Supplementary Information (SI) for Journal of Analytical Atomic Spectrometry.
This journal is © The Royal Society of Chemistry 2026

Supplementary Information

A Data-Driven, Post-Acquisition Quality Diagnostic Pipeline for Isotope Analysis by MC-ICP-MS

Chufan Zhou[a,b], *Qiang Huang*[a,*], *Yang Tang*[a], *Ying Zhong*[a], *Xinbin Feng*[a]

[a] State Key Laboratory of Environmental Geochemistry, Institute of Geochemistry, Chinese Academy of Sciences, Guiyang, Guizhou 550081, China

[b] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author

Email: huangqiang@mail.gyig.ac.cn

Table of Contents

S1. Automated data export and calculation 1
S1.1 Data export script 1
S1.2 Data calculation principle 1
S2. Confirm the 2SD range and segmentation method 3
S3. Exploration of abnormal data 4
S4. Introduction to ML model and algorithms 6
S4.1 Algorithm descriptions 6
S4.1.1 Logistic Regression (LR) 6
S4.1.2 Random Forest (RF) 6
S4.1.3 Support Vector Machine (SVM) 6
S4.1.4 K-Nearest Neighbors (K-NN) 6
S4.1.5 Decision Tree (DT) 6
S4.1.6 Gradient Boosting (GB) 6
S4.1.7 Naive Bayes (NB) 6
S4.1.8 XGBoost (XGB) 7
S4.1.9 Bagging Random Forest (BagRF) 7
S4.2 Sampling methodologies for imbalanced data 7
S4.2.1 SMOTE (Synthetic Minority Over-sampling Technique) 7
S4.2.2 ADASYN (Adaptive Synthetic Sampling) 7
S4.2.3 SMOTEENN (SMOTE + Edited Nearest Neighbors) 7
S4.2.4 UnderSampling 7
S4.2.5 Cost-Sensitive Learning 8
S4.3 Enhanced feature description 9
S4.4 Train your own ML expert experience model 10
S5. Supplement to the results 11
S5.1 Binary classification section 11
S5.2 Diagnostic multi-class classification of anomalies section 11
REFERENCES 13

S1. Automated data export and calculation

S1.1 Data export script

To enable high-throughput and reproducible data processing, we developed a desktop automation tool (HG MC Auto software, Option 1) for batch conversion of raw measurement files (.dat format from Data Evaluation software) into structured .csv files. The tool integrates instrumental parameters (.log files) with analytical data while calculating essential statistics (mean and absolute standard error).Given the proprietary nature of the Data Evaluation software, we implemented two complementary automation strategies:

1. Image recognition-based navigation using template matching algorithms;
2. Deterministic coordinate-based scripting employing predefined screen coordinates;

Table S1 compares these approaches across three operational criteria relevant to analytical chemistry workflows.

Table S1. Comparison of different export methods

	Convenience	Cross-instrument portability	Later maintenance costs
Image Recognition	Poor	Good	Low
Coordinate positioning	Good	Poor	High

The automated pipeline comprises two sequential modules:

Export Module (Option 1): Executes GUI automation to activate the Data Evaluation interface, sequentially open .dat files, trigger ASCII export functions, and save outputs as .csv format

The system incorporates configurable parameters for screen coordinate calibration, comprehensive operation logging, and automated error recovery mechanisms (including limited retry attempts and modal window handling via Esc key simulation). This approach maintains compatibility with the original software interface while significantly reducing manual intervention.

S1.2 Data calculation principle

Hg isotope mass fractionation is usually expressed as δxxxHg. Using NIST SRM 3133Hg (abbreviated as NIST 3133 in the formulas) as the reference standard, the calculation formulas are as follows (1) or (2),

with all relevant formulas referencing previous studies[1-4]:

$$\delta^{xxx}Hg(‰) = \left\{\left[\frac{(^{xxx}Hg/^{198}Hg)_{Sample}}{(^{xxx}Hg/^{198}Hg)_{NIST\ 3133}}\right] - 1\right\} \times 1000 \quad (1)$$

or[5]

$$\delta^{xxx}Hg(‰) = \left\{\left[\frac{2 \times (^{xxx}Hg/^{198}Hg)_{Sample}}{(^{xxx}Hg/^{198}Hg)_{NIST\ 3133\_Prev} + (^{xxx}Hg/^{198}Hg)_{NIST\ 3133\_Next}}\right] - 1\right\} \times 1000 \quad (2)$$

