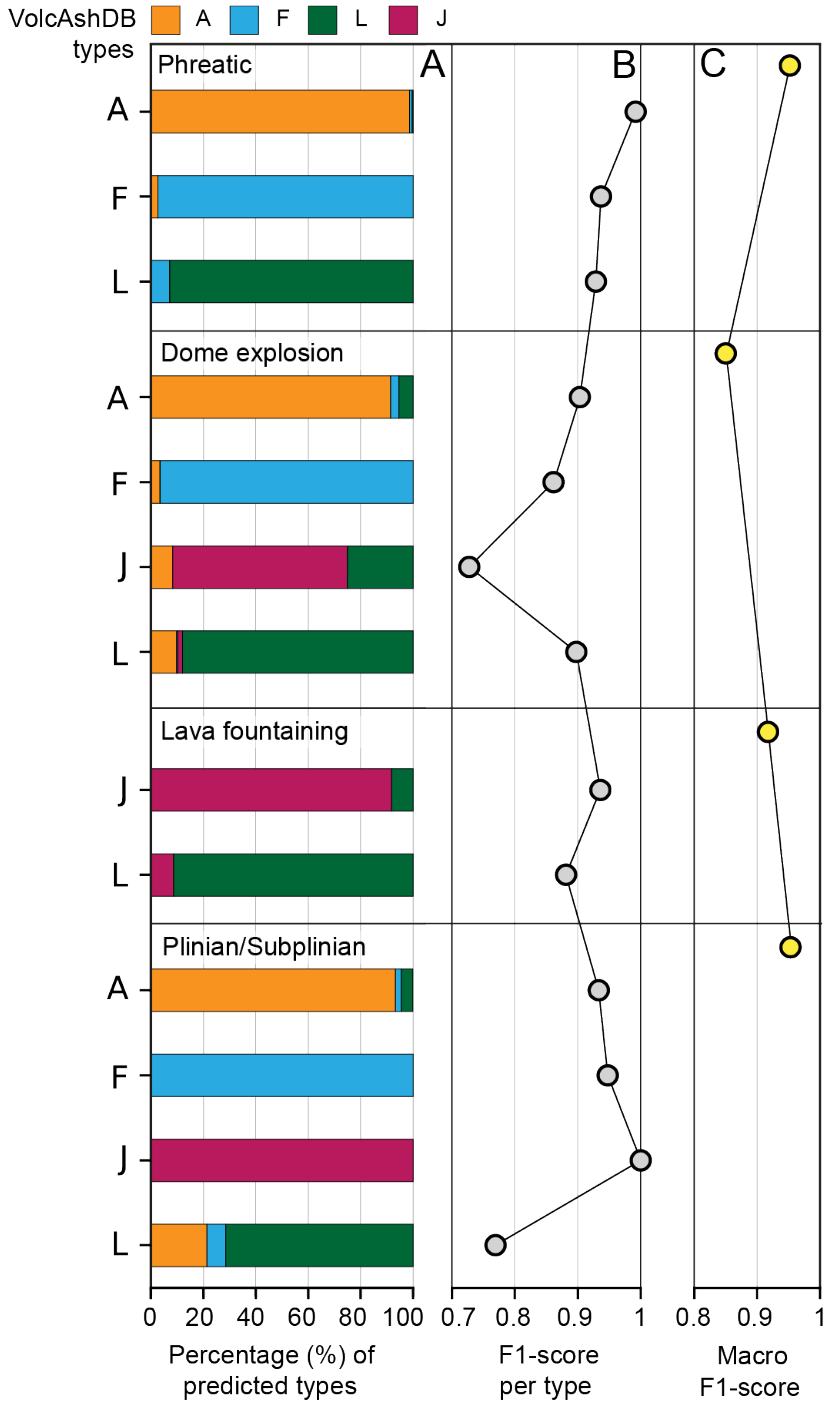

Figure 14. (A) Bar charts showing the percentage of predicted types for each particle type in VolcAshDB. If all predictions were the same as in the database, each bar would be single-colored as follows: orange for altered material (A), light blue for free-crystal (F), magenta for juvenile (J), and dark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

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

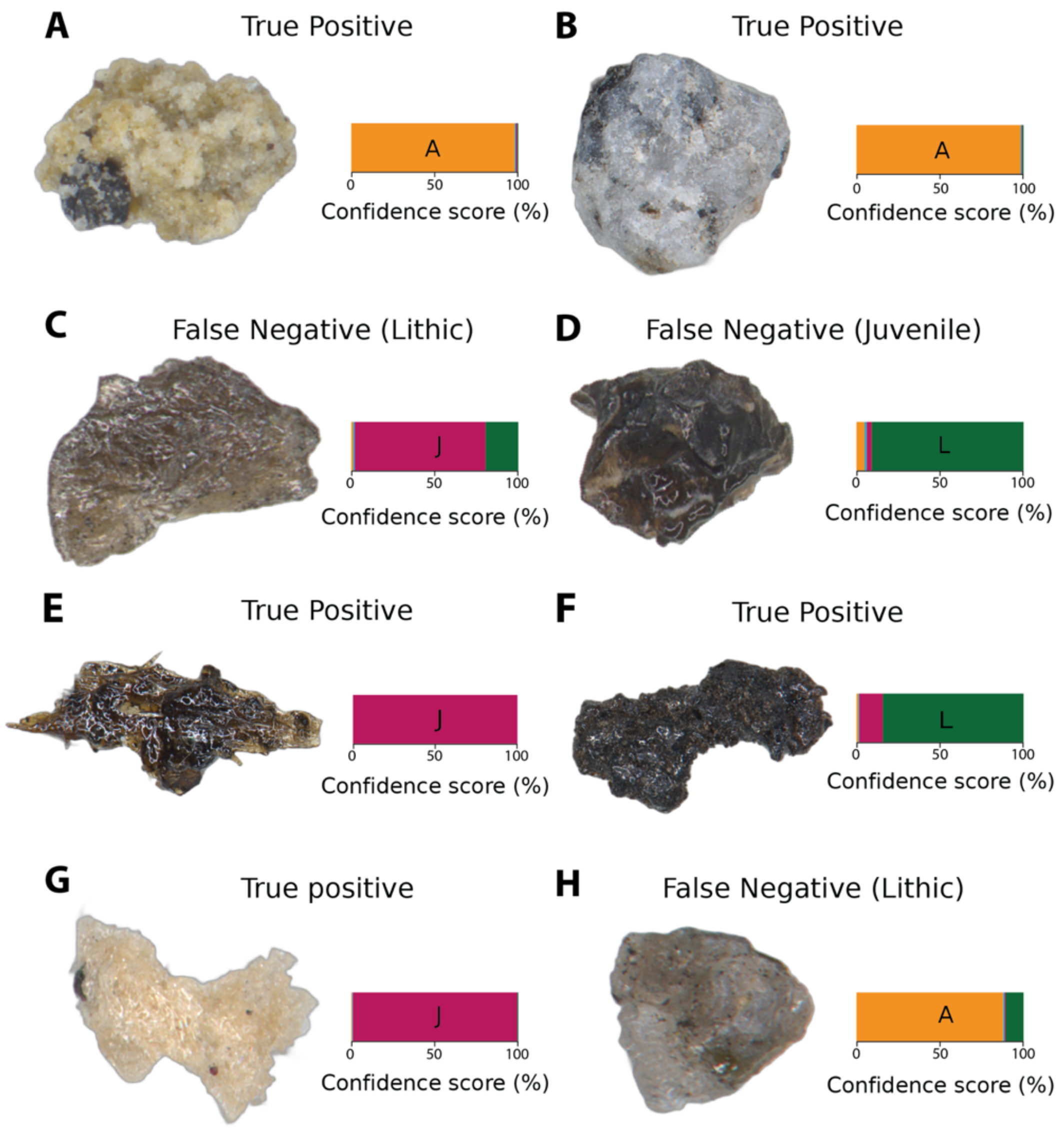

Figure 15. Representative examples of particle images and the predictions and their associated confidence score across eruptive styles, including phreatic (A,B), dome explosion (C,D), lava fountaining (E,F), and subplinian/plinian (G,H). Note that False Negatives contain in brackets the particle type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

4 Discussion

4.1 Comparison between classification using particle's features versus images

We found that, overall, the ViT classifies more accurately with particle images (0.93 of *macro F1-score*) than the XGBoost classifies with the particle features (0.77 of *macro F1-score*). This difference is unlikely to be the XGBoost model itself, which is very popular in the literature and has had best performances amongst models for complex classification tasks (Brownlee, 2016; Chen & Guestrin, 2016; Dhaliwal et al., 2018). One possibility is that the extracted features don't retain certain discriminant information from the images, and as a result, the XGBoost is unable to classify particles such as free crystals (0.57 of *F1-score*). On the other hand, maintaining the physical information associated with features makes the model's outcomes more interpretable (e.g., in classification of volcano-seismic signals; Falcin et al., 2021; Malfante et al., 2018) with xAI methods. This is an important advantage over Vision Transformers, whose main xAI tool consists in a heatmap of the region(s) of attention by the model (Dosovitskiy et al., 2020) but appears insufficient to obtain well founded classification insights for ash particles (Figure 16).

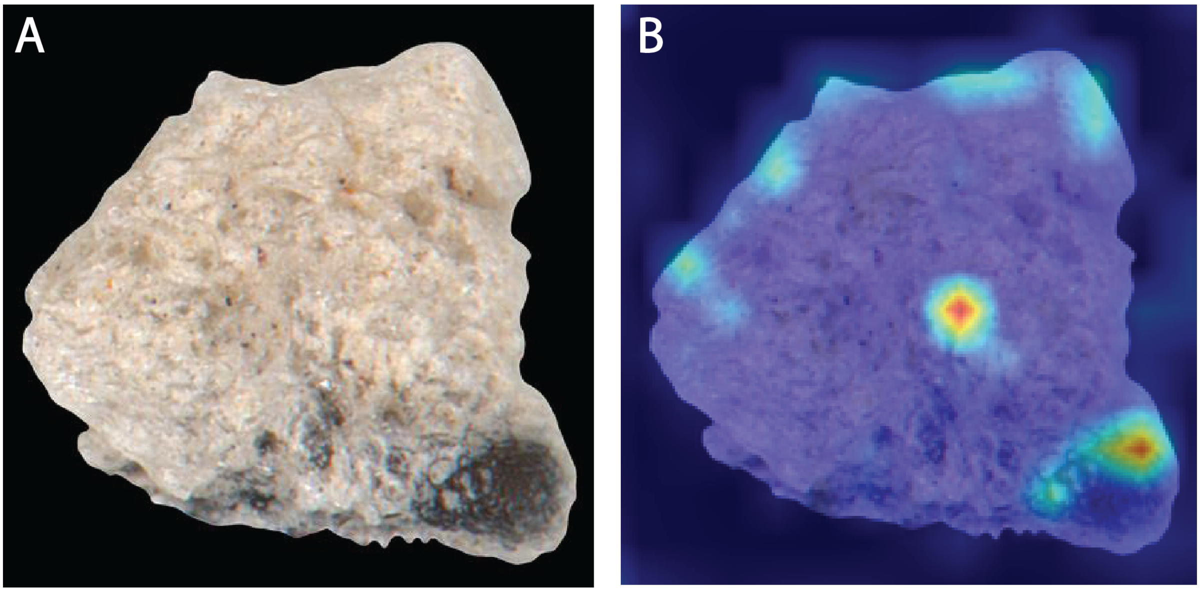

Figure 16. Example of (A) one multi-focused binocular image of a pumice particle from Mount St. Helens (1980), which is overlain by (B) a heatmap of the regions of attention by the base Vision Transformer (Dosovitskiy et al., 2020), typically used for interpreting image classifier's predictions. It does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

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

The XGBoost model gave a medium to high classification performance with *macro F1-score* of 0.77, and using the Shapley values we identified the most discriminant features of each particle type (Table 4). For instance, lithic particles can be distinguished with low values of *value_mode* which correspond to the luster of the particle according to the high Shapley values. This finding agrees with previous studies that use a dull luster (which corresponds to low values of *value_mode*) to identify lithic particles (Miwa et al., 2013). On the other hand, juvenile particles have high Shapley values for the *saturation_mode*. This feature is related to high color intensities as observed under the binocular, but it was not recognized before as a diagnostic observation of the particle type. These two examples belong to particle types that are well classified and for which the Shapley values are reliable. Shapley values obtained from particles that yielded lower accuracies, such as the free crystals, are not reliable, and thus overall performances should be improved. This could be achieved by enhancing the quality and quantity of VolcAshDB dataset by (i) adding particles to balance the dataset, (ii) refining the particle contour in the multi-focused images, so that shape features can measure micro-scaled cavities (Benet et al., *preprint*), and (iii) the inclusion of a new feature that measures the density of lines on the surface, which could be sensitive to planar structures of free crystals.

4.3 Deploying the ViT for automatic particle classification

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

4.4 A ViT particle classifier for volcano monitoring

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

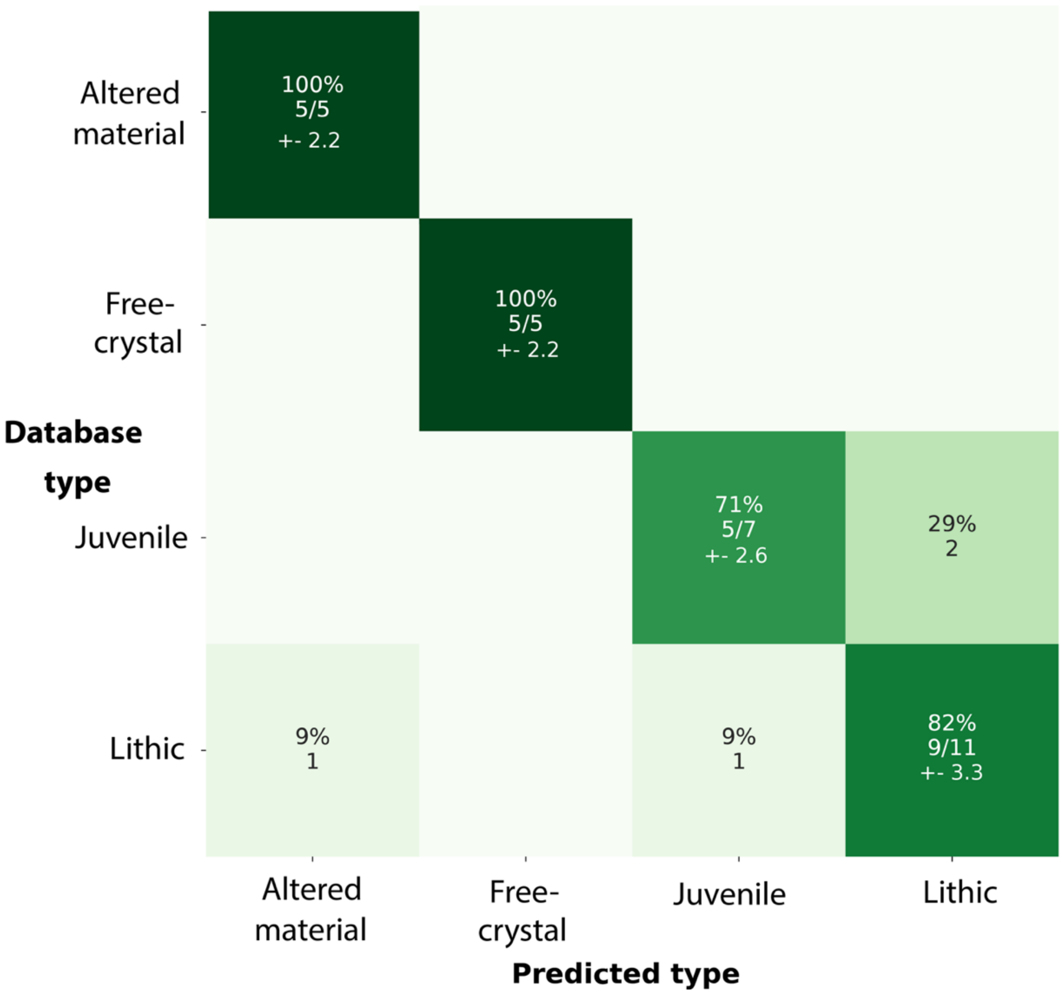

Figure 17. Confusion matrix of the predictions by the ViT image classifier after being fine-tuned with a single-focused, small training set (~370 particles from Benet et al., 2021). The percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and otherwise, the False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

5 Conclusions

Classification of the different particles that make up volcanic ash is not straightforward because diagnostic criteria are not standardized and thus reliable, and systematic identification of a given particle type is not straightforward. In this contribution, we attempt to alleviate this situation by exploring the use of state-of-the-art machine learning-based models to identify the most discriminant features of each particle type, and to evaluate their ability to classify particles. The identified features provide new insights on the recognition of juvenile and lithic particles towards a standardized classification. The image classifier performs at very high accuracies, although the variability across eruption and types shows that its capability to generalize to new samples is still unclear. Higher numbers of particles from a wider variety of eruptions and volcanoes into VolcAshDB coupled to ML models should allow for unbiased comparison of ash samples, and reproducible classification of their particles as a tool for volcano monitoring studies.

