

Ash Particle Classification Data

This directory contains classification results, quality metrics, and reference data for your submitted ash particles, organized into the following subdirectories:

Directory Structure

```
.
  README.md
  README.pdf
  ood_analysis/          # Out-of-distribution detection metrics
  results/               # Classification results and reference data
  visualizations/        # Visual outputs and predictions
```

Files

results/results.csv

This file contains the classification results for user-provided ash particles. Each row corresponds to one particle and includes:

- **Physical Features**
 - **Shape descriptors:**
'aspect_rat', 'solidity', 'convexity', 'circularity_cioni', 'circularity_dellino', 'rectangularity', 'compactness', 'elongation', 'roundness', 'eccentricity_moments', 'eccentricity_ellipse', 'circ_rect', 'comp_elon', 'circ_elon', 'rect_comp'
 - **Texture descriptors:**
'contrast', 'dissimilarity', 'homogeneity', 'energy', 'correlation', 'asm'
 - **Color descriptors (RGB & HSV statistics):**
'red_mean', 'red_std', 'red_mode',
'green_mean', 'green_std', 'green_mode',
'blue_mean', 'blue_std', 'blue_mode',
'hue_mean', 'hue_std', 'hue_mode',
'saturation_mean', 'saturation_std', 'saturation_mode',
'value_mean', 'value_std', 'value_mode'
- **Model Outputs**
 - 'Particle type associated by the model': The predicted class label.
 - 'Model probability of being Altered_material': The probability that the ash particle belong to the class Altered material.
 - 'Model probability of being Free_crystal': The probability that the ash particle belong to the class Free Crystal.
 - 'Model probability of being Juvenile': The probability that the ash particle belong to the class Juvenile.
 - 'Model probability of being Lithic': The probability that the ash particle belong to the class Lithic.
- **Model Information**

- ‘publication_doi’: The DOI of the model used to classify the ash particles.

results/volcashedb_particles.csv

This file includes a reference dataset of particles from the VolcAshDB collection, with similar physical features to the ones in ‘results.csv’. These particles, however, have been **annotated by human experts**, and therefore do not contain the model-generated classification or associated probabilities.

- **Physical Features:**
Same descriptors as in ‘results.csv’ (shape, texture, color).
- **Human Annotations:**
 - ‘publication_names’: A list of labels given to the ash particles.
 - ‘publication_dois’: A list of DOIs for the models used to classify the ash particles.

These columns enable users to compare model results against expert annotations and access the original data sources.

ood_analysis/ood_scores.csv (Out-of-Distribution Detection)

This file provides quality metrics to assess how similar your submitted particles are to the training data used to build the classification model. Each row corresponds to one particle and includes:

- **‘knn_score’:** K-Nearest Neighbor distance score. Measures the average distance to the k-nearest training samples in feature space. Lower values indicate particles more similar to the training set.
- **‘md_score’:** Mahalanobis Distance score. A statistical measure of how far a particle is from the training distribution center, accounting for feature correlations. Lower values indicate better alignment with training data characteristics.
- **‘warning’:** A color-coded quality flag indicating reliability:
 - **GREEN:** Images whose scores fall within the 95th percentile of the training data distribution - particle is well within the training distribution
 - **AMBER:** Images whose scores fall between the 95th and 99th percentiles of the training data distribution - particle shows some deviation from training data

- **RED**: Images whose scores fall above the 99th percentile of the training data distribution - particle is significantly different from training data; predictions should be interpreted with caution
- **‘ood_flag’**: Boolean flag (‘True’/‘False’) indicating whether the particle is considered out-of-distribution (OOD). ‘True’ means the particle differs substantially from the training set.
- **‘filepath’**: Path to the corresponding particle image.

Interpretation Guide:

Particles flagged as AMBER or RED may have less reliable classification results because they differ from the particles the model was trained on. Consider these predictions as preliminary and verify them with additional analysis or expert review.

`ood_analysis/ood_summary.csv`

This summary file provides an overview of the out-of-distribution analysis for your entire submission:

- **‘warning’**: The quality flag category (GREEN, AMBER, or RED).
- **‘count’**: Number of particles in each category.
- **‘percentage’**: Percentage of your total submission in each category.

This allows you to quickly assess the overall reliability of the classification results for your dataset. A high percentage of GREEN particles indicates a distribution close to the training data distribution, while a high percentage of RED particles suggests your samples may be from a different population than the training data.

`ood_analysis/umap_ood_visualization.png`

This visualization shows how submitted particles relate to the training dataset in a 2D projection of a high-dimensional feature space using UMAP (Uniform Manifold Approximation and Projection).

- **Black crosses (+)**: Training set particles - these form the reference distribution.
- **Colored dots**: Your submitted particles, color-coded by their quality flag:
 - **Green dots**: Images whose scores fall within the 95th percentile of the training data distribution
 - **Orange/Amber dots**: Images whose scores fall between the 95th and 99th percentiles of the training data distribution

- **Red dots:** Images whose scores fall above the 99th percentile of the training data distribution

How to interpret this visualization (with caution)

UMAP is a **non-linear dimensionality reduction technique** that aims to preserve **local neighborhood structure** when projecting data from an N-dimensional feature space into two dimensions. As a consequence, **global distances and cluster sizes are not preserved**.

Therefore, the following points are critical for correct interpretation:

- **Visual proximity in UMAP does not guarantee statistical similarity** in the original feature space. A submitted particle may appear embedded within a dense cluster of training data in 2D, yet still be **out-of-distribution** according to its high-dimensional distance metrics.
- **Conversely, apparent isolation in the UMAP projection does not necessarily imply out-of-distribution behavior.** Some particles may be visually separated due to projection artifacts, even though they remain within the training distribution in the original space.
- **Color coding (GREEN / AMBER / RED) reflects quantitative OOD scores**, computed in the original feature space relative to the training distribution, **not the UMAP geometry**.
- **UMAP compresses and distorts distances** in order to optimize neighborhood preservation; as a result, **relative distances between distant points should not be interpreted literally**.

`visualizations/particle_predictions_part*.png`

One or more image files named like `particle_predictions_part1.png`, `particle_predictions_part2.png`, ..., `particle_predictions_partN.png` (pattern: `particle_predictions_part*.png`) display your submitted particles with their predicted classifications overlaid as colored labels. Each file contains a subset of particles and shows the model-predicted type (Juvenile, Lithic, Free Crystal, or Altered Material) for quick visual inspection of the classification results.

`visualizations/legend.png`

A legend showing the color-coding used for each particle type in the predictions visualization. This helps interpret the `particle_predictions.png` image.

Notes

- All particles share the same feature space (shape, texture, color) to allow direct comparison.
- Model predictions are **only** available in `results/results.csv`.
- Human-labeled references are **only** available in `results/volcashdb_particles.csv`.

References

- For further reading, please refer to the DOIs included in the reference dataset to access published sources and annotation protocols.

Contact

If you have any questions or need assistance, please reach out to us at volcashdb@ipgp.fr.