Hg is an element that exhibits both odd and even non-mass-dependent fractionation[3], and its calculation formula is as follows[4, 6, 7]:

$$\Delta^{199}Hg = \delta^{199}Hg - (\delta^{202}Hg \times 0.2520) \quad (3)$$

$$\Delta^{200}Hg = \delta^{200}Hg - (\delta^{202}Hg \times 0.5024) \quad (4)$$

$$\Delta^{201}Hg = \delta^{201}Hg - (\delta^{202}Hg \times 0.7520) \quad (5)$$

Users can automatically perform this part of the calculation task in the HG MC Auto software interface (Option 2): Parses exported .csv files to extract mean and standard error values, chronologically sorts instrumental logs, and merges datasets to generate comprehensive summary tables ("Before_Fractionation_Calculation.xlsx")

S2. Confirm the 2SD range and segmentation method

Two standard deviations (2SD) are commonly used to represent a 95% confidence interval (assuming the data are normally distributed), meaning that there is a 95% probability that the measured values fall within this range. In mercury isotope analysis, 2SD is typically used to represent: the external reproducibility of isotope ratios (such as δ^{202}Hg), the uncertainty range of standard or unknown samples, and the assessment of long-term instrument stability.

Calculation steps (using δ^{202}Hg as an example)

(1) Calculate the average.

$$x_a = \text{mean}(\delta^{202}Hg_{1\ldots k}) \quad (6)$$

(2) Calculate the sample standard deviation (using n-1 in the denominator).

$$s = \sqrt{[\Sigma(x_i - x_a)^2/(k-1)]} \quad (7)$$

(3) 2SD.

$$2SD = 2 \times s \quad (8)$$

(4) Record this 2SD as the uncertainty of δ^{202}Hg for the entire batch of unknown samples in the results column.

$$\delta^{202}Hg = x_a \pm 2SD(2SD, n = k) \quad (9)$$

For the “3133”, “3177”, and “8610” standard samples, the deviation of each test result from the known value is calculated, followed by the standard deviation of these deviations. If the standard deviation of the deviations falls within an acceptable range (for example, less than the threshold of the mean ± 2SD for all 3133 samples in this section), it is marked as “Normal”; otherwise, it is marked as “Abnormal”.

For the “Sample” section, if the values do not fall within the ranges specified, it is marked as “Potential anomaly detected; please pay attention to the source of the sample”; if they do, it is marked as “Normal”.

The preliminary assessments of the above data are noted in the “Check Status” column.

S3. Exploration of abnormal data

It can be observed that certain instrument parameters—"Extraction [V]", "Focus [V]", "Source Quad1 [V]", "Rot-Quad1 [V]", "Foc-Quad1 [V]", "Rot-Quad2 [V]", "Source Offset [V]", "Matsuda Plate [V]", "Cool Gas [l/min]", "Aux Gas [l/min]", "Sample Gas [l/min]", "Add Gas [l/min]", "Org Gas [l/min]", "Operation Power [W]", "X-Pos [mm]", "Y-Pos [mm]", "Z-Pos [mm]", "Ampl.-Temp [°C]", "Fore Vacuum [mbar]", "High Vacuum [mbar]", "IonGetter-Press [mbar]"—show certain correlations with the internal precision indicators "StdErr(abs)^{202}Hg/^{198}Hg" "StdErr(abs)^{201}Hg/^{198}Hg" "StdErr(abs)^{200}Hg/^{198}Hg" and "StdErr(abs)^{199}Hg/^{198}Hg", which is consistent with earlier studies that used internal precision as an indicator of instrument stability at this stage.