Acknowledgments

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

Open Research

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

References

Ahmed, S. N. (2015). Essential statistics for data analysis. In *Physics and Engineering of Radiation Detection*. https://doi.org/10.1016/b978-0-12-801363-2.00009-7

Alvarado, G. E., Mele, D., Dellino, P., de Moor, J. M., & Avard, G. (2016). Are the ashes from the latest eruptions (2010–2016) at Turrialba volcano (Costa Rica) related to phreatic or phreatomagmatic events? *Journal of Volcanology and Geothermal Research*, *327*, 407–415. https://doi.org/10.1016/j.jvolgeores.2016.09.003

Ayyadevara, V. K., & Reddy, Y. (2020). *Modern Computer Vision with PyTorch: Explore deep learning concepts and implement over 50 real-world image applications.* Packt Publishing Ltd.

Bebbington, M. S., & Jenkins, S. F. (2019). *Intra-eruption forecasting*.

Benet, D., Costa, F., Pedreros, G., & Cardona, C. (2021). The volcanic ash record of shallow magma intrusion and dome emplacement at Nevados de Chillán Volcanic complex, Chile. *Journal of Volcanology and Geothermal Research*, *417*. https://doi.org/10.1016/j.jvolgeores.2021.107308

Benet, D., Costa, F., Widiwijayanti, C., Pallister, J., Pedreros, G., Allard, P., Humaida, H., Aoki, Y., & Maeno, F. (2023). VolcashDB: Volcanic ash particle image and classification database. *Preprint in EarthArxiv*. https://doi.org/10.31223/X53659

Biass, S., Jenkins, S. F., Aeberhard, W. H., Delmelle, P., & Wilson, T. (2022). Insights into the vulnerability of vegetation to tephra fallouts from interpretable machine learning and big Earth observation data. *Natural Hazards and Earth System Sciences*, *22*(9), 2829–2855. https://doi.org/10.5194/nhess-22-2829-2022

Brownlee, J. (2016). XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn. *Machine Learning Mastery*.

Brownlee, J. (2020). Imbalanced Classification with Python. *Machine Learning Mastery*, 463.

Cashman, K. V., & Hoblitt, R. P. (2004). Magmatic precursors to the 18 May 1980 eruption of Mount St. Helens, USA. *Geology*, *32*(2), 141–144. https://doi.org/10.1130/G20078.1

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Cioni, R., Pistolesi, M., Bertagnini, A., Bonadonna, C., Hoskuldsson, A., & Scateni, B. (2014). Insights into the dynamics and evolution of the 2010 Eyjafjallajökull summit eruption (Iceland) provided by volcanic ash textures. *Earth and Planetary Science Letters*, *394*(May 2010), 111–123. https://doi.org/10.1016/j.epsl.2014.02.051

D'Oriano, C., Bertagnini, A., Cioni, R., & Pompilio, M. (2014). Identifying recycled ash in basaltic eruptions. *Scientific Reports*, *4*. https://doi.org/10.1038/srep05851

D'Oriano, C., Del Carlo, P., Andronico, D., Cioni, R., Gabellini, P., Cristaldi, A., & Pompilio, M. (2022). Syn-Eruptive Processes During the January–February 2019 Ash-Rich Emissions Cycle at Mt. Etna (Italy): Implications for Petrological Monitoring of Volcanic Ash. *Frontiers in Earth Science*, *10*(February 2019). https://doi.org/10.3389/feart.2022.824872

Dellino, P., & La Volpe, L. (1996). Image processing analysis in reconstructing fragmentation and transportation mechanisms of pyroclastic deposits. The case of Monte Pilato-Rocche Rosse eruptions, Lipari (Aeolian islands, Italy). *Journal of Volcanology and Geothermal Research*, *71*(1), 13–29. https://doi.org/10.1016/0377-0273(95)00062-3

Dhaliwal, S. S., Nahid, A. Al, & Abbas, R. (2018). Effective intrusion detection system using XGBoost. *Information (Switzerland)*, *9*(7). https://doi.org/10.3390/info9070149

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Dürig, T., Bowman, M. H., White, J. D. L., Murch, A., Mele, D., Verolino, A., & Dellino, P. (2018). Particle shape analyzer Partisan - An open source tool for multi-standard two-dimensional particle morphometry analysis. *Annals of Geophysics*, *61*(6). https://doi.org/10.4401/ag-7865

Dürig, T., Ross, P. S., Dellino, P., White, J. D. L., Mele, D., & Comida, P. P. (2021). A review of statistical tools for morphometric analysis of juvenile pyroclasts. *Bulletin of Volcanology*, *83*(11). https://doi.org/10.1007/s00445-021-01500-0

Falcin, A., Métaxian, J. P., Mars, J., Stutzmann, É., Komorowski, J. C., Moretti, R., Malfante, M., Beauducel, F., Saurel, J. M., Dessert, C., Burtin, A., Ucciani, G., de Chabalier, J. B., & Lemarchand, A. (2021). A machine-learning approach for automatic classification of volcanic seismicity at La Soufrière Volcano, Guadeloupe. *Journal of Volcanology and Geothermal Research*, *411*. https://doi.org/10.1016/j.jvolgeores.2020.107151

Feuillard, M., Allegre, C. J., Brandeis, G., Gaulon, R., Le Mouel, J. ., Mercier, J. C., Pozzi, J. P., & Semet, M. . (1983). The 1975–1977 crisis of la Soufriere de Guadeloupe (F.W.I): A still-born magmatic eruption. *Journal of Volcanology and Geothermal Research*, *16*, 317–334.

Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning*, *73*(2), 133–153. https://doi.org/10.1007/s10994-008-5064-8

Gaunt, H. E., Bernard, B., Hidalgo, S., Proaño, A., Wright, H., Mothes, P., Criollo, E., & Kueppers, U. (2016). Juvenile magma recognition and eruptive dynamics inferred from the analysis of ash time series: The 2015 reawakening of Cotopaxi volcano. *Journal of Volcanology and Geothermal Research*, *328*, 134–146. https://doi.org/10.1016/j.jvolgeores.2016.10.013

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. In *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*.

Gianfagna, L., & Di Cecco, A. (2021). Explainable AI. In *Berlin/Heidelberg, Germany: Springer*. https://doi.org/10.3233/FAIA190100

Hall-Beyer, M. (2017). GLCM Texture: A Tutorial. *17th International Symposium on Ballistics*, *2*(March), 18–19.

Haralick, R. M., Dinstein, I., & Shanmugam, K. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Herrera, F., Charte, F., Rivera, A. J., & del Jesus, M. J. (2016). *Multi-Label Classification* (Springer,). https://doi.org/10.4018/jdwm.2007070101

Hincks, T. K., Komorowski, J. C., Sparks, S. R., & Aspinall, W. P. (2014). Retrospective analysis of uncertain eruption precursors at La Soufrière volcano, Guadeloupe, 1975-77: Volcanic hazard assessment using a Bayesian Belief Network approach. *Journal of Applied Volcanology*, *3*(1). https://doi.org/10.1186/2191-5040-3-3

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/cvprw.2009.5206848

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-

Torres, M. Á., & Carvalhais, N. (2022). Wildfire Danger Prediction and Understanding With Deep Learning. *Geophysical Research Letters*, *49*(17), 1–11. https://doi.org/10.1029/2022GL099368

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Le Guern, F., Bernard, A., & Chevrier, R. M. (1980). Soufrière of guadeloupe 1976–1977 eruption — mass and energy transfer and volcanic health hazards. *Bulletin Volcanologique*, *43*(3), 577–593. https://doi.org/10.1007/BF02597694

Lee, J. J., Aime, M. C., Rajwa, B., & Bae, E. (2022). Machine Learning-Based Classification of Mushrooms Using a Smartphone Application. *Applied Sciences (Switzerland)*, *12*(22). https://doi.org/10.3390/app122211685

Leibrandt, S., & Le Pennec, J. L. (2015). Towards fast and routine analyses of volcanic ash morphometry for eruption surveillance applications. *Journal of Volcanology and Geothermal Research*, *297*, 11–27. https://doi.org/10.1016/j.jvolgeores.2015.03.014

Liu, E. J., Cashman, K. V., Miller, E., Moore, H., Edmonds, M., Kunz, B. E., Jenner, F., & Chigna, G. (2020). Petrologic monitoring at Volcán de Fuego, Guatemala. *Journal of Volcanology and Geothermal Research*, *405*(August 2019), 107044. https://doi.org/10.1016/j.jvolgeores.2020.107044

Liu, E. J., Cashman, K. V., & Rust, A. C. (2015). Optimising shape analysis to quantify volcanic ash morphology. *GeoResJ*, *8*, 14–30. https://doi.org/10.1016/j.grj.2015.09.001

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the 2020s*. 11966–11976. https://doi.org/10.1109/cvpr52688.2022.01167

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). *Consistent Individualized Feature Attribution for Tree Ensembles*. *2*. http://arxiv.org/abs/1802.03888

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Maeno, F., Nakada, S., Yoshimoto, M., Shimano, T., Hokanishi, N., Zaennudin, A., & Iguchi, M. (2019). A sequence of a plinian eruption preceded by dome destruction at Kelud volcano, Indonesia, on February 13, 2014, revealed from tephra fallout and pyroclastic density current deposits. *Journal of Volcanology and Geothermal Research*, *382*, 24–41. https://doi.org/10.1016/j.jvolgeores.2017.03.002

Malfante, M., Mura, M. D., Métaxian, J., & Mars, J. I. (2018). *Machine Learning for Volcano-Seismic Signals*. *March*, 20–30.

Mandal, S., Mones, S. M. B., Das, A., Balas, V. E., Shaw, R. N., & Ghosh, A. (2021). Single shot detection for detecting real-time flying objects for unmanned aerial vehicle. In *Artificial Intelligence for Future Generation Robotics*. INC. https://doi.org/10.1016/B978-0-323-85498-6.00005-8

Marzocchi, W., Newhall, C., & Woo, G. (2012). The scientific management of volcanic crises.

Journal of Volcanology and Geothermal Research, *247–248*, 181–189. https://doi.org/10.1016/j.jvolgeores.2012.08.016

Mishra, P. (2022). Practical Explainable AI Using Python. In *Practical Explainable AI Using Python*. https://doi.org/10.1007/978-1-4842-7158-2

Miwa, T., Geshi, N., & Shinohara, H. (2013). Temporal variation in volcanic ash texture during a vulcanian eruption at the sakurajima volcano, Japan. *Journal of Volcanology and Geothermal Research*, *260*, 80–89. https://doi.org/10.1016/j.jvolgeores.2013.05.010

Miwa, T., Toramaru, A., & Iguchi, M. (2009). Correlations of volcanic ash texture with explosion earthquakes from vulcanian eruptions at Sakurajima volcano, Japan. *Journal of Volcanology and Geothermal Research*, *184*(3–4), 473–486. https://doi.org/10.1016/j.jvolgeores.2009.05.012

Miyagi, I., Geshi, N., Hamasaki, S., Oikawa, T., & Tomiya, A. (2020). Heat source of the 2014 phreatic eruption of Mount Ontake, Japan. *Bulletin of Volcanology*, *82*(4). https://doi.org/10.1007/s00445-020-1358-x

Molnar, C. (2021). Interpretable Machine Learning. *Queue*, *19*(6), 28–56. https://doi.org/10.1145/3511299

Moran, S. C., Newhall, C., & Roman, D. C. (2011). Failed magmatic eruptions: Late-stage cessation of magma ascent. *Bulletin of Volcanology*, *73*(2), 115–122. https://doi.org/10.1007/s00445-010-0444-x

Newhall, C. G., & Punongbayan, R. S. (1996). The narrow margin of successful volcanic-risk mitigation. In *Monitoring and mitigation of volcano hazards* (pp. 807–838). Springer Science & Business Media.

Nurfiani, D., & Bouvet de Maisonneuve, C. (2018). Furthering the investigation of eruption styles through quantitative shape analyses of volcanic ash particles. *Journal of Volcanology and Geothermal Research*, *354*, 102–114. https://doi.org/10.1016/j.jvolgeores.2017.12.001

Owen, L. (2022). *Hyperparameter Tuning with Python*.

Paladio-Melasantos, M. L., Solidum, R. U., Scott, W. E., Quiambao, R. B., Umbal, J. V, Rodolfo, K. S., Tubianosa, B. S., Delos Reyes, P. J., Alonso, R. A., & Ruerlo, H. B. (1996). Tephra falls of the 1991 eruptions of Mount Pinatubo. *In: Newhall, C.G. (Editor) & Others, Fire and Mud; Eruptions and Lahars of Mount Pinatubo, Philippines, Philippine Institute of Volcanology and Seismology, Quezon City*, *layer D*, 413–535. https://doi.org/10.1159/000153100

Panati, C., Wagner, S., & Bruggenwirth, S. (2022). Feature Relevance Evaluation using Grad-CAM, LIME and SHAP for Deep Learning SAR Data Classification. *Proceedings International Radar Symposium*, *2022-Septe*, 457–462.

Pardo, N., Avellaneda, J. D., Rausch, J., Jaramillo-Vogel, D., Gutiérrez, M., & Foubert, A. (2020). Decrypting silicic magma/plug fragmentation at Azufral crater lake, Northern Andes: insights from fine to extremely fine ash morpho-chemistry. *Bulletin of Volcanology*, *82*(12). https://doi.org/10.1007/s00445-020-01418-z

Pardo, N., Cronin, S. J., Németh, K., Brenna, M., Schipper, C. I., Breard, E., White, J. D. L., Procter, J., Stewart, B., Agustín-Flores, J., Moebis, A., Zernack, A., Kereszturi, G., Lube, G., Auer, A., Neall, V., & Wallace, C. (2014). Perils in distinguishing phreatic from phreatomagmatic ash; insights into the eruption mechanisms of the 6 August 2012 Mt. Tongariro eruption, New Zealand. *Journal of Volcanology and Geothermal Research*, *286*, 397–414. https://doi.org/10.1016/j.jvolgeores.2014.05.001

Re, G., Corsaro, R. A., D'Oriano, C., & Pompilio, M. (2021). Petrological monitoring of active volcanoes: A review of existing procedures to achieve best practices and operative protocols during eruptions. *Journal of Volcanology and Geothermal Research*, *419*, 107365. https://doi.org/10.1016/j.jvolgeores.2021.107365

Romero, J. E., Burton, M., Cáceres, F., Taddeucci, J., Civico, R., Ricci, T., Pankhurst, M. J., Hernández, P. A., Bonadonna, C., Llewellin, E. W., & Pistolesi, M. (2022). *The initial phase of the 2021 Cumbre Vieja ridge eruption (Canary Islands): Products and dynamics controlling edifice growth and collapse*. *431*(July). https://doi.org/10.1016/j.jvolgeores.2022.107642

Ross, P. S., Dürig, T., Comida, P. P., Lefebvre, N., White, J. D. L., Andronico, D., Thivet, S., Eychenne, J., & Gurioli, L. (2022). Standardized analysis of juvenile pyroclasts in comparative studies of primary magma fragmentation; 1. Overview and workflow. *Bulletin of Volcanology*, *84*(1), 1–29. https://doi.org/10.1007/s00445-021-01516-6

Rowe, M. C., Thornber, C. R., & Kent, A. J. R. (2008). Identification and Evolution of the Juvenile Component in. *A Volcano Rekindled: The Renewed Eruption of Mount St. Helens, 2004–2006*, 2004–2006.

Shapley, L. S. (1953). A Value for n-Person Games, in: Contributions to the Theory of Games II. *Contributions to the Theory of Games*, 307–318. https://doi.org/https://doi.org/10.1515/9781400881970-018

Shoji, D., Noguchi, R., Otsuki, S., & Hino, H. (2018). *Classification of volcanic ash particles using a convolutional neural network and probability*. 1–12. https://doi.org/10.1038/s41598-018-26200-2

Sujatha, R., Chatterjee, J. M., Jhanjhi, N. Z., & Brohi, S. N. (2021). Performance of deep learning vs machine learning in plant leaf disease detection. *Microprocessors and Microsystems*, *80*(November 2020). https://doi.org/10.1016/j.micpro.2020.103615

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *30th International Conference on Machine Learning, ICML 2013*, *PART 3*, 2176–2184.

Suzuki, Y., Nagai, M., Maeno, F., Yasuda, A., Hokanishi, N., Shimano, T., Ichihara, M., Kaneko, T., & Nakada, S. (2013). Precursory activity and evolution of the 2011 eruption of Shinmoe-dake in Kirishima volcano-insights from ash samples. *Earth, Planets and Space*, *65*(6), 591–607. https://doi.org/10.5047/eps.2013.02.004

Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J. A., Zou, J., & Rubin, D. L. (2021). Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. *Scientific Reports*, *11*(1), 1–9. https://doi.org/10.1038/s41598-021-87762-2

Tilling, R. ~I. (2008). The critical role of volcano monitoring in risk reduction. *Advances in Geosciences*, *14*, 3–11.