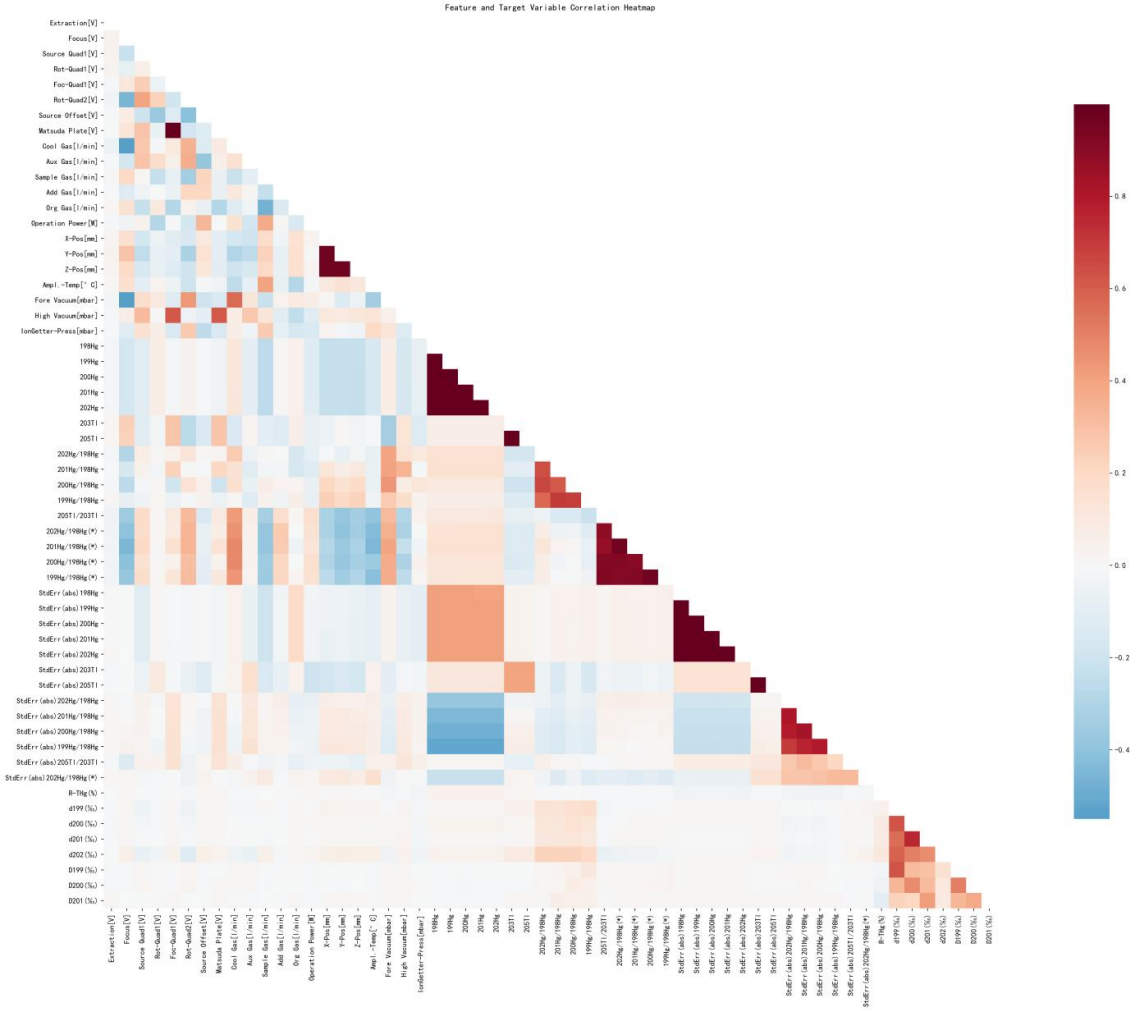

Figure S1 Correlation analysis chart, which presents correlation analyses of instrument parameters, internal precision indicators, isotope signal values, ratios, concentrations, and fractionation values using all the exported data (The correlation coefficients can be found in the attached Correlation matrix.xlsx).

During the process of testers assessing the causes of erroneous data, the concentration metric “RCM, (R-THg(%))” is also considered a potential factor that may affect the data. The correlation diagram (Figure S1) demonstrates that these assessments are reasonably grounded. Likewise, the ratio section is divided into values marked with (*) and those without (*); values with (*) represent pre-correction data, while values without (*) represent post-correction data. The correlation diagram indicates that after correction (current correction methods include power function correction, balance correction, Russell correction, and Baxter correction; in MC-ICP-MS, the Russell correction is typically used[5, 8, 9]), the effects of instrument parameter fluctuations on isotope ratios are substantially reduced.

Concentration Anomaly (R-THg(%)): The accuracy of SSB correction is highly sensitive to concentration matching between the sample and the bracketing standard. Significant deviations can induce non-linear mass bias effects. The threshold of ±10% for R-THg(%) is adopted based on the reference,[10] which demonstrates that Hg isotopic measurements within this range can achieve optimal precision and accuracy. Samples exceeding this threshold are flagged for "It might be an abnormal concentration"

Instrument Instability (StdErr(abs)): The internal precision (StdErr(abs)) of isotope ratios is a direct measure of signal noise during integration. Elevated noise is a hallmark of transient instrumental instability (e.g., plasma fluctuations, intermittent introduction issues). Literature-based thresholds for 2StdErr(abs) (Table 2) were established from the long-term performance data of well-behaved certified reference materials (CRMs; NIST 3133, 3177, and 8610) under optimal conditions. Measurements where 2StdErr(abs) exceeds these CRM-derived limits are flagged for "Potential Instrument Instability."

Integrated Diagnosis and "Other Factors": Critically, samples not explained by these two prevalent issues are categorized as "Other Reasons, Retesting Recommended." This category is intentionally broad and serves as a crucial acknowledgment of the framework's inherent and important limitation: it cannot diagnose failures arising from sample-specific matrix effects (e.g., spectral interferences from concomitant elements, non-spectral matrix-induced sensitivity shifts) because such information is not encoded in the standard isotopic ratio and internal precision outputs of a typical MC-ICP-MS run. Diagnosing matrix effects requires independent elemental characterization data which falls outside the scope of this data-driven diagnostic model.

S4. Introduction to ML model and algorithms

S4.1 Algorithm descriptions

S4.1.1 Logistic Regression (LR)

As a fundamental statistical classification algorithm, LR employs a sigmoid function to model the probability of categorical outcomes.[11] In isotope geochemistry applications, this method provides interpretable coefficients that quantify feature importance, offering insights into the relative contributions of various isotopic ratios to classification decisions.[12]

S4.1.2 Random Forest (RF)

This ensemble method constructs multiple decision trees during training and aggregates their predictions through majority voting.[13, 14] For complex isotopic datasets with non-linear relationships, RF demonstrates robust performance by reducing overfitting while handling high-dimensional feature spaces effectively.[15, 16]

S4.1.3 Support Vector Machine (SVM)

SVM identifies optimal hyperplanes to maximize separation between classes in high-dimensional space.[17] When applied to isotopic discrimination problems, its kernel functions adeptly capture intricate patterns in multi-isotope systems, particularly beneficial for datasets with clear margin separation.[18]

S4.1.4 K-Nearest Neighbors (K-NN)

K-NN Operating on the distance-based similarity principle, K-NN classifies samples according to their proximity to labeled neighbors in feature space.[17] This method proves particularly valuable for isotopic fingerprinting applications where local sample relationships provide critical classification information.[19]

S4.1.5 Decision Tree (DT)

Through recursive binary splitting of feature space, DT create interpretable classification rules.[20] Their transparent structure facilitates geochemical interpretation by explicitly showing decision boundaries based on isotopic threshold values.[17]

S4.1.6 Gradient Boosting (GB)

This sequential ensemble technique builds trees iteratively, with each new model correcting errors of its predecessors.[11] For isotopic classification tasks involving subtle geochemical signatures, gradient boosting excels at capturing complex feature interactions through its stage-wise optimization approach.[21]

S4.1.7 Naive Bayes (NB)

Based on Bayesian probability theory with feature independence assumptions, this probabilistic

classifier efficiently handles high-dimensional isotopic datasets.[19] Its computational efficiency makes it suitable for preliminary screening of large geochemical databases.[22]

S4.1.8 XGBoost (XGB)

As an optimized gradient boosting implementation, XGBoost incorporates regularization techniques to prevent overfitting while supporting parallel processing.[13] In isotope geochemistry applications, it consistently demonstrates superior performance in handling mixed-type isotopic data with missing values.

S4.1.9 Bagging Random Forest (BagRF)

By combining bootstrap aggregation with random feature selection, this ensemble variant enhances prediction stability and accuracy. For isotopic provenance studies requiring robust classification, bagging RF reduces variance while maintaining model interpretability.[23]

S4.2 Sampling methodologies for imbalanced data

S4.2.1 SMOTE (Synthetic Minority Over-sampling Technique)

This approach generates synthetic minority class samples through linear interpolation between existing instances.[24] In geochemical contexts where rare isotopic signatures are underrepresented, SMOTE effectively balances class distributions while preserving the multivariate relationships within isotopic feature space.[11, 25]

S4.2.2 ADASYN (Adaptive Synthetic Sampling)

Building upon SMOTE principles, ADASYN adaptively creates synthetic samples based on learning difficulty, with greater emphasis on minority class examples that are harder to learn.[11, 26] This method proves particularly advantageous for isotopic datasets containing complex, overlapping geochemical end-members.

S4.2.3 SMOTEENN (SMOTE + Edited Nearest Neighbors)

This hybrid technique combines synthetic oversampling with data cleaning using nearest neighbors rule. For isotopic datasets containing noisy measurements or analytical artifacts, SMOTEENN enhances classification performance by both augmenting minority classes and removing ambiguous majority class samples.[11]

S4.2.4 UnderSampling

By randomly reducing majority class instances to match minority class cardinality, this approach directly addresses class imbalance.[11, 24] In isotope geochemistry applications where majority classes dominate databases, strategic under-sampling helps prevent model bias while maintaining essential

geochemical relationships.