Verdhan, V. (2020). Supervised Learning with Python. In *Supervised Learning with Python*. https://doi.org/10.1007/978-1-4842-6156-9

Watanabe, K., Danhara, T., Watanabe, K., Terai, K., & Yamashita, T. (1999). Juvenile volcanic glass erupted before the appearance of the 1991 lava dome, Unzen volcano, Kyushu, Japan. *Journal of Volcanology and Geothermal Research*, *89*(1–4), 113–121. https://doi.org/10.1016/S0377-0273(98)00127-9

Volcanic ash classification through Machine Learning

Damià Benet[1,2,3,†], Fidel Costa[1], Christina Widiwijayanti[2]

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

[2] EOS, Earth Observatory of Singapore, Nanyang Technological University, Singapore.

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

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

Key Points:

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

Abstract

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

1 Introduction

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

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

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

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

2 Materials and Methods

2.1 VolcAshDB dataset

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

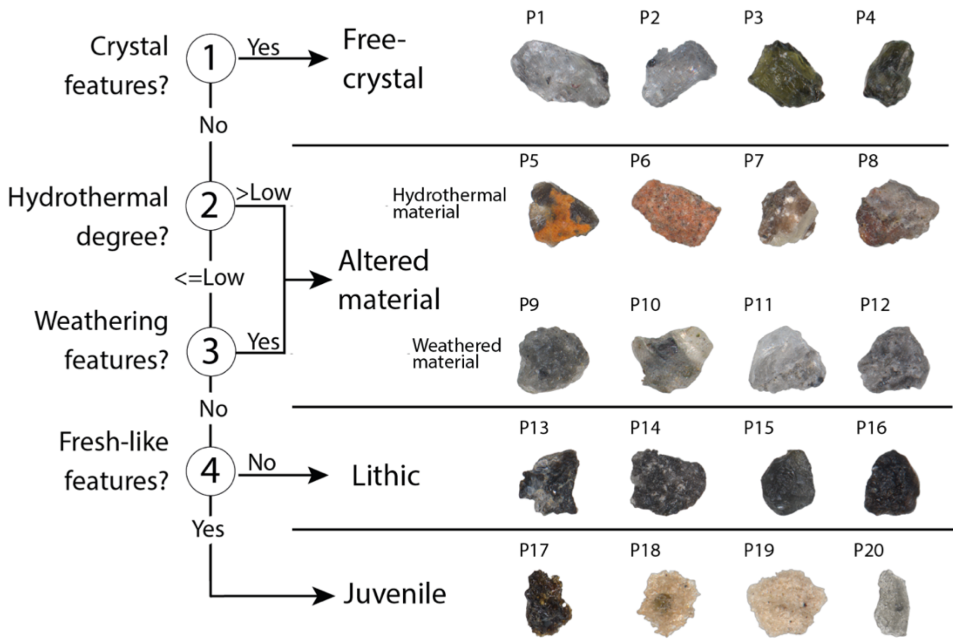

Figure 1. Example of classification process and particle images in VolcAshDB based on the steps for petrographic classification in Benet et al., (*preprint*). Note that the particle type altered material comprises both hydrothermal and weathered material.

Table 1. Main sample characteristics, and proportion of main particle types in VolcAshDB. The associated error is calculated using the equation of margin of error Benet et al., (*preprint*) 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	
Cumbre Vieja									
CV-DB1	19/10/21	Mafic	Cinder cone	Lava fountaining	3 (±0.3)	1(±0.2)	719(±2.8)	352(±1.4)	1075
Kelud									
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
Merapi									
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
Sourfière de Guadeloupe									
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
Nevados de Chillán									
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
Ontake									

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

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

2.2 Development of a particle classifier using the measured particle features

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

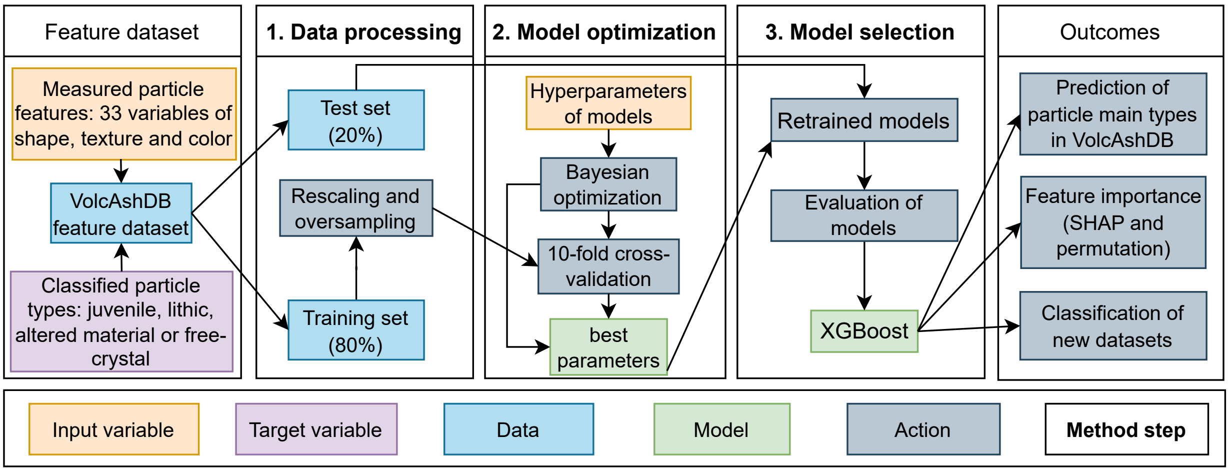

Figure 2. Illustration of the steps involved from the dataset to the outcomes, including those to obtain the best optimized model, XGBoost. (1) Data processing of the full dataset (features and particle types), including the oversampling of the training set. (2) hyperparameter optimization and cross-validation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

2.2.1 Data processing

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

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

2.2.2 Hyperparameter optimization

Hyperparameters control the model learning process and are explicitly defined by the 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

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

2.2.3 Model evaluation and selection

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

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\text{-}score = \frac{2*TP}{2*TP+FP+FN} \tag{4}$$

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

2.2.4 Explaining the model's predictions

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

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

2.2.5 Classification strategies

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

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

As noted above, we first split the dataset into a training (80% of all particle features in VolcAshDB) and a test set (20%) and used the latter to evaluate the XGBoost's performance. However, as splitting process is random it may affect the precision and accuracy of the measured *F1-scores*. To estimate this error, we re-trained and evaluated the XGBoost at 1,000 different values of random state, i.e., the hyperparameter that controls randomness. We obtained an average accuracy (*macro F1-score* of *0.76*; Table S4) that is like the accuracy from the test set (*macro F1-score* of *0.75*). The free-crystal type shows the widest variability (standard deviation of 0.04) and is the most inaccurate (*F1*-score of 0.57; Figure 3) amongst the particle types. This is likely because it is the least abundant type, and its classification is challenging given the different types of minerals and lack of a discriminant feature as explained below (Section 3.1). Accuracies of the three other types are higher (*F1*-score of 0.73–0.88) and with better precision (standard deviation is < 0.02; Table S4).

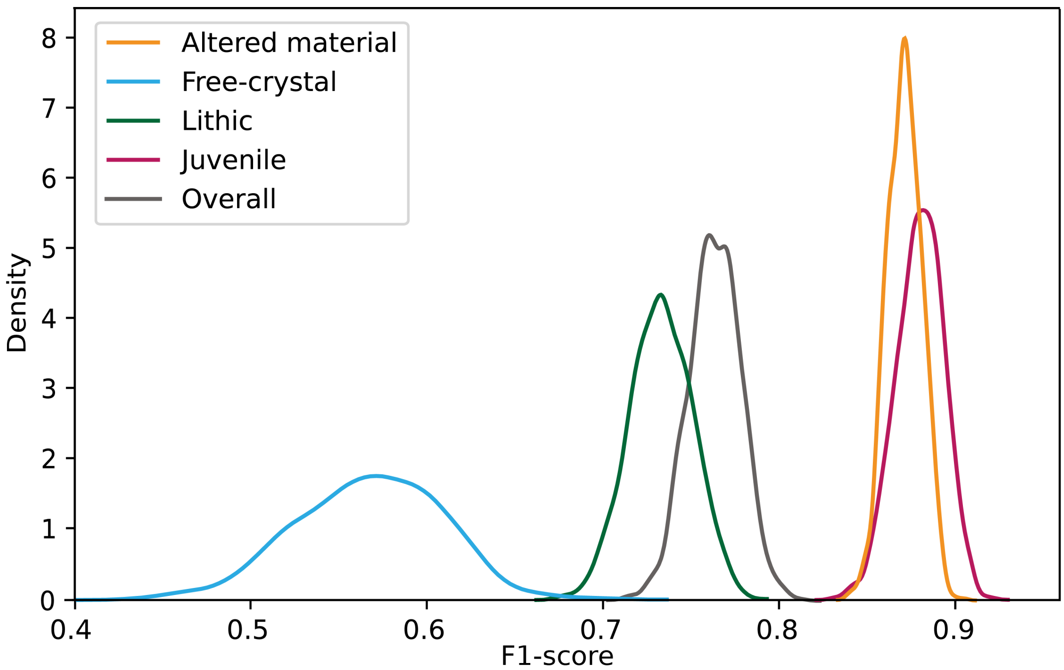

Figure 3. Density plots of the *F1-scores* obtained from 1,000 runs of the XGBoost at different random state across particle types and aggregated as *macro F1-score* (Overall).

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

$$\alpha = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \text{ x2} \qquad (5)$$

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

2.3 Development of a particle classifier using VolcAshDB image dataset

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

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

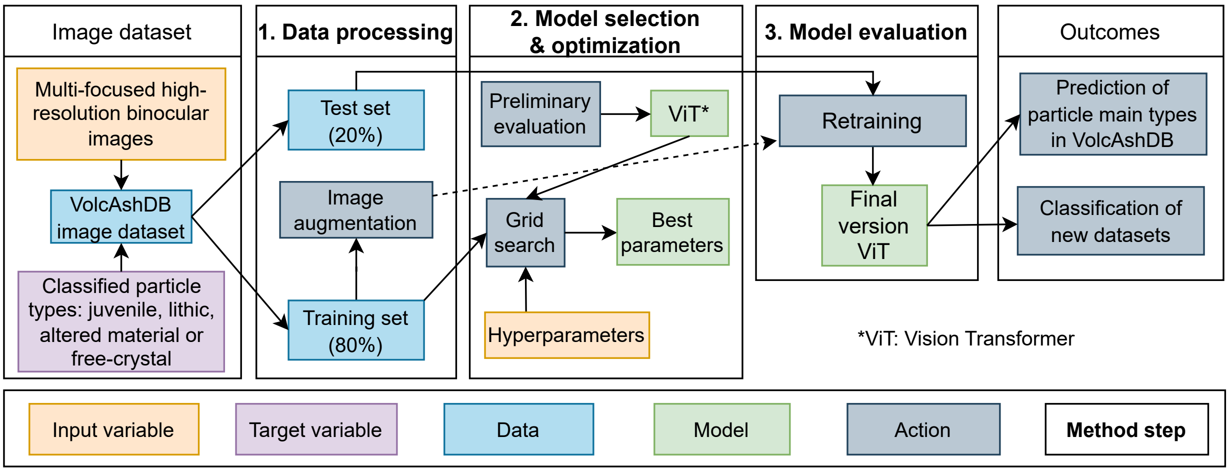

Figure 4. Illustration of the steps involved from the dataset to the outcomes, including those to fine-tune the Vision Transformer (ViT). (1) Data processing of the full dataset (images and particle types). (2) preliminary evaluation of the models using the base hyperparameters, selection of ViT and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

2.3.1 Image augmentation and processing

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

2.3.2 Fine-tuning, preliminary evaluation, and model selection

We fine-tuned the classifiers and did a preliminary round of evaluations to choose the best-performing model. Fine-tuning consists in making small adjustments to an already trained classifier, as opposed to training, where the data drives the model's learning process without any prior exposure. We selected the model before hyperparameter optimization because each run is time consuming (lasting about 14–18 hours) and because the authors of each model already provide the base hyperparameters (Table S5). The fine-tuned model that yielded the highest accuracy is ViT (0.88), followed by ConvNext and ResNet, both with an accuracy of 0.86.

2.3.3 ViT Hyperparameter optimization

We obtained the optimal hyperparameters following the grid search technique for two ranges of batch size and learning rate. In grid search, each hyperparameter is modified one step at a time, while the other hyperparameters remain fixed, throughout all the possible combinations (Owen, 2022). We found that the optimal batch size and learning rate are 16 and $3x10^{-5}$, respectively (accuracies obtained from grid search are reported in Table S6). Using these values, we tested three different optimizers, AdamW (Loshchilov & Hutter, 2019), Stochastic Gradient Descent (Sutskever et al., 2013) and Adagrad (Duchi et al., 2011) with the former performing the best (Table S7). We also tested and an increasing number of epochs (i.e., 5, 10, 15, 20), which didn't improve performance above 10, probably because the model had already converged.

2.3.4 Model evaluation

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

3 Results

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

Table 2. *F1-score* values for the whole database (unweighted average or *macro*) and particle types obtained from various models, including XGBoost multilabel, One-vs-One (OVO), One-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

3.1 XGBoost quantitative evaluation

Overall, the XGBoost shows rather accurate *F1-scores* across classification strategies: 0.76 for multilabel, 0.75 for OVO, and 0.76 for OVR (Table 2). Computation of the confusion matrix (Figure 5) shows that the model classifies best the altered material type (*F1-score* of 0.9), closely followed by the juvenile type (*F1-score* of 0.88), and less accurately the lithic type (*F1-score* of 0.74), and significantly less the free-crystal type (*F1-score* of 0.57).

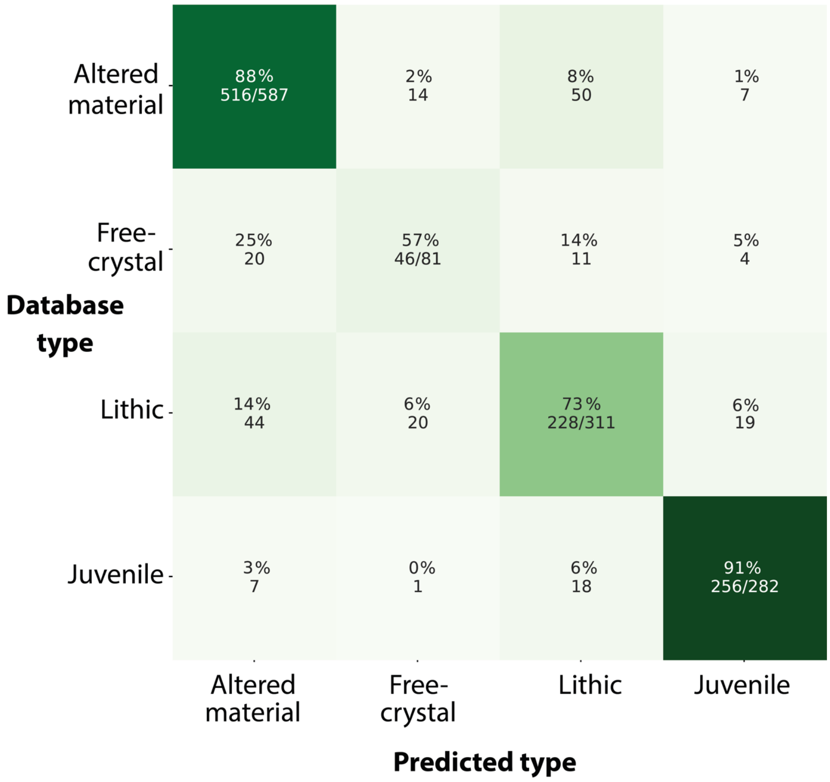

Figure 5. Confusion matrix of the predictions by the XGBoost multilabel classifier. The percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and otherwise, the False Negative rate (lighter), all percentages with the corresponding number of particles per predicted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

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

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

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

	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

[1]*F*: Free-crystal [2]*A*: Altered material [3]*L*: Lithic [4]*J*: Juvenile

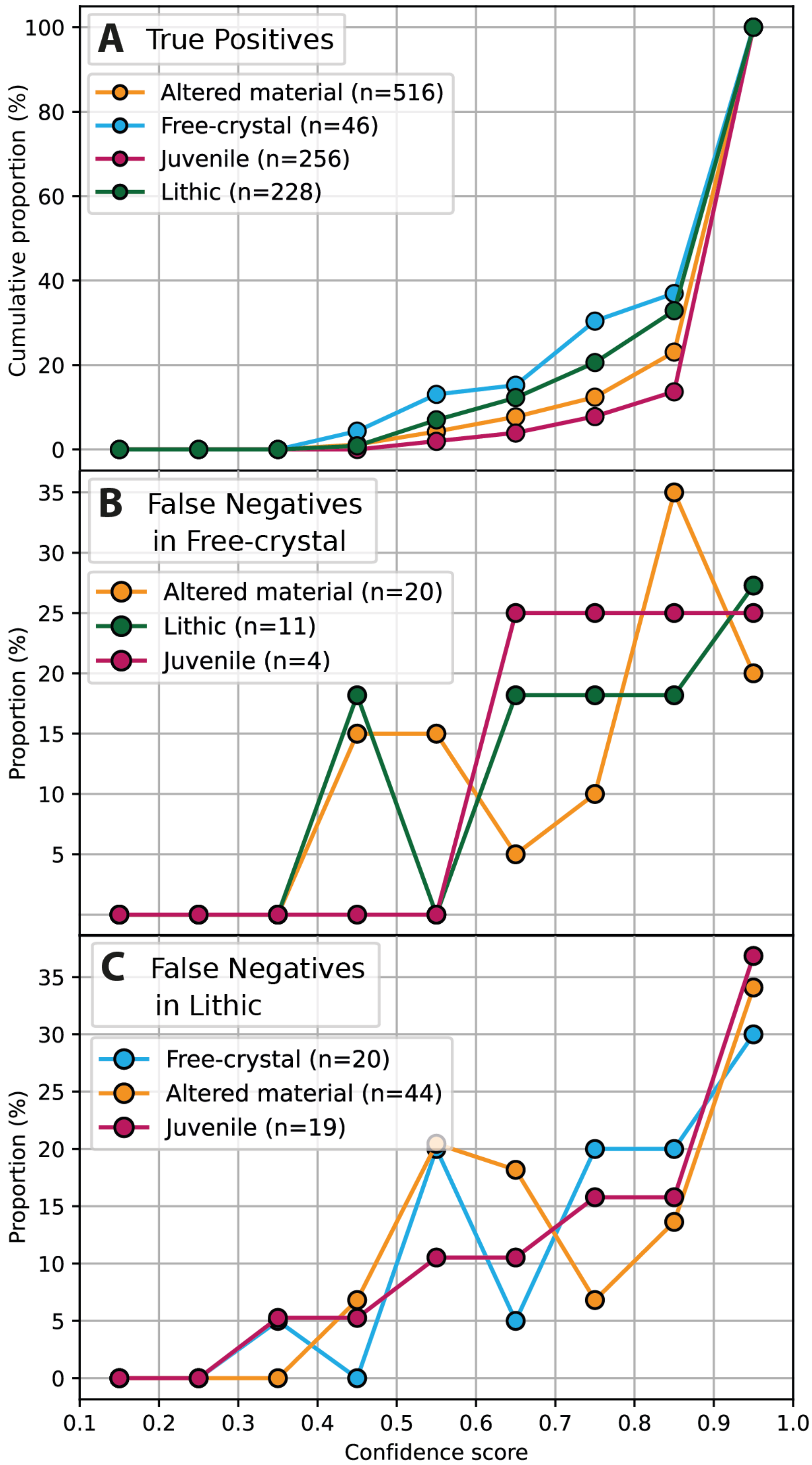

Figure 6. Line plots of the confidence score versus (A) the cumulative proportion of True Positives (TP), (B) False Negatives (FN) in free-crystal, and (C) lithic types. The distribution of the data have been plotted into 9 bins of size 0.1. We don't use cumulative proportion in (B) and (C) given the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

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

3.2 What features drive XGBoost ash particle type predictions?

3.2.1 Global feature importance

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

Table 4. Feature importance identification based on mean of Shapley values and feature permutation. These two methods calculate the feature importance values differently and can't be directly compared. The relative ranking of the features importance is similar (top ten ranked features in bold) with the same top two ranked features (*hue_mean* and *correlation*). We used the Shapley mean value for feature importance per particle type (shown as a plot in Figure 7), the top three of which are underlined. For the meaning of the abbreviations of each feature please see Table S1. The permutation feature values have been multiplied by ten for better readability, as the importance lies on the relative values across 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	**0.78**	**0.86**	**0.12**	**1.15**	**2.91**	**0.91**	**0.41**	**0.15**	**0.91**	**1.22**
correlation	**0.46**	**0.33**	**0.33**	**0.55**	**1.68**	**0.34**	0.02	**0.19**	**0.04**	**0.29**
blue_mode	**0.31**	0.10	**0.48**	**0.54**	**1.43**	0.06	**0.04**	0.00	**0.01**	**0.10**
value_mode	**0.28**	**0.23**	**0.60**	0.20	**1.31**	0.05	**0.05**	**0.24**	0.00	0.00
saturation_mode	**0.10**	**0.27**	-0.01	**0.80**	**1.17**	0.02	**0.06**	**0.10**	**0.10**	**0.13**
convexity	0.02	**0.52**	0.06	**0.48**	**1.10**	0.01	**0.06**	0.00	**0.03**	0.03
red_mean	**0.16**	0.18	**0.53**	**0.21**	**1.07**	0.03	0.03	0.01	**0.01**	0.04
blue_std	-0.06	**0.81**	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	

3.2.2 Local feature importance across particle types

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

(1) <u>Altered material</u> has the highest classification success with a *F1-score* of 0.90 and is predicted through color (*hue_mean* and *blue_std*), texture (*correlation*) and shape (*convexity*) (Figure 8A). A group of True Positives (*TP*) with *hue_mean* values between -3 and -2 (rescaled as described in Section 2.3.1) is revealed by the Shapley dependence plot (Figure 8B), which relates feature values (*hue_mean*) and their associated Shapley values for each particle (Lundberg et al., 2018). Such *TP* have almost 100% of confidence scores and consist of white (Figure 8C), red (predicted by *red_mode*, Figure 8D), rounded, hydrothermally altered material.

(2) The <u>juvenile</u> particles are accurately classified with a *F1-score* of 0.88 with color (*hue_mean*, *saturation_mode)*, texture (*correlation*), and shape (*convexity*) (Figure 9A). The *saturation_mode* feature, which relates to the intensity of color, is discriminant (Shapley values > 1) with values of 0–2 (Figure 9B). The *value_mode,* which measures the amount of reflected light, or gloss, and which is considered characteristic of juvenile particles under the binocular (Miwa et al., 2009) is also very important. Low values of *convexity* are also relevant for discrimination, as could be expected by the presence of vesicles on the particles' surfaces (Figure 9C). Moreover, the XGBoost predicts instances with lower *hue_mean* and *saturation_mode* as lithic (i.e., False Negative, FN), which correspond to darker, mid to high crystallinity juvenile particles from dome explosions (Figure 9D).

(3) The <u>lithic</u> particles are moderately well classified with a *F1-score* of 0.74, and is mainly discriminated through color (*value_mode* and *read_mean*) and texture (*homogeneity* and *correlation*) features (Figure 10A). Low values of *value_mode*, ranging between of -1.7 to 0 (Figure 10B), discriminate lithic particles. These features together with relatively high values of *correlation* reflect dark lithic particles with uniform texture (Figure 10C). In contrast, instances with higher pixel intensity-based features (*hue_mean* and *green_mean*) are a source of *FN*, as suggested by negative Shapley values, and are classified as altered material (Figure 10D).

(4) <u>Free-crystals</u> are the least accurately classified with *F1-score* of 0.54, and is mainly discriminated by color (*blue_std, hue_mean),* shape (*convexity)* and textural (*correlation*; Figure 11A*).* Unlike the other types, the most discriminant feature doesn't cluster particles as shown by the *blue_std* values as a function of the Shapley values doesn't yield any cluster of *TP* (Figure 11B), and those with Shapley values > 1.5 overlap with altered material (Figure 11C). Thus, the XGBoost has limited predictability of free crystals, which is consistent with low a *F1-score* yielded from Free-crystals vs Rest binary classification (Table S8). Possible causes for this, besides the lack of a discriminant feature, include the presence of glass films on the crystal's surface, the wide range of aspects of different minerals (mostly plagioclase and

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).

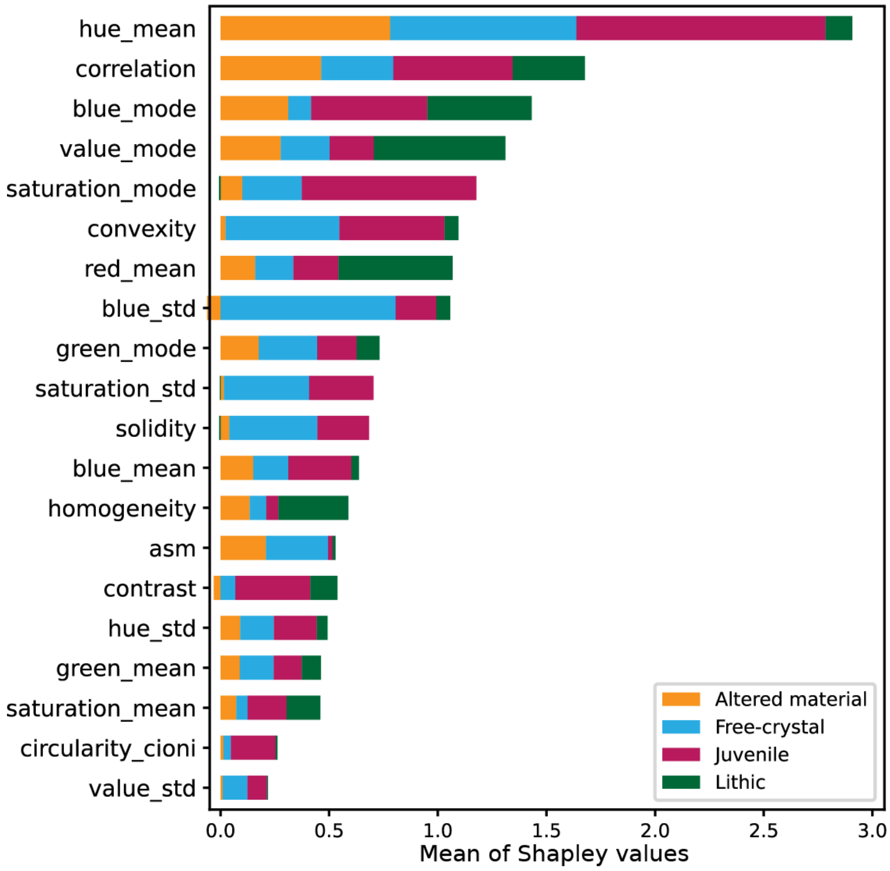

Figure 7. Aggregated mean of the Shapley values by particle type. Note that some features are important for discrimination of multiple particle types (e.g., *hue_mean*) and other features are more discriminant of a specific type (e.g., *value_mode* for lithic type). Meaning of the abbreviations can be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

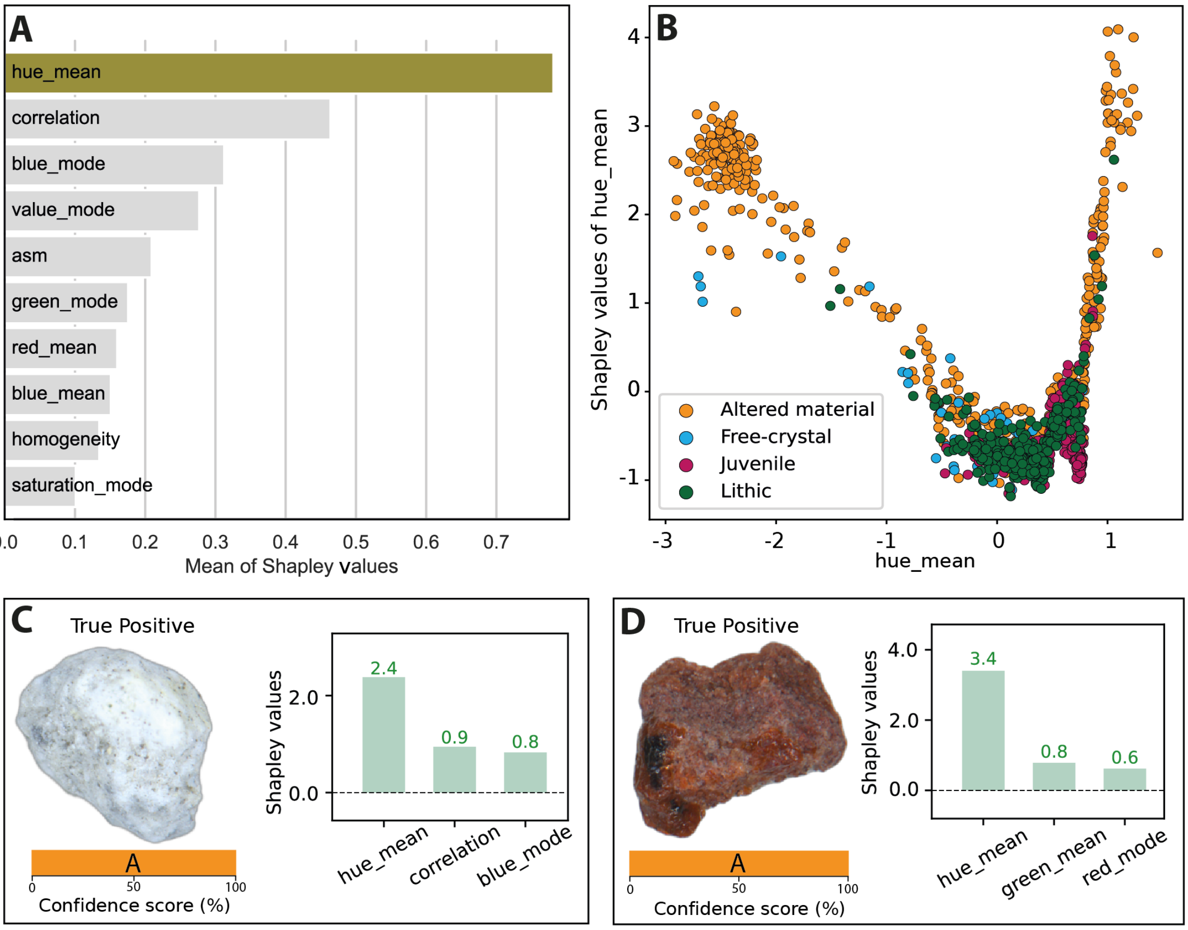

Figure 8. Summary plots to explain predictions of the altered material particle main type. (A) Feature importance according to the mean of the Shapley values, the higher the value the more the importance of the feature in the correct prediction. In (B) the Shapley dependence plot shows the relation of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

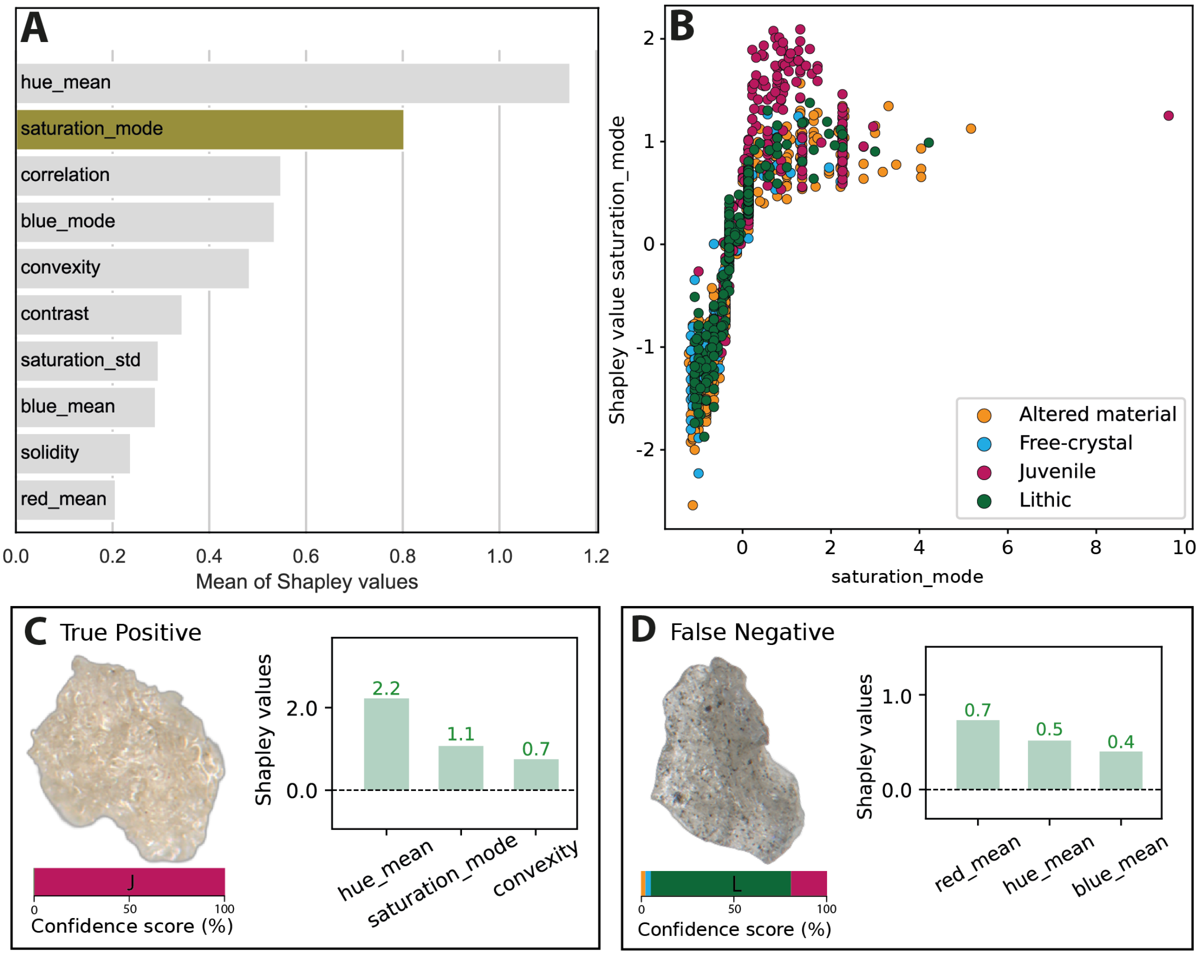

Figure 9. Summary plots to illustrate the features that contribute the most to the correct predictions of the juvenile particles. (A) Feature importance based on the mean of the Shapley values. (B) Shapley dependence plot. Note a cluster of juvenile particles around *saturation_mode* values between 1–3.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