S4.2.5 Cost-Sensitive Learning

Instead of resampling training data, cost-sensitive learning incorporates imbalance handling directly into the model's optimization process by assigning higher misclassification costs to minority classes. This approach preserves the original data distribution while strategically weighting learning objectives to prioritize accurate minority class identification.[27] In mercury isotope analysis, where extreme class imbalance characterizes anomaly distributions, cost-sensitive learning provides robust multi-class discrimination without synthetic data generation or information loss from undersampling.

Table S2. Hyperparameter settings for binary classification models

Model	Key hyperparameters	Class balancing	Random state
LR	C=0.1, penalty='l2', solver='liblinear', max_iter=1000	class_weight='balanced'	42
RF	n_estimators=50, max_depth=8, min_samples_split=15, min_samples_leaf=10	class_weight='balanced'	42
SVM	C=0.1, kernel='rbf', gamma='scale'	class_weight='balanced'	42
K-NN	n_neighbors=7, weights='distance'	-	-
DT	max_depth=6, min_samples_split=20, min_samples_leaf=10, max_features=0.8	class_weight='balanced'	42
GB	n_estimators=50, max_depth=4, min_samples_split=20, min_samples_leaf=10, learning_rate=0.05, subsample=0.8	-	42
NB	var_smoothing: Float = 1e-9	-	-
XGB	n_estimators=50, max_depth=4, learning_rate=0.05, reg_alpha=0.1, reg_lambda=0.1	scale_pos_weight=1	42
BagRF	base_estimator=RF(n_estimators=30, max_depth=6), n_estimators=10, max_samples=0.8	-	42

Table S3. Hyperparameter settings for multiclass anomaly diagnostic models

Model	Key hyperparameters	Class balancing	Random state
RF	n_estimators=200, max_depth=10	Cost-Sensitive	42
SVM	kernel='rbf', probability=True	Cost-Sensitive	42
LR	C=0.1, max_iter=1000	Cost-Sensitive	42

S4.3 Enhanced feature description

The current basic features are all set based on the judgment and experience of the experimenters. We also hope to further explore the impact of other factors on anomalous data, model accuracy, and prediction reliability. To this end, we have added some enhanced features (features for which reasonable ranges are currently unknown). This part is provided as an optional feature in the Hg MC Auto software interface (Option 5.2) for users to choose from. In addition to the more commonly used RSD metrics, we incorporated some new features, such as the Data_Quality_Score, by combining basic characteristics, in order to simulate users constructing new features and explore their impact on the model.

$$RSD_{^{xxx}Hg} = \frac{StdErr(abs)^{xxx}Hg/^{198}Hg}{|^{xxx}Hg/^{198}Hg|} \tag{10}$$

RSD features normalize the absolute standard error by the measured isotope ratio, providing concentration-independent precision metrics essential for comparing samples across varying mercury concentrations—a critical consideration in environmental sample analysis.

$$StdErr\_ratio_{xxx/202} = \frac{StdErr(abs)^{xxx}Hg/^{198}Hg}{StdErr(abs)^{202}Hg/^{198}Hg} \tag{11}$$

These ratios quantify precision relationships between different mercury isotopes, potentially revealing mass-dependent analytical behaviors or detector-specific anomalies that affect odd and even mass isotopes differentially.

$$Total\_StdErr = \frac{\sum StdErr(abs)^{xxx}\,Hg/^{198}Hg}{4} \tag{12}$$

The average standard error across all four monitored isotope ratios provides a robust, integrated measure of overall measurement noise, reducing the impact of sporadic single-isotope anomalies.

$$Data\_Quality\_Score = \frac{1}{1 + Total\_StdErr \times R - THg(\%)} \tag{13}$$

This novel composite metric simultaneously considers both precision (through Total_StdErr) and accuracy (through R-THg(%)), recognizing that deficiencies in either dimension substantially compromise analytical quality. Scores approach 1.0 for optimal measurements and decrease asymptotically with deteriorating quality.

$$CV_{^{xxx}Hg} = \frac{StdErr(abs)^{xxx}Hg/^{198}Hg}{Total\_StdErr} \tag{14}$$

CV features quantify the relative contribution of each isotope's precision to the overall measurement uncertainty, aiding in diagnostic assessment of which mass stations may be contributing disproportionately

to noise.

The “Cause of the anomaly” was designated as the target variable. Five-fold cross-validation was employed to assess model stability, with the most robust model selected based on performance metrics such as stability, F1-score, precision, and recall.