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

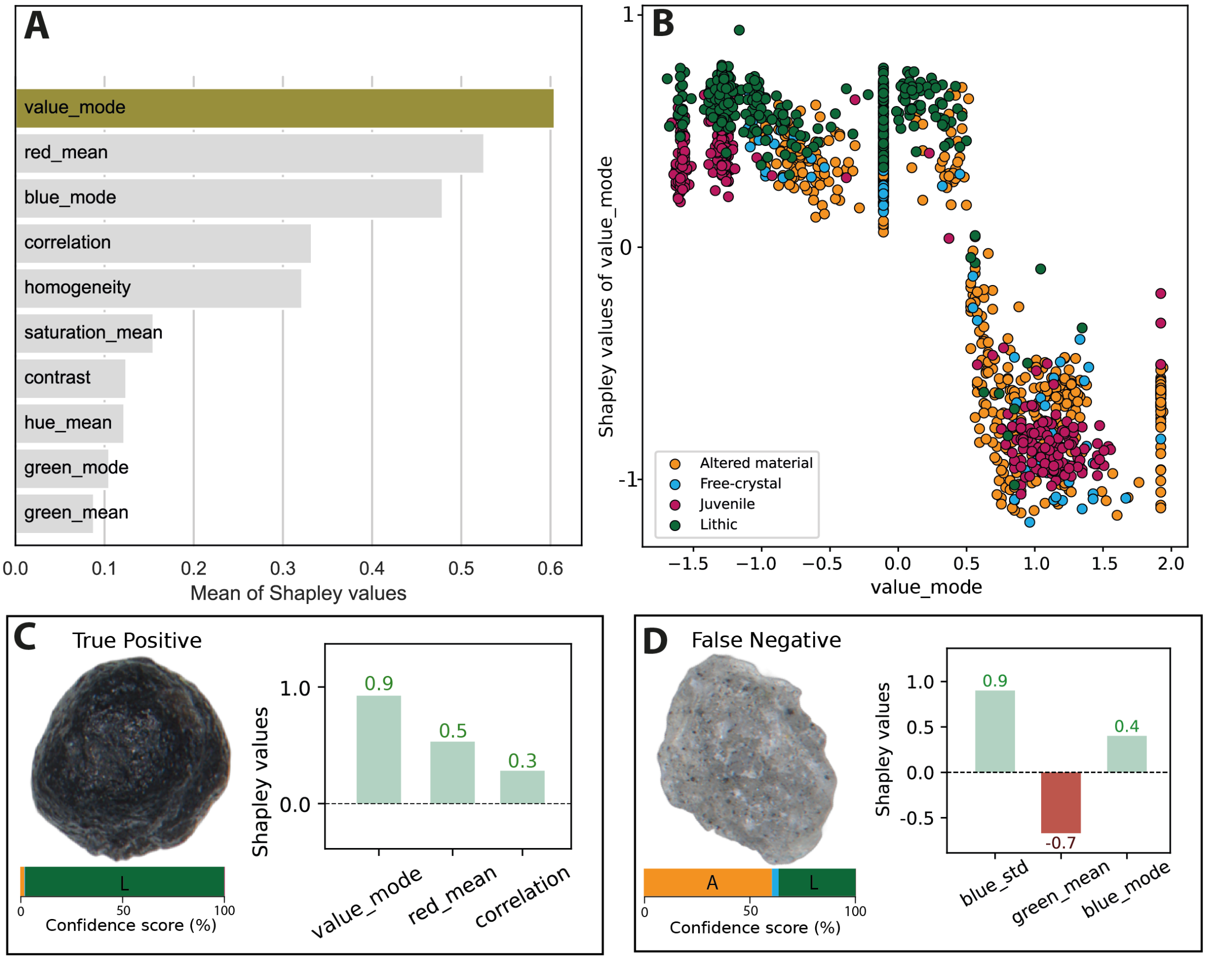

Figure 10. Summary plots to explain predictions of the lithic type. (A) Ranking of the features according to the mean of the Shapley values. (B) The Shapley dependence plot shows correct predictions of lithic particles with high Shapley values at negative values of *value_mode*. (C) and (D) show for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

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

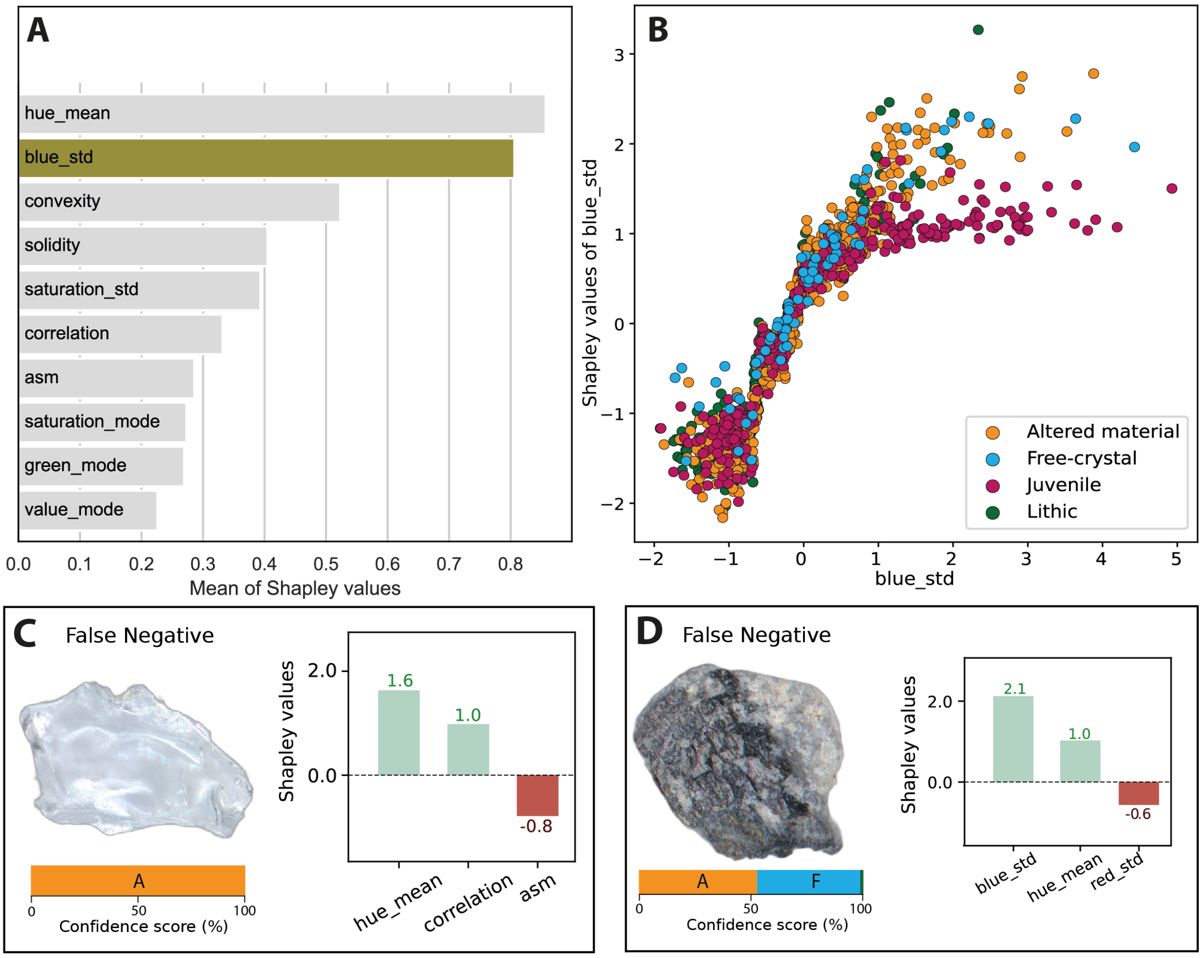

Figure 11. Summary plots to explain predictions of the models for the free-crystal type. (A) Feature importance based on the mean of the Shapley values. (B) Shapley dependence plot. Note that the feature values have been rescaled by a standard scaler. (C) and (D) show for each prediction the particle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

3.3 ViT quantitative evaluation

3.3.1 General evaluation

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

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

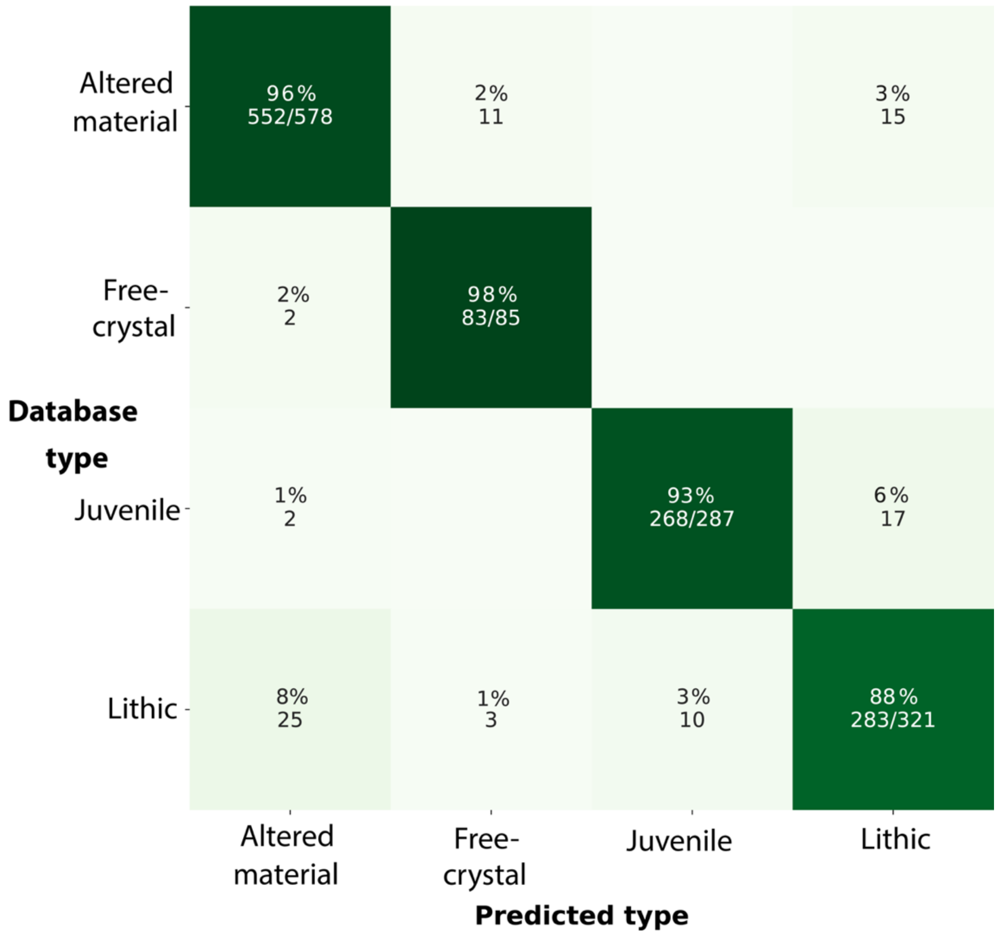

Figure 12. Confusion matrix of the predictions by the ViT image classifier. The percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and otherwise, the False Negative rate (lighter), all percentages with the corresponding number of particles

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

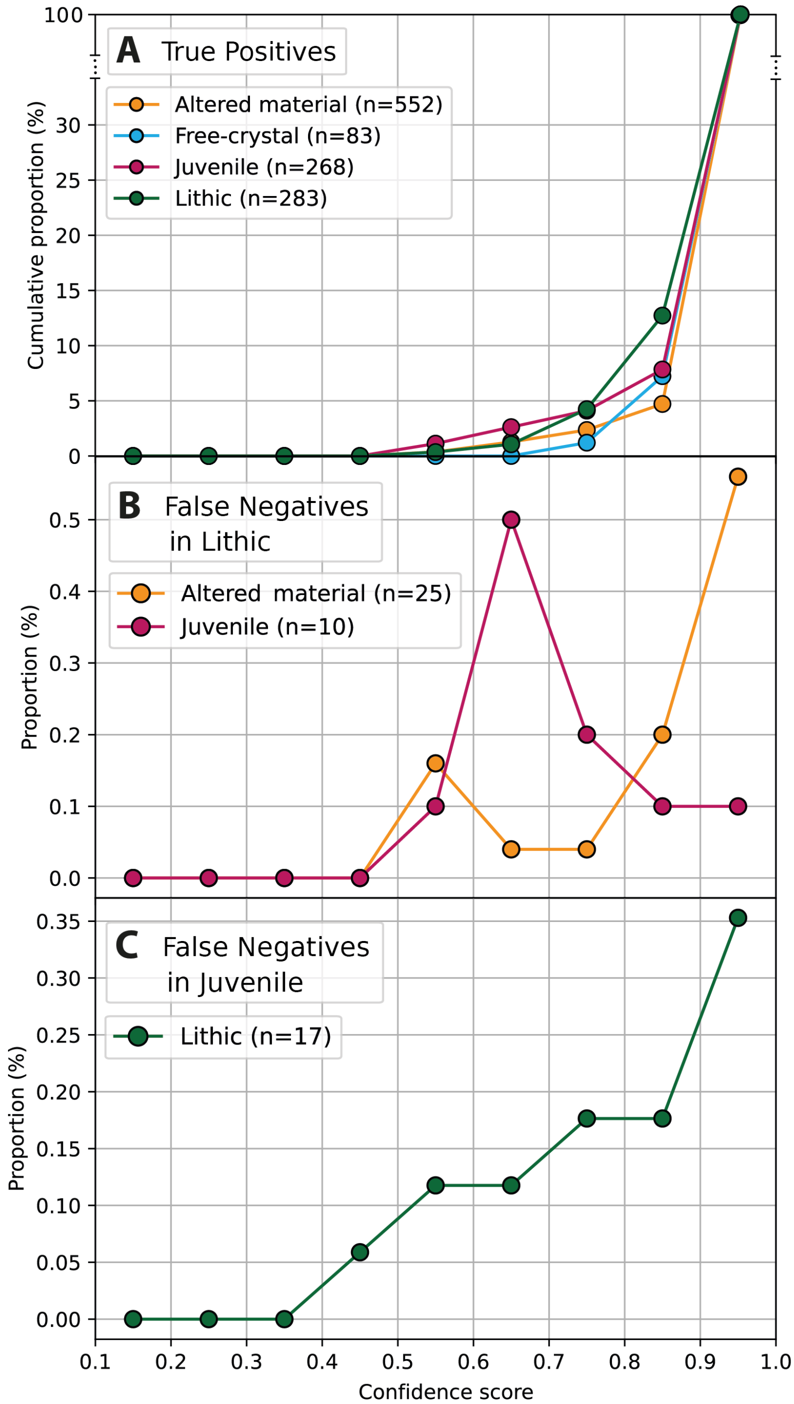

Figure 13. Line plots of the confidence score versus (A) the cumulative proportion of True Positives (TP), (B) False Negatives (FN) in free-crystal and (C) lithic types. The distribution of the data have been plotted into 9 bins of size 0.1. We don't use cumulative proportion in (B) and (C) given the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

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

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

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

(1) Ash particles from phreatic events are in general well classified (*macro F1-score* of 0.95), including the particle main types: altered material (*F1-score* of 0.99), free-crystal (*F1-score* of 0.94) and lithic (*F1-score* of 0.93). The ViT successfully classifies the most common groups of particles in these samples such as hydrothermal aggregates (Figure 15A) and weathered material (Figure 15B).

(2) Particles from samples of dome explosions are classified with the lowest accuracy (*macro F1-score* of 0.85) among the eruptive styles. The ViT accurately classifies free-crystal (*F1-score* of 0.86), altered material (*F1-score* of 0.90) and lithic (*F1-score* of 0.90) types, but is less accurate (*F1-score* of 0.73) for the juvenile type with most False Negatives (*FN*) classified as lithics. However, the confidence scores of some *FN* show a transition between the juvenile and lithic types that has explanatory value. This means that particles may have both juvenile and lithic traits, and thus a measure on the types' prevalence seems more realistic than using mutually exclusive types like in VolcAshDB. Particles with combined traits are common in samples from Nevados de Chillán Volcanic Complex (Figure 15C), which originated from a relatively long-lived dome-forming eruption cycle. An additional challenge is that the ViT confidently classifies as lithics some particles that are labelled as juvenile and, since petrographic classification was not always straightforward (Benet et al., *preprint*), it is difficult to decide whether these are False Negatives, or instead, petrographic classification errors (Figure 15D), especially when ML-based image classifiers have surpassed human performances in other fields (He et al., 2015).

(3) Ash particles from lava fountaining are generally accurately classified (*macro F1-score* of 0.94), between juvenile (*F1-score* of 0.94) and lithic (*F1-score* of 0.88) types. Most of the lithic particles belong to recycled juvenile particles, which are critical to avoid overestimating the amount of juvenile component (D'Oriano et al., 2022) and their identification typically requires examination in the SEM (D'Oriano et al., 2014). The high score suggests that the ViT can discriminate between them to some extent (Figure 15E), but a more robust labelling by a team of experts and a larger dataset containing SEM images is necessary to obtain more robust conclusions. On the other hand, the juvenile particles consist of glossy, smoothed surface, vesicular, elongated glass shards and are accurately classified (Figure 15F).

(4) The ViT accurately classifies ash particles from plinian and subplinian eruptive styles (*macro F1-score* of 0.95), including free crystals (*F1-score* of 0.92), altered material (*F1-score* of 0.93) and juvenile (1.0), but less accurate for lithics (*F1-score* of 0.77). The juvenile particles consist of fragments of pumice and all particles are successfully classified (Figure 15G). In contrast, the lithic particles mostly consist of dull grey fragments with rounded edges, and most of the *FN* are classified as altered material, which may reflect the challenge of classifying particles with incipient weathering into weathered material or lithic (Figure 15H).

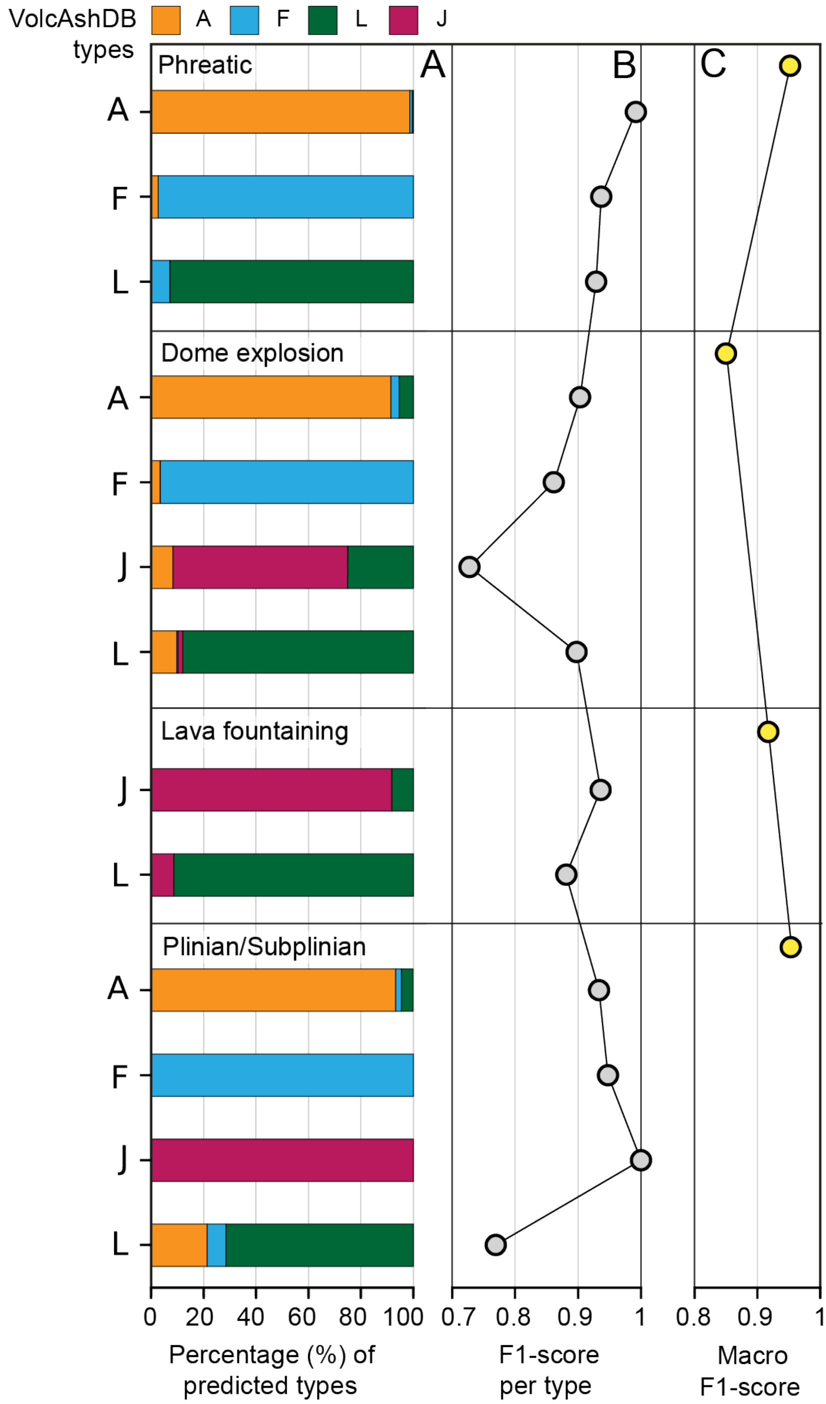

Figure 14. (A) Bar charts showing the percentage of predicted types for each particle type in VolcAshDB. If all predictions were the same as in the database, each bar would be single-colored as follows: orange for altered material (A), light blue for free-crystal (F), magenta for juvenile (J), and dark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

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

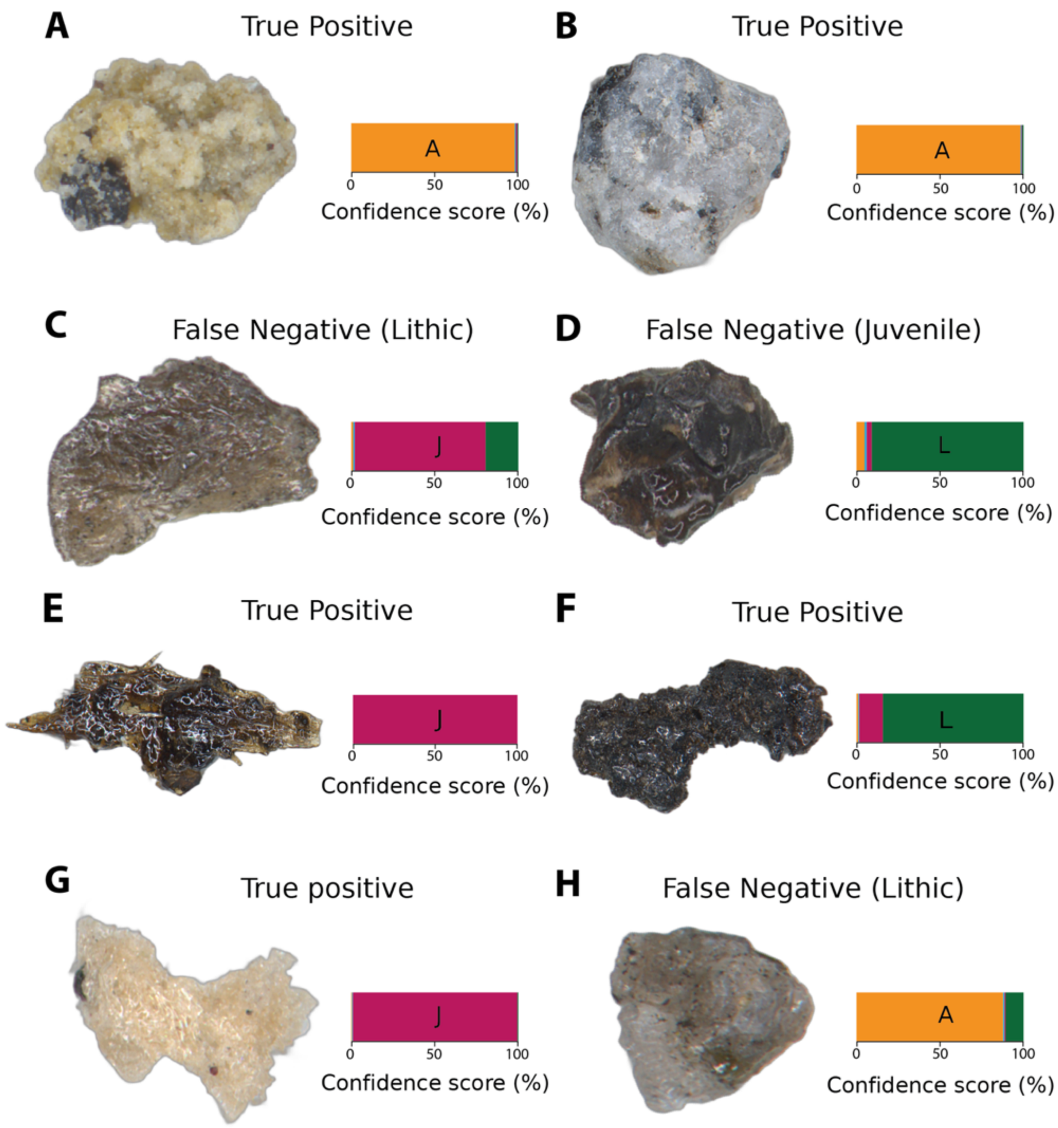

Figure 15. Representative examples of particle images and the predictions and their associated confidence score across eruptive styles, including phreatic (A,B), dome explosion (C,D), lava fountaining (E,F), and subplinian/plinian (G,H). Note that False Negatives contain in brackets the particle type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

4 Discussion

4.1 Comparison between classification using particle's features versus images

We found that, overall, the ViT classifies more accurately with particle images (0.93 of *macro F1-score*) than the XGBoost classifies with the particle features (0.77 of *macro F1-score*). This difference is unlikely to be the XGBoost model itself, which is very popular in the literature and has had best performances amongst models for complex classification tasks (Brownlee, 2016; Chen & Guestrin, 2016; Dhaliwal et al., 2018). One possibility is that the extracted features don't retain certain discriminant information from the images, and as a result, the XGBoost is unable to classify particles such as free crystals (0.57 of *F1-score*). On the other hand, maintaining the physical information associated with features makes the model's outcomes more interpretable (e.g., in classification of volcano-seismic signals; Falcin et al., 2021; Malfante et al., 2018) with xAI methods. This is an important advantage over Vision Transformers, whose main xAI tool consists in a heatmap of the region(s) of attention by the model (Dosovitskiy et al., 2020) but appears insufficient to obtain well founded classification insights for ash particles (Figure 16).

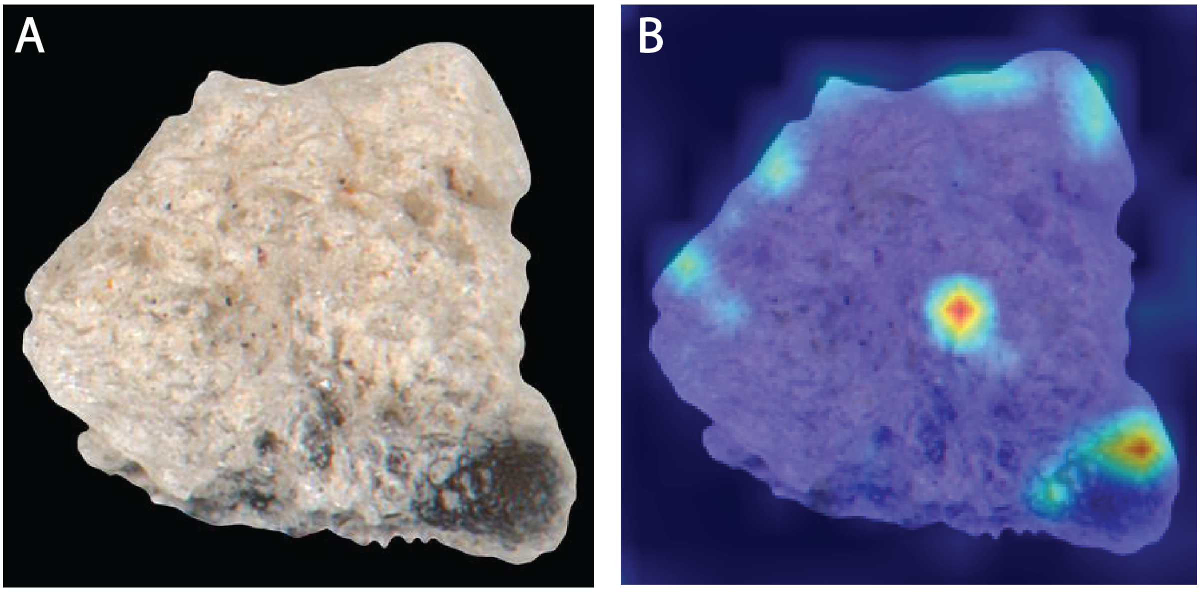

Figure 16. Example of (A) one multi-focused binocular image of a pumice particle from Mount St. Helens (1980), which is overlain by (B) a heatmap of the regions of attention by the base Vision Transformer (Dosovitskiy et al., 2020), typically used for interpreting image classifier's predictions. It does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

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

The XGBoost model gave a medium to high classification performance with *macro F1-score* of 0.77, and using the Shapley values we identified the most discriminant features of each particle type (Table 4). For instance, lithic particles can be distinguished with low values of *value_mode* which correspond to the luster of the particle according to the high Shapley values. This finding agrees with previous studies that use a dull luster (which corresponds to low values of *value_mode*) to identify lithic particles (Miwa et al., 2013). On the other hand, juvenile particles have high Shapley values for the *saturation_mode*. This feature is related to high color intensities as observed under the binocular, but it was not recognized before as a diagnostic observation of the particle type. These two examples belong to particle types that are well classified and for which the Shapley values are reliable. Shapley values obtained from particles that yielded lower accuracies, such as the free crystals, are not reliable, and thus overall performances should be improved. This could be achieved by enhancing the quality and quantity of VolcAshDB dataset by (i) adding particles to balance the dataset, (ii) refining the particle contour in the multi-focused images, so that shape features can measure micro-scaled cavities (Benet et al., *preprint*), and (iii) the inclusion of a new feature that measures the density of lines on the surface, which could be sensitive to planar structures of free crystals.

4.3 Deploying the ViT for automatic particle classification

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

4.4 A ViT particle classifier for volcano monitoring

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

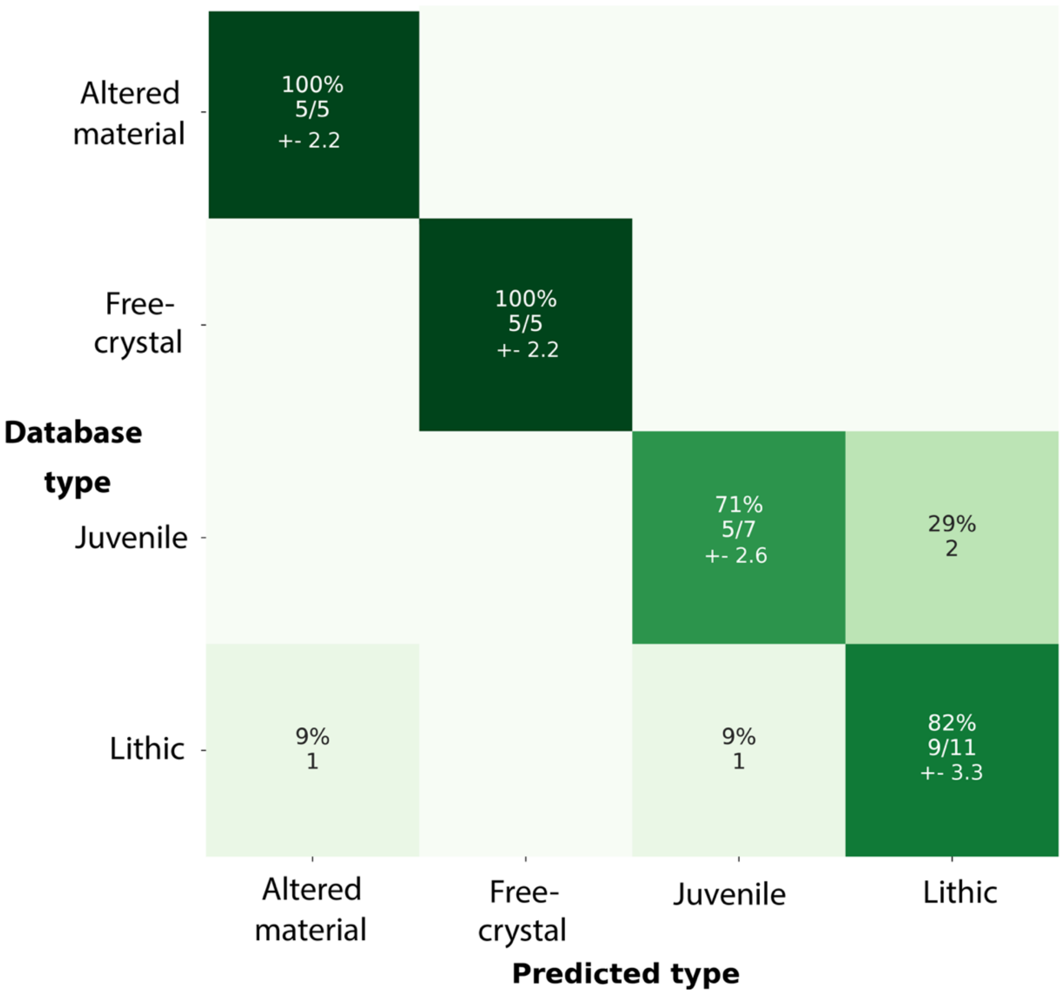

Figure 17. Confusion matrix of the predictions by the ViT image classifier after being fine-tuned with a single-focused, small training set (~370 particles from Benet et al., 2021). The percentages show the True Positive rate if positioned in the diagonal matrix (darker green), and otherwise, the False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

5 Conclusions

Classification of the different particles that make up volcanic ash is not straightforward because diagnostic criteria are not standardized and thus reliable, and systematic identification of a given particle type is not straightforward. In this contribution, we attempt to alleviate this situation by exploring the use of state-of-the-art machine learning-based models to identify the most discriminant features of each particle type, and to evaluate their ability to classify particles. The identified features provide new insights on the recognition of juvenile and lithic particles towards a standardized classification. The image classifier performs at very high accuracies, although the variability across eruption and types shows that its capability to generalize to new samples is still unclear. Higher numbers of particles from a wider variety of eruptions and volcanoes into VolcAshDB coupled to ML models should allow for unbiased comparison of ash samples, and reproducible classification of their particles as a tool for volcano monitoring studies.