S4.4 Train your own ML expert experience model

You only needs to provide the file path of the data set they have labeled with their own defined empirical range in the HG MC Auto software interface (Option 3.2). This data set includes a binary “Check Status” column (all_empirical_model_classified_data.xlsx) and a multiclass “Cause of the anomaly” column (all_abnormal_data_processed.xlsx). Afterwards, You can select the HG MC Auto software interface (Option 4) again to perform predictions using their own trained ML model.

S5. Supplement to the results

S5.1 Binary classification section

The Figure S2 is a supplement to Figure 4, Mainly reflects the model's prediction confidence and accuracy, It can be observed that the top three models generally achieve a prediction accuracy of 0.75 or higher, while their minimum prediction confidence also exceeds 0.5.

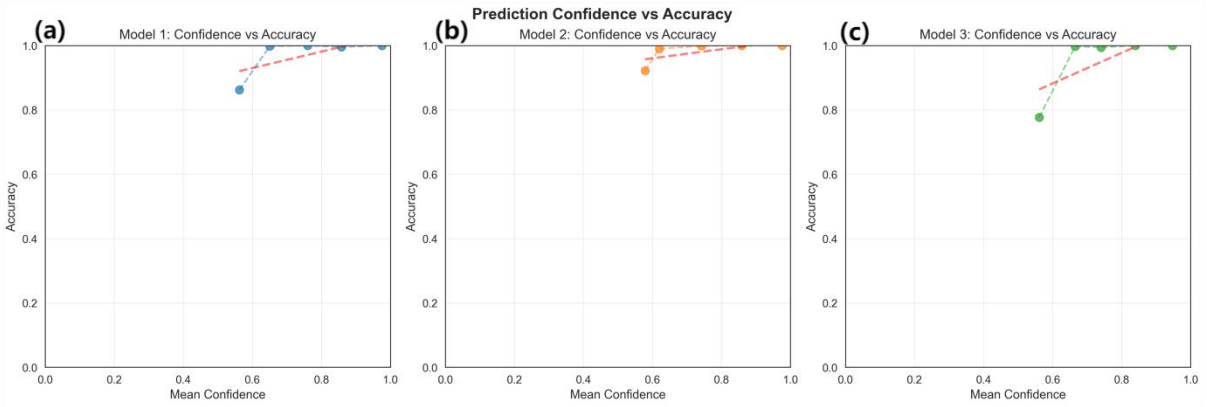

Figure S2 Confidence versus accuracy for (a) RF_S, (b) BagRF_U and (c) XGB_U (N=4674, 0.05-wide bins)

S5.2 Diagnostic multi-class classification of anomalies section

The confusion matrix (Figure S3) analysis revealed particularly strong performance in identifying instrument instability events (recall: 0.9984; 0.9979), which is especially valuable for proactive laboratory maintenance. The model’s ability to distinguish concentration-related anomalies from instrumental issues provides actionable diagnostic information that directly supports troubleshooting and method optimization—a significant advancement over binary normal/abnormal classification. The above image uses basic features, while the image below uses enhanced features.

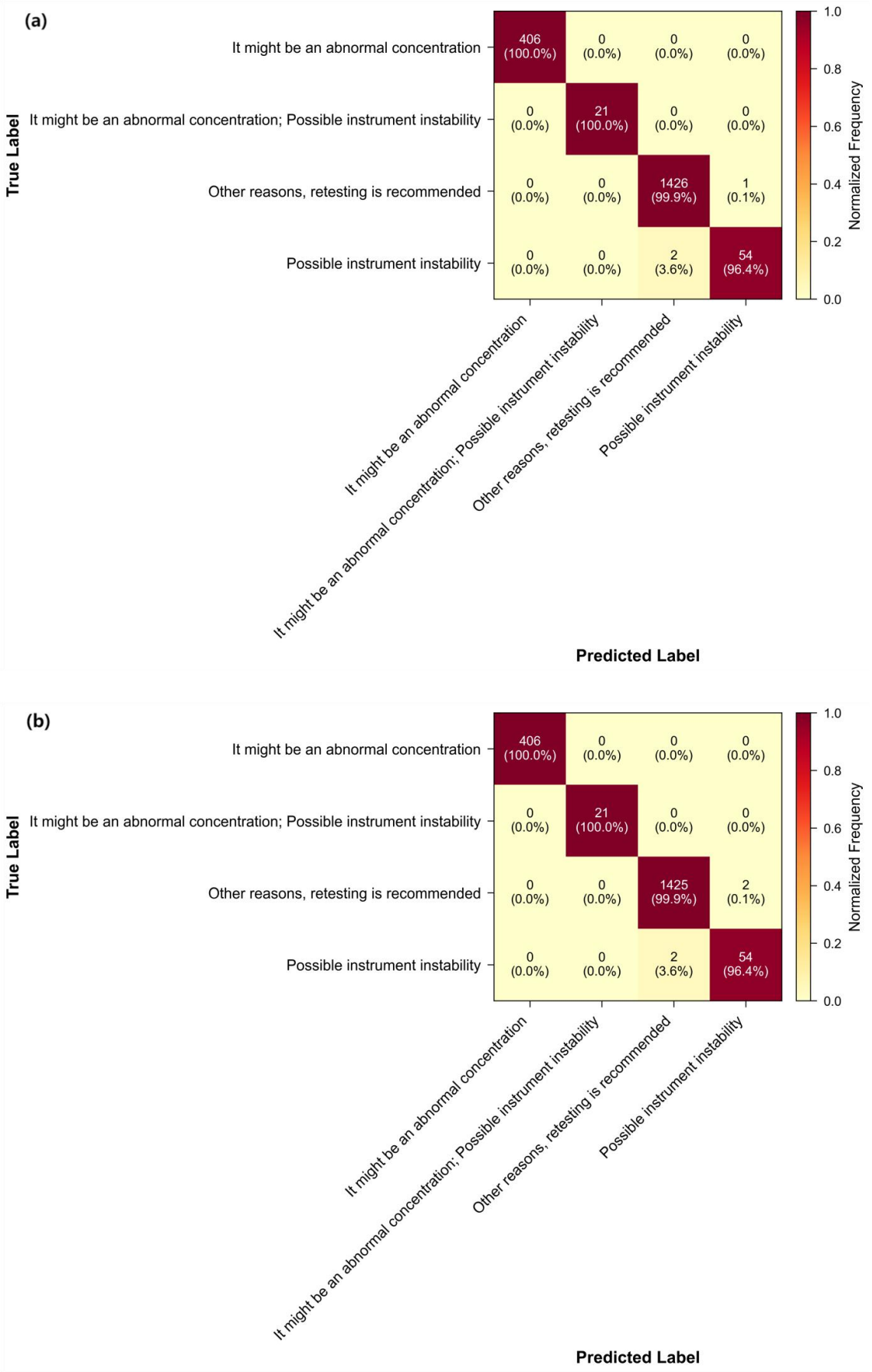

Figure S3. Confusion matrix analysis, (a) basic features, (b) enhance features

REFERENCES

1. W. F. Shi, X. B. Feng, G. Zhang, L. L. Ming, R. S. Yin, Z. Q. Zhao, et al., High-precision measurement of mercury isotope ratios of atmospheric deposition over the past 150 years recorded in a peat core taken from Hongyuan, Sichuan Province, China, *Chin. Sci. Bull.*, 2011, **56**(9), 877-882.
2. J.-H. Yang, J.-H. Wu, M.-F. Zhou, R.-Z. Hu, J.-H. Zhao, A. E. Williams-Jones, et al., Mantle contributions to global tungsten recycling and mineralization, *Commun. Earth Environ.*, 2025, **6**(1), 510.
3. Z. Wang, J. Chen, X. Feng and H. Cai, Progress in the Study of Stable Hg Isotope Geochemistry, *Earth and Environment*, 2012, **40**(4), 599-610.
4. H. Cai and J. Chen, Mass-independent fractionation of even mercury isotopes, *Sci. Bull.*, 2016, **61**(2), 116-124.
5. A. Rua-Ibarz, E. Bolea-Fernandez and F. Vanhaecke, An in-depth evaluation of accuracy and precision in Hg isotopic analysis via pneumatic nebulization and cold vapor generation multi-collector ICP-mass spectrometry, *Anal. Bioanal. Chem.*, 2016, **408**(2), 417-429.
6. D. Y. Jia, K. Luo, Z. D. Xu, X. H. Xu, C. Li, H. M. Wu, et al., Mercury accumulation,distribution,and isotopic composition in tissues of the Collared Scops Owl (Otus lettia), *Acta Geochim.*, 2023, **42**(4), 637-647.
7. G. E. Gehrke, J. D. Blum, D. G. Slotton and B. K. Greenfield, Mercury Isotopes Link Mercury in San Francisco Bay Forage Fish to Surface Sediments, *Environ. Sci. Technol.*, 2011, **45**(4), 1264-1270.
8. L. Yang, S. Y. Tong, L. Zhou, Z. C. Hu, Z. Mester and J. Meija, A critical review on isotopic fractionation correction methods for accurate isotope amount ratio measurements by MC-ICP-MS, *J. Anal. At. Spectrom.*, 2018, **33**(11), 1849-1861.
9. L. Suárez-Criado, S. Queipo-Abad, P. Rodriguez-González and J. I. G. Alonso, Comparison of different mass bias correction procedures for the measurement of mercury species-specific isotopic composition by gas chromatography coupled to multicollector ICP-MS, *J. Anal. At. Spectrom.*, 2024, **39**(2), 508-517.
10. R. S. Yin, D. P. Krabbenhoft, B. A. Bergquist, W. Zheng, R. F. Lepak and J. P. Hurley, Effects of mercury and thallium concentrations on high precision determination of mercury isotopic composition by Neptune Plus multiple collector inductively coupled plasma mass spectrometry, *J. Anal. At. Spectrom.*, 2016, **31**(10), 2060-2068.
11. T. C. C. Lui, D. D. Gregory, M. Anderson, W. S. Lee and S. A. Cowling, Applying machine learning methods to predict geology using soil sample geochemistry, *Appl. Comput. Geosci.*, 2022, **16**, 100094.
12. G. P. Wilson, On the application of contemporary bulk sediment organic carbon isotope and geochemical datasets for Holocene sea-level reconstruction in NW Europe, *Geochim. Cosmochim. Acta*, 2017, **214**, 191-208.