Acknowledgments

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

Open Research

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

References

Ahmed, S. N. (2015). Essential statistics for data analysis. In *Physics and Engineering of Radiation Detection*. https://doi.org/10.1016/b978-0-12-801363-2.00009-7

Alvarado, G. E., Mele, D., Dellino, P., de Moor, J. M., & Avard, G. (2016). Are the ashes from the latest eruptions (2010–2016) at Turrialba volcano (Costa Rica) related to phreatic or phreatomagmatic events? *Journal of Volcanology and Geothermal Research*, *327*, 407–415. https://doi.org/10.1016/j.jvolgeores.2016.09.003

Ayyadevara, V. K., & Reddy, Y. (2020). *Modern Computer Vision with PyTorch: Explore deep learning concepts and implement over 50 real-world image applications.* Packt Publishing Ltd.

Bebbington, M. S., & Jenkins, S. F. (2019). *Intra-eruption forecasting*.

Benet, D., Costa, F., Pedreros, G., & Cardona, C. (2021). The volcanic ash record of shallow magma intrusion and dome emplacement at Nevados de Chillán Volcanic complex, Chile. *Journal of Volcanology and Geothermal Research*, *417*. https://doi.org/10.1016/j.jvolgeores.2021.107308

Benet, D., Costa, F., Widiwijayanti, C., Pallister, J., Pedreros, G., Allard, P., Humaida, H., Aoki, Y., & Maeno, F. (2023). VolcashDB: Volcanic ash particle image and classification database. *Preprint in EarthArxiv*. https://doi.org/10.31223/X53659

Biass, S., Jenkins, S. F., Aeberhard, W. H., Delmelle, P., & Wilson, T. (2022). Insights into the vulnerability of vegetation to tephra fallouts from interpretable machine learning and big Earth observation data. *Natural Hazards and Earth System Sciences*, *22*(9), 2829–2855. https://doi.org/10.5194/nhess-22-2829-2022

Brownlee, J. (2016). XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn. *Machine Learning Mastery*.

Brownlee, J. (2020). Imbalanced Classification with Python. *Machine Learning Mastery*, 463.

Cashman, K. V., & Hoblitt, R. P. (2004). Magmatic precursors to the 18 May 1980 eruption of Mount St. Helens, USA. *Geology*, *32*(2), 141–144. https://doi.org/10.1130/G20078.1

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Cioni, R., Pistolesi, M., Bertagnini, A., Bonadonna, C., Hoskuldsson, A., & Scateni, B. (2014). Insights into the dynamics and evolution of the 2010 Eyjafjallajökull summit eruption (Iceland) provided by volcanic ash textures. *Earth and Planetary Science Letters*, *394*(May 2010), 111–123. https://doi.org/10.1016/j.epsl.2014.02.051

D'Oriano, C., Bertagnini, A., Cioni, R., & Pompilio, M. (2014). Identifying recycled ash in basaltic eruptions. *Scientific Reports*, *4*. https://doi.org/10.1038/srep05851

D'Oriano, C., Del Carlo, P., Andronico, D., Cioni, R., Gabellini, P., Cristaldi, A., & Pompilio, M. (2022). Syn-Eruptive Processes During the January–February 2019 Ash-Rich Emissions Cycle at Mt. Etna (Italy): Implications for Petrological Monitoring of Volcanic Ash. *Frontiers in Earth Science*, *10*(February 2019). https://doi.org/10.3389/feart.2022.824872

Dellino, P., & La Volpe, L. (1996). Image processing analysis in reconstructing fragmentation and transportation mechanisms of pyroclastic deposits. The case of Monte Pilato-Rocche Rosse eruptions, Lipari (Aeolian islands, Italy). *Journal of Volcanology and Geothermal Research*, *71*(1), 13–29. https://doi.org/10.1016/0377-0273(95)00062-3

Dhaliwal, S. S., Nahid, A. Al, & Abbas, R. (2018). Effective intrusion detection system using XGBoost. *Information (Switzerland)*, *9*(7). https://doi.org/10.3390/info9070149

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Dürig, T., Bowman, M. H., White, J. D. L., Murch, A., Mele, D., Verolino, A., & Dellino, P. (2018). Particle shape analyzer Partisan - An open source tool for multi-standard two-dimensional particle morphometry analysis. *Annals of Geophysics*, *61*(6). https://doi.org/10.4401/ag-7865

Dürig, T., Ross, P. S., Dellino, P., White, J. D. L., Mele, D., & Comida, P. P. (2021). A review of statistical tools for morphometric analysis of juvenile pyroclasts. *Bulletin of Volcanology*, *83*(11). https://doi.org/10.1007/s00445-021-01500-0

Falcin, A., Métaxian, J. P., Mars, J., Stutzmann, É., Komorowski, J. C., Moretti, R., Malfante, M., Beauducel, F., Saurel, J. M., Dessert, C., Burtin, A., Ucciani, G., de Chabalier, J. B., & Lemarchand, A. (2021). A machine-learning approach for automatic classification of volcanic seismicity at La Soufrière Volcano, Guadeloupe. *Journal of Volcanology and Geothermal Research*, *411*. https://doi.org/10.1016/j.jvolgeores.2020.107151

Feuillard, M., Allegre, C. J., Brandeis, G., Gaulon, R., Le Mouel, J. ., Mercier, J. C., Pozzi, J. P., & Semet, M. . (1983). The 1975–1977 crisis of la Soufriere de Guadeloupe (F.W.I): A still-born magmatic eruption. *Journal of Volcanology and Geothermal Research*, *16*, 317–334.

Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning*, *73*(2), 133–153. https://doi.org/10.1007/s10994-008-5064-8

Gaunt, H. E., Bernard, B., Hidalgo, S., Proaño, A., Wright, H., Mothes, P., Criollo, E., & Kueppers, U. (2016). Juvenile magma recognition and eruptive dynamics inferred from the analysis of ash time series: The 2015 reawakening of Cotopaxi volcano. *Journal of Volcanology and Geothermal Research*, *328*, 134–146. https://doi.org/10.1016/j.jvolgeores.2016.10.013

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. In *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*.

Gianfagna, L., & Di Cecco, A. (2021). Explainable AI. In *Berlin/Heidelberg, Germany: Springer*. https://doi.org/10.3233/FAIA190100

Hall-Beyer, M. (2017). GLCM Texture: A Tutorial. *17th International Symposium on Ballistics*, *2*(March), 18–19.

Haralick, R. M., Dinstein, I., & Shanmugam, K. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Herrera, F., Charte, F., Rivera, A. J., & del Jesus, M. J. (2016). *Multi-Label Classification* (Springer,). https://doi.org/10.4018/jdwm.2007070101

Hincks, T. K., Komorowski, J. C., Sparks, S. R., & Aspinall, W. P. (2014). Retrospective analysis of uncertain eruption precursors at La Soufrière volcano, Guadeloupe, 1975-77: Volcanic hazard assessment using a Bayesian Belief Network approach. *Journal of Applied Volcanology*, *3*(1). https://doi.org/10.1186/2191-5040-3-3

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/cvprw.2009.5206848

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-

Torres, M. Á., & Carvalhais, N. (2022). Wildfire Danger Prediction and Understanding With Deep Learning. *Geophysical Research Letters*, *49*(17), 1–11. https://doi.org/10.1029/2022GL099368

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Le Guern, F., Bernard, A., & Chevrier, R. M. (1980). Soufrière of guadeloupe 1976–1977 eruption — mass and energy transfer and volcanic health hazards. *Bulletin Volcanologique*, *43*(3), 577–593. https://doi.org/10.1007/BF02597694

Lee, J. J., Aime, M. C., Rajwa, B., & Bae, E. (2022). Machine Learning-Based Classification of Mushrooms Using a Smartphone Application. *Applied Sciences (Switzerland)*, *12*(22). https://doi.org/10.3390/app122211685

Leibrandt, S., & Le Pennec, J. L. (2015). Towards fast and routine analyses of volcanic ash morphometry for eruption surveillance applications. *Journal of Volcanology and Geothermal Research*, *297*, 11–27. https://doi.org/10.1016/j.jvolgeores.2015.03.014

Liu, E. J., Cashman, K. V., Miller, E., Moore, H., Edmonds, M., Kunz, B. E., Jenner, F., & Chigna, G. (2020). Petrologic monitoring at Volcán de Fuego, Guatemala. *Journal of Volcanology and Geothermal Research*, *405*(August 2019), 107044. https://doi.org/10.1016/j.jvolgeores.2020.107044

Liu, E. J., Cashman, K. V., & Rust, A. C. (2015). Optimising shape analysis to quantify volcanic ash morphology. *GeoResJ*, *8*, 14–30. https://doi.org/10.1016/j.grj.2015.09.001

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the 2020s*. 11966–11976. https://doi.org/10.1109/cvpr52688.2022.01167

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). *Consistent Individualized Feature Attribution for Tree Ensembles*. *2*. http://arxiv.org/abs/1802.03888

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Maeno, F., Nakada, S., Yoshimoto, M., Shimano, T., Hokanishi, N., Zaennudin, A., & Iguchi, M. (2019). A sequence of a plinian eruption preceded by dome destruction at Kelud volcano, Indonesia, on February 13, 2014, revealed from tephra fallout and pyroclastic density current deposits. *Journal of Volcanology and Geothermal Research*, *382*, 24–41. https://doi.org/10.1016/j.jvolgeores.2017.03.002

Malfante, M., Mura, M. D., Métaxian, J., & Mars, J. I. (2018). *Machine Learning for Volcano-Seismic Signals*. *March*, 20–30.

Mandal, S., Mones, S. M. B., Das, A., Balas, V. E., Shaw, R. N., & Ghosh, A. (2021). Single shot detection for detecting real-time flying objects for unmanned aerial vehicle. In *Artificial Intelligence for Future Generation Robotics*. INC. https://doi.org/10.1016/B978-0-323-85498-6.00005-8

Marzocchi, W., Newhall, C., & Woo, G. (2012). The scientific management of volcanic crises.

Journal of Volcanology and Geothermal Research, *247–248*, 181–189. https://doi.org/10.1016/j.jvolgeores.2012.08.016

Mishra, P. (2022). Practical Explainable AI Using Python. In *Practical Explainable AI Using Python*. https://doi.org/10.1007/978-1-4842-7158-2

Miwa, T., Geshi, N., & Shinohara, H. (2013). Temporal variation in volcanic ash texture during a vulcanian eruption at the sakurajima volcano, Japan. *Journal of Volcanology and Geothermal Research*, *260*, 80–89. https://doi.org/10.1016/j.jvolgeores.2013.05.010

Miwa, T., Toramaru, A., & Iguchi, M. (2009). Correlations of volcanic ash texture with explosion earthquakes from vulcanian eruptions at Sakurajima volcano, Japan. *Journal of Volcanology and Geothermal Research*, *184*(3–4), 473–486. https://doi.org/10.1016/j.jvolgeores.2009.05.012

Miyagi, I., Geshi, N., Hamasaki, S., Oikawa, T., & Tomiya, A. (2020). Heat source of the 2014 phreatic eruption of Mount Ontake, Japan. *Bulletin of Volcanology*, *82*(4). https://doi.org/10.1007/s00445-020-1358-x

Molnar, C. (2021). Interpretable Machine Learning. *Queue*, *19*(6), 28–56. https://doi.org/10.1145/3511299

Moran, S. C., Newhall, C., & Roman, D. C. (2011). Failed magmatic eruptions: Late-stage cessation of magma ascent. *Bulletin of Volcanology*, *73*(2), 115–122. https://doi.org/10.1007/s00445-010-0444-x

Newhall, C. G., & Punongbayan, R. S. (1996). The narrow margin of successful volcanic-risk mitigation. In *Monitoring and mitigation of volcano hazards* (pp. 807–838). Springer Science & Business Media.

Nurfiani, D., & Bouvet de Maisonneuve, C. (2018). Furthering the investigation of eruption styles through quantitative shape analyses of volcanic ash particles. *Journal of Volcanology and Geothermal Research*, *354*, 102–114. https://doi.org/10.1016/j.jvolgeores.2017.12.001

Owen, L. (2022). *Hyperparameter Tuning with Python*.

Paladio-Melasantos, M. L., Solidum, R. U., Scott, W. E., Quiambao, R. B., Umbal, J. V, Rodolfo, K. S., Tubianosa, B. S., Delos Reyes, P. J., Alonso, R. A., & Ruerlo, H. B. (1996). Tephra falls of the 1991 eruptions of Mount Pinatubo. *In: Newhall, C.G. (Editor) & Others, Fire and Mud; Eruptions and Lahars of Mount Pinatubo, Philippines, Philippine Institute of Volcanology and Seismology, Quezon City*, *layer D*, 413–535. https://doi.org/10.1159/000153100

Panati, C., Wagner, S., & Bruggenwirth, S. (2022). Feature Relevance Evaluation using Grad-CAM, LIME and SHAP for Deep Learning SAR Data Classification. *Proceedings International Radar Symposium*, *2022-Septe*, 457–462.

Pardo, N., Avellaneda, J. D., Rausch, J., Jaramillo-Vogel, D., Gutiérrez, M., & Foubert, A. (2020). Decrypting silicic magma/plug fragmentation at Azufral crater lake, Northern Andes: insights from fine to extremely fine ash morpho-chemistry. *Bulletin of Volcanology*, *82*(12). https://doi.org/10.1007/s00445-020-01418-z

Pardo, N., Cronin, S. J., Németh, K., Brenna, M., Schipper, C. I., Breard, E., White, J. D. L., Procter, J., Stewart, B., Agustín-Flores, J., Moebis, A., Zernack, A., Kereszturi, G., Lube, G., Auer, A., Neall, V., & Wallace, C. (2014). Perils in distinguishing phreatic from phreatomagmatic ash; insights into the eruption mechanisms of the 6 August 2012 Mt. Tongariro eruption, New Zealand. *Journal of Volcanology and Geothermal Research*, *286*, 397–414. https://doi.org/10.1016/j.jvolgeores.2014.05.001

Re, G., Corsaro, R. A., D'Oriano, C., & Pompilio, M. (2021). Petrological monitoring of active volcanoes: A review of existing procedures to achieve best practices and operative protocols during eruptions. *Journal of Volcanology and Geothermal Research*, *419*, 107365. https://doi.org/10.1016/j.jvolgeores.2021.107365

Romero, J. E., Burton, M., Cáceres, F., Taddeucci, J., Civico, R., Ricci, T., Pankhurst, M. J., Hernández, P. A., Bonadonna, C., Llewellin, E. W., & Pistolesi, M. (2022). *The initial phase of the 2021 Cumbre Vieja ridge eruption (Canary Islands): Products and dynamics controlling edifice growth and collapse*. *431*(July). https://doi.org/10.1016/j.jvolgeores.2022.107642

Ross, P. S., Dürig, T., Comida, P. P., Lefebvre, N., White, J. D. L., Andronico, D., Thivet, S., Eychenne, J., & Gurioli, L. (2022). Standardized analysis of juvenile pyroclasts in comparative studies of primary magma fragmentation; 1. Overview and workflow. *Bulletin of Volcanology*, *84*(1), 1–29. https://doi.org/10.1007/s00445-021-01516-6

Rowe, M. C., Thornber, C. R., & Kent, A. J. R. (2008). Identification and Evolution of the Juvenile Component in. *A Volcano Rekindled: The Renewed Eruption of Mount St. Helens, 2004–2006*, 2004–2006.

Shapley, L. S. (1953). A Value for n-Person Games, in: Contributions to the Theory of Games II. *Contributions to the Theory of Games*, 307–318. https://doi.org/https://doi.org/10.1515/9781400881970-018

Shoji, D., Noguchi, R., Otsuki, S., & Hino, H. (2018). *Classification of volcanic ash particles using a convolutional neural network and probability*. 1–12. https://doi.org/10.1038/s41598-018-26200-2

Sujatha, R., Chatterjee, J. M., Jhanjhi, N. Z., & Brohi, S. N. (2021). Performance of deep learning vs machine learning in plant leaf disease detection. *Microprocessors and Microsystems*, *80*(November 2020). https://doi.org/10.1016/j.micpro.2020.103615

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *30th International Conference on Machine Learning, ICML 2013*, *PART 3*, 2176–2184.