13. J. N. Yin and N. Li, Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping, *Ore Geol. Rev.*, 2022, **145**, 104916.
14. L. Breiman, Random forests, *Machine Learning*, 2001, **45**(1), 5-32.
15. M.-Y. Fan, Y. Hong, Y.-L. Zhang, T. Sha, Y.-C. Lin, F. Cao, et al., Increasing nonfossil fuel contributions to atmospheric nitrate in urban china from observation to prediction, *Environ. Sci. Technol.*, 2023, **57**(46), 18172-18182.
16. X. Qin, Q. Guo, P. Martens and T. Krafft, Mercury stable isotopes revealing the atmospheric mercury circulation: a review of particulate bound mercury in China, *Earth Sci. Rev.*, 2024, **250**, 104681.
17. G. Sun, Q. Zeng and J.-X. Zhou, Machine learning coupled with mineral geochemistry reveals the origin of ore deposits, *Ore Geol. Rev.*, 2022, **142**, 104753.
18. M. Petrelli and D. Perugini, Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data, *Contrib. Mineral. Petrol.*, 2016, **171**(10), 81.

19. M. C. Figueroa, D. D. Gregory, K. H. Williford, D. J. Fike and T. W. Lyons, A Machine-Learning Approach to Biosignature Exploration on Early Earth and Mars Using Sulfur Isotope and Trace Element Data in Pyrite, *Astrobiology*, 2024, **24**(11), 1110-1127.
20. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and regression trees (CART)*, Wadsworth, 1st edn., 1984.
21. M. A. Gul, H. S. Zhang, Y. G. Li, X. Y. Yang, C. Sun, X. J. Zhao, et al., Machine learning-driven classification of Pb-Zn ore deposits using pyrite trace elements and isotopic signatures: A case study of the Gunga deposit, *J. Geochem. Explor.*, 2025, **272**, 107693.
22. Y. D. Chen, Z. K. Liu, R. C. Wang, B. Yang and X. C. Mao, New insights into the metallogenic genesis of the Xiadian Au deposit, Jiaodong Peninsula, Eastern China: Constraints from integrated rutile in-situ geochemical analysis and machine learning discrimination, *Ore Geol. Rev.*, 2024, **171**, 106184.
23. J. Xia, P. Ghamisi, N. Yokoya and A. Iwasaki, Random Forest Ensembles and Extended Multiextinction Profiles for Hyperspectral Image Classification, *IEEE Trans. Geosci. Remote Sens.*, 2018, **56**(1), 202-216.
24. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, 2002, **16**, 321-357.
25. S. C. Xue, Y. Y. Niu, Z. S. Yao, L. Y. Wang, X. H. Zhang and Q. F. Wang, Predicting olivine formation environments using machine learning and implications for magmatic sulfide prospecting, *Am. Mineral.*, 2024, **109**(3), 510-520.
26. H. B. He, Y. Bai, E. A. Garcia, S. T. Li and Ieee, ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, International Joint Conference on Neural Networks, Hong Kong, PEOPLES R CHINA, 2008.
27. J. J. Li, Y. Diao, R. Song, B. B. Xi, Y. S. Li and Q. Du, Class-Specific Autoaugment Architecture Based on Schmidt Mathematical Theory for Imbalanced Hyperspectral Classification, *IEEE Trans. Geosci. Remote Sens.*, 2023, **61**, 15, 5525315.