Suzuki, Y., Nagai, M., Maeno, F., Yasuda, A., Hokanishi, N., Shimano, T., Ichihara, M., Kaneko, T., & Nakada, S. (2013). Precursory activity and evolution of the 2011 eruption of Shinmoe-dake in Kirishima volcano-insights from ash samples. *Earth, Planets and Space*, *65*(6), 591–607. https://doi.org/10.5047/eps.2013.02.004

Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J. A., Zou, J., & Rubin, D. L. (2021). Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. *Scientific Reports*, *11*(1), 1–9. https://doi.org/10.1038/s41598-021-87762-2

Tilling, R. ~I. (2008). The critical role of volcano monitoring in risk reduction. *Advances in Geosciences*, *14*, 3–11.

Verdhan, V. (2020). Supervised Learning with Python. In *Supervised Learning with Python*. https://doi.org/10.1007/978-1-4842-6156-9

Watanabe, K., Danhara, T., Watanabe, K., Terai, K., & Yamashita, T. (1999). Juvenile volcanic glass erupted before the appearance of the 1991 lava dome, Unzen volcano, Kyushu, Japan. *Journal of Volcanology and Geothermal Research*, *89*(1–4), 113–121. https://doi.org/10.1016/S0377-0273(98)00127-9

Journal of Geochemistry, Geophysics, Geosystems

Supporting Information for

Volcanic ash classification through Machine Learning

Damià Benet[1,2,3], Fidel Costa[1], Christina Widiwijayanti[2]

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

[2] EOS, Earth Observatory of Singapore, Nanyang Technological University, Singapore.

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

Contents of this file

Introduction

The following information below includes details on the used features and their abbreviations, a series of tables with the accuracies obtained through different experiments to choose amongst the different types of machine learning-based models and to choose the optimal hyperparameters. Other supplementary information includes the results of classification from the One-Versus-One and One-Versus-Rest classification techniques, two figures to summarize classification scores across different types of models, and an example of a confusion matrix.

Table S1. List of measured features, and their abbreviation and calculation. The reader is referred to Benet et al., (*preprint*) for more details and a reference.

Feature	Abbreviation	Equation
Convexity	convexity	P_h/P_p
Rectangularity	rectangularity	$\frac{P_p}{2H+2W}$
Elongation	elongation	$\frac{{D_{MaxFeret}}^2}{E_{maj}}$
Roundness	roundness	$\frac{4A_p}{\pi {D_{MaxFeret}}^2}$
Circularity by Dellino and la Volpe (1996)	circ_dellino	$\frac{P_p}{2\sqrt{\pi A_p}}$

Circularity by Cioni et al. (2014)	circ_cioni	$\frac{4\pi A_p}{{P_p}^2}$
Solidity	solidity	$\frac{A_p}{2H + 2W}$
Aspect ratio	aspect_rat	W/H
Compactness	compactness	$\frac{A_p}{HW}$
Contrast	contrast	$\sum_{i,j=0}^{levels-1} P_d^\theta (i-j)^2$
Dissimilarity	dissimilarity	$\sum_{i,j=0}^{levels-1} P_d^\theta \lvert i-j \rvert$
Homogeneity	homogeneity	$\sum_{i,j=0}^{levels-1} \frac{P_d^\theta(i,j)}{1+(i-j)^2}$
ASM	asm	$\sum_{i,j=0}^{levels-1} P_d^\theta(i,j)^2$
Energy	energy	$\sqrt{ASM}$
Correlation	correlation	$\sum_{i,j=0}^{levels-1} P_d^\theta \left[\frac{(i-\mu_i)\left(j-\mu_j\right)}{\sqrt{(\sigma_i^2)\left(\sigma_j^2\right)}}\right]$
Channel[1] mean	channel_mean (e.g., hue_mean)	$\frac{1}{N}\sum_{i=i}^{n} x_i$
Channel standard dev	channel_std (e.g., value_std)	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2}$
Channel mode	channel_mode (e.g., red_mode)	Computationally found as the most common value in the array by Scipy's stats.mode function

Symbols used: A_p, particle area; P_p, particle perimeter; A_h, hull area; P_h, hull perimeter; W, width of bounding box; H, height of bounding box; $D_{MaxFeret}$, Feret maximum diameter the maximum distance between two parallel lines tangential to the particle outline; E_{maj}, major ellipse axis; levels, pixel intensities from the ROI used for Grey-Level Cooccurrence-Matrix (GLCM) calculation; $P_d^\theta(i,j)$, probability of pixel pairs at a given distance (d) and angle (θ) in GLCM; μ_i, GLCM mean; σ_i^2, standard deviation; N, number of pixels per channel; x_i, pixel value; $\overline{x}$, mean of pixel values.

Table S2: Optimal hyperparameter obtained from the highest cross-validation score for various models.

Hyperparameter	XGB	RF	DTC	KNN	GBC
colsample_bytree	0.47	–	–	–	–

learning_rate	0.01	–	–	–	0.01
max_depth	10	7	7	–	10
n_estimators	45	22	–	–	48
reg_alpha	1	–	–	–	–
reg_lambda	1	–	–	–	–
min_samples_split	–	22	25	–	30
n_neighbors	–	–	–	5	–

Table S3: Evaluation of optimized models. Support indicates the number of particles used for evaluation.

	XGB	RF	DT	KNN	GBC
precision	0.75	0.68	0.65	0.64	0.69
recall	0.75	0.72	0.69	0.68	0.72
F1-score	0.75	0.69	0.65	0.64	0.70
accuracy	0.81	0.75	0.70	0.70	0.77
support	315	315	315	315	315

Table S4: Statistical measures of mean, first and second standard deviations of the distribution of F1-scores by particle type and their aggregated macro F1-score.

	Altered material	Free-crystal	Lithic	Juvenile	Overall
Mean	0.87	0.57	0.73	0.88	0.76
Standard deviation	0.01	0.04	0.02	0.01	0.015
Second standard deviation	0.02	0.09	0.04	0.03	0.03
Particles in train	2310	326	1122	1281	5040
Particles in test	577	81	280	320	1260

Table S5. List of the base hyperparameters for each model provided by their authors. Note, in bold, the name of the model according to the authors.

Hyperparameter	Value
Vision Transformer (ViT-B{16,32})	**Dosovitskiy et al., 2020**
Learning rate	$8x10^{-4}$
Epochs	7
Residual neural network (R50x{1,2})	**He et al., 2016**
Learning rate	10^{-3}
Epochs	7

Convolutional neural network (ConvNeXt-T/S/B/L/XL	**Liu et al., 2022**
Optimizer	Adam
Learning rate	$5x10^{-5}$
Epochs	30

Table S6. Accuracies obtained from grid search at varying learning rate and batch size.

Batch size \ Learning rate	6e-4	8e-4	1e-5	3e-5
4	86.66	87.32	87.18	86.66
8	86.55	87.79	86.55	86.55
16	86.93	87.50	86.97	87.25
32	86.13	86.99	87.07	87.08
64	86.34	87.21	86.87	87.08

Table S7. Comparison of optimizers' performance based on accuracy.

Optimizer	**Accuracy**
AdamW	87.50%
SGD	81.72%
Adagrad	85.59%

Table S8. *F1-scores* obtained from the OVO and OVR strategies for each particle type, and their unweighted average (i.e., macro), for all particles in the test set (Overall) and across the associated binary classifiers. These measurements have an estimated precision of ±0.03 (see 'Effect of the train and test split' in Section 2.2.6 for its calculation).

	One-vs-One (OVO)							One-vs-Rest (OVR)				
	Overall	*F* vs *A*	*F* vs *L*	*F* vs *J*	*A* vs *L*	*A* vs *J*	*L* vs *J*	Overall	*A* vs Rest	*F* vs Rest	*L* vs Rest	*J* vs Rest
F1-score (macro)	0.75	0.81	0.78	0.9	0.88	0.97	0.84	0.76	0.89	0.74	0.82	0.92
F[1]	0.56	0.67	0.64	0.82	–	–	–	0.55	–	0.52	–	–
A[2]	0.9	0.95	–	–	0.92	0.98	–	0.88	0.88	–	–	–
L[3]	0.71	–	0.92	–	0.86	–	0.84	0.73	–	–	0.73	–
J[4]	0.85	–	–	0.96	–	0.97	0.85	0.88	–	–	–	0.88
Rest[5]								–	0.89	0.97	0.9	0.96

[1]*F*: Free-crystal [2]*A*: Altered material [3]*L*: Lithic [4]*J*: Juvenile

[5]Rest includes all the particles that do not belong to the class of interest (e.g., Lithic vs Non-lithic)

Table S9. *F1-scores* obtained from the ViT classifier of each particle type and their unweighted *F1-score* average (i.e., macro) for all particles in the test set (Overall), and across volcanoes and eruptive styles.

		Volcano					Eruptive style			
	Overall	Soufrière de Guadeloupe	Merapi	Nevados de Chillán	Cumbre Vieja	Kelud	Phreatic	Dome explosion	Lava fountaining	Sub-plinian / Plinian
F1-score	0.93	0.95	0.80	0.85	0.91	0.91	0.95	0.85	0.91	0.95
F[1]	0.91	0.90	0.72	0.95	–	0.92	0.94	0.86	–	0.92
A[2]	0.95	0.99	0.95	0.80	–	0.93	0.99	0.90	–	0.93
L[3]	0.89	0.96	0.75	0.91	0.88	0.77	0.93	0.90	0.88	0.77
J[4]	0.95	–	–	0.72	0.94	1	–	0.73	0.94	1

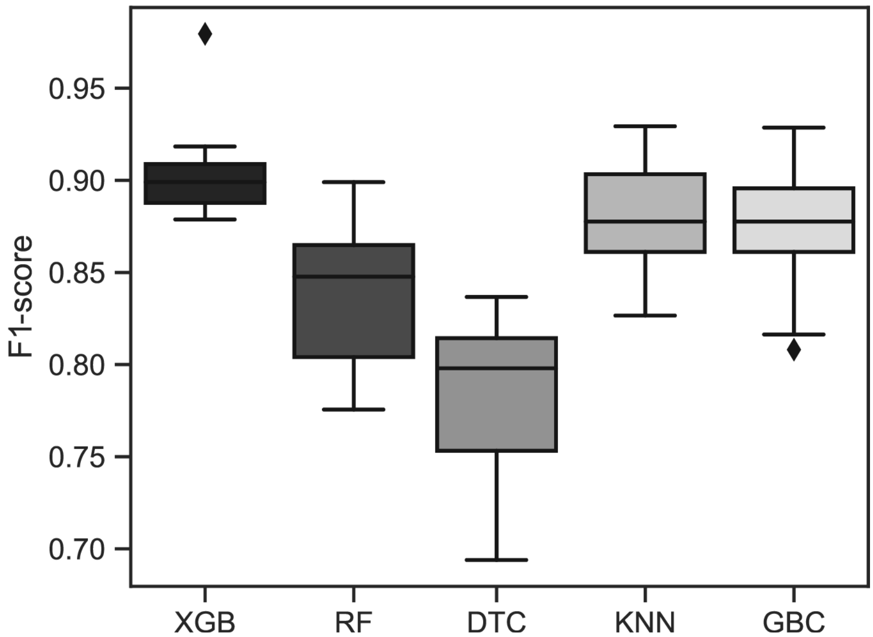

Figure S1. Whisker plots of the *F1-score* values obtained from 10-fold cross validation (see 'Hyperparameter optimization' in Section 2.2.2 for an explanation of this technique) of Extreme Gradient Boosting (XGB), Random Forest (RF), Decision Tree Classifier (DTC), K-Nearest Neighbor (KNN) and Gradient Boost Classifier (GBC). Performances are measured with the *F1-score* (see 'Model evaluation' in Section 2.2.3 for its calculation).

Each whisker plot shows the median (horizontal line), 25th and 75th percentiles (box upper and lower side). Whisker lengths are at 1.5 times the interquartile ranges, beyond which are the outliers (diamonds).

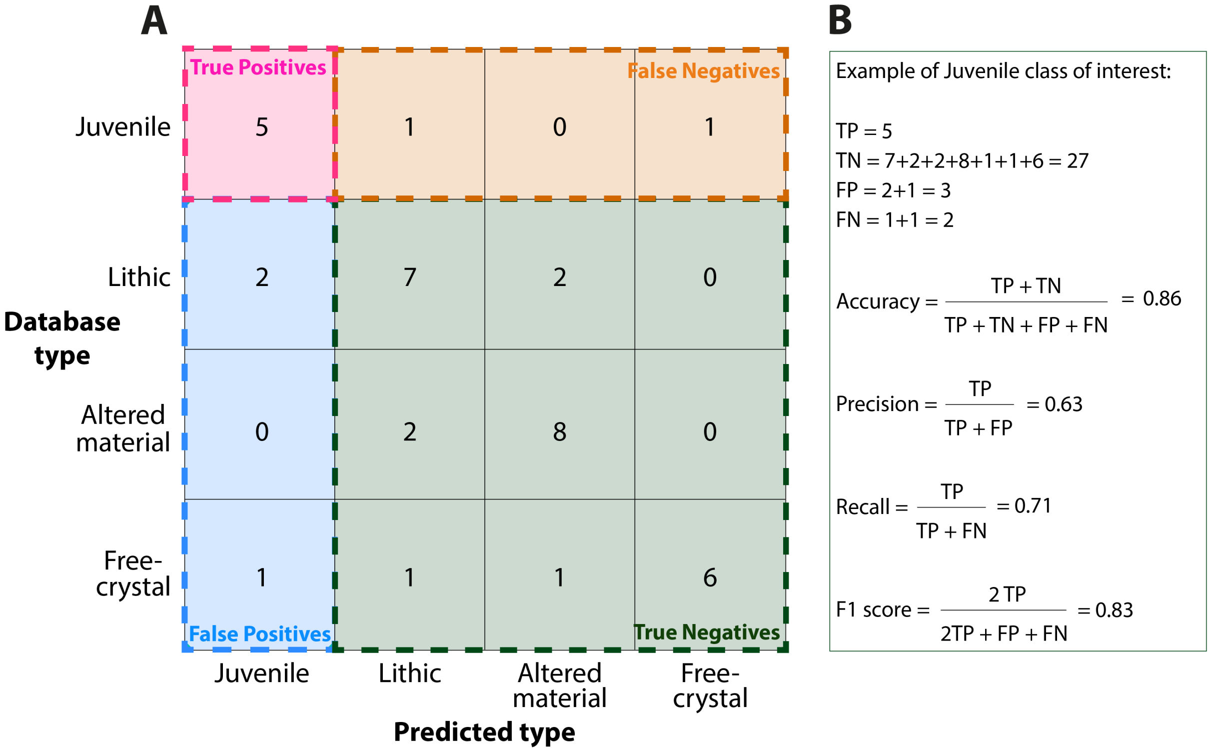

Figure S2. (A) Example of a confusion matrix for a four particle-classes classifier and (B) calculation of the main metrics taking juvenile as the class of interest.

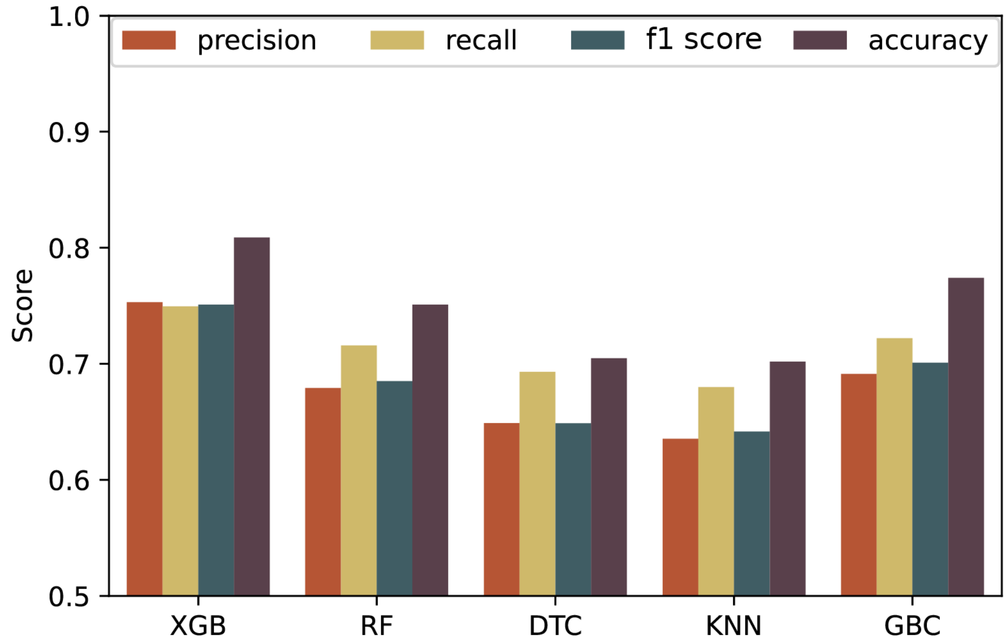

Figure S3. Evaluation of the models' performance with the test set after hyperparameter optimization based on the precision, recall, F1-score and accuracy.