Supplementary Information

Machine Learning-Driven Multidimensional Tea Profiling from a Single SERS Spectrum: Toward Practical Application

Jincheng Ni[1], Yanyan Lu[2,3], Xuewen Chen[2], Haoming Bao[2*], Yu Qiao[2,3], Shuai Zhang[2,3], Daoshuan Ding[2], Jiamei Jin [2,3], Huaiyuan Zhao[2], Qian Zhao[2], Jinhua Wang[1,4,5*], Hongwen Zhang[2,3*], Weiping Cai[2]

[1]Anhui Provincial Key Laboratory of Environmental Pollution Control and Resource Reuse, Anhui Jianzhu University, Hefei, Anhui, 230601, PR China
[2]Key Lab of Materials Physics, Anhui Key Lab of Nanomaterials and Nanotechnology, Institute of Solid State Physics, HFIPS, Chinese Academy of Sciences, Hefei 230031, P.R. China
[3]University of Science and Technology of China, Hefei 230026, P.R. China
[4]Pollution Control and Resource Utilization in Industrial Parks Joint Laboratory of Anhui Province, Anhui Jianzhu University, Hefei, Anhui, 230601, PR China
[5]Anhui Research Academy of Ecological Civilization, Anhui Jianzhu University, Hefei, Anhui, 230601, PR China

* To whom all correspondence should be addressed.

Email: baohm@issp.ac.cn, huawangjin@163.com, hwzhang@issp.ac.cn

Table of Contents

S1. FDTD Simulation of Electromagnetic Field Distribution

FDTD simulations were conducted to investigate the electromagnetic field distributions of the gold NP system, including both an isolated NP and a nanoparticle film (NP film). The optical properties of gold were defined using experimental dielectric constants from Johnson and Christy, and the surrounding medium was set to air (n = 1.0). Two orthogonal linearly polarized plane wave at 785 nm along with Z axis was used as the excitation source, and perfectly matched layer (PML) boundaries were applied in all directions to eliminate spurious reflections. The mesh size was refined to 2 nm around the nanoparticle region to accurately capture the localized surface plasmon resonance (LSPR) effects. The spatial distributions of the near-field intensity ($|E/E_0|$) were extracted to visualize the electromagnetic "hot spots" and to compare the field enhancement between the NP film and the isolated NP configurations (Figure S2).

S2. SERS EF Calculation

The SERS EF was calculated using 4-NTP as the probe molecule. The EF was determined according to the following equation due to the identical measurement conditions:

$$EF = \frac{\left(\frac{I_{\mathrm{SERS}}}{N_{\mathrm{SERS}}}\right)}{\left(\frac{I_{\mathrm{Raman}}}{N_{\mathrm{Raman}}}\right)} = \frac{\left(\frac{I_{\mathrm{SERS}}}{C_{\mathrm{SERS}}}\right)}{\left(\frac{I_{\mathrm{Raman}}}{C_{\mathrm{Raman}}}\right)}$$

where $I_{\mathrm{SERS}}$and I_{Raman} are the Raman intensities of the characteristic 4-NTP band (typically at 1340 cm^{-1}) obtained from the SERS substrate and the normal Raman measurement, respectively. $N_{\mathrm{SERS}}$and N_{Raman}represent the estimated numbers of 4-NTP molecules contributing to the SERS and normal Raman signals. In practice, C_{SERS}corresponds to the 4-NTP concentration used on the gold nanoparticle substrate (10^{-6} M), while C_{Raman}represents the concentration of 4-NTP solution used for normal Raman measurement (0.1 M).

S3. Prediction Accuracy Calculation

1. Overall Definition

To comprehensively evaluate the performance of the model across multiple prediction tasks, a **Sample–Dimension Averaged Accuracy** was adopted.

Let:

- N: total number of samples,
- M: number of prediction dimensions (e.g., *Types*, *Grade*, *Quality*, *GLY*, *DDVP*, *TMTD*),

- a_{ij}: prediction accuracy of the i^{th}sample on the j^{th}dimension.

Then, the overall prediction accuracy is defined as:

$$\text{Accuracy}_{overall} = \frac{1}{N \times M}\sum_{i=1}^{N}\sum_{j=1}^{M}a_{ij}$$

2. Definition of Per-Dimension Accuracy

(1) Categorical Variables (Types, Grade)

For categorical attributes, the accuracy is defined as:

$$a_{ij} = \begin{cases} 1, & \text{if } y^{pred}_{ij} = y^{real}_{ij} \\ 0, & \text{otherwise} \end{cases}$$

(2) Numerical Variables (Quality, GLY, DDVP, TMTD)

For numerical attributes, an improved relative accuracy formulation was used to ensure stability when the real value is not zero. The accuracy is defined as:

$$a_{ij} = 1 - \frac{|y^{pred}_{ij} - y^{real}_{ij}|}{y^{real}_{ij}}$$

When the real value is zero, it is given by:

$$a_{ij} = 1 - \frac{y^{pred}_{ij}}{y^{real}_{(i+1)j} - y^{real}_{ij}}$$

where:

- y^{real}_{ij} is the true (real) value,

- $y^{pred}_{(i+1)j}$ is the predicted value,

These expressions also show how closely the measured value fits the true value.

3. Handling Missing Values

If either the real or predicted value for a given dimension is missing (represented as "–"), that dimension is excluded from the accuracy computation:

a_{ij} = NaN⇒excluded from aggregation.

4. Practical Computation Procedure

1. For each sample i: compute accuracy a_{ij}across all available dimensions j.

2. Aggregate all valid a_{ij}values:

$$\text{Accuracy}_{overall} = \frac{\sum a_{ij}}{\text{number of valid (i, j) pairs}}$$

Convert to percentage form (multiply by 100%) for easier interpretation.

According to Table S1, N=23, M=6, $\text{Accuracy}_{overall}$ is calculated as 98.2%.

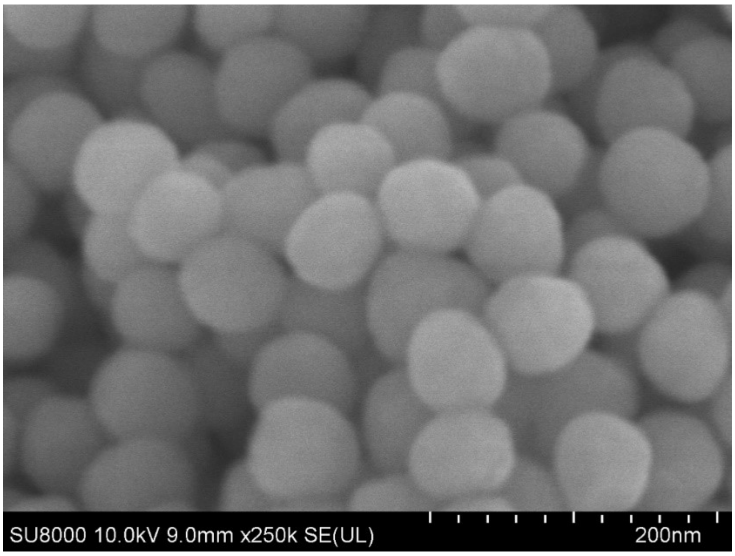


Figure S1. SEM image of the used gold NPs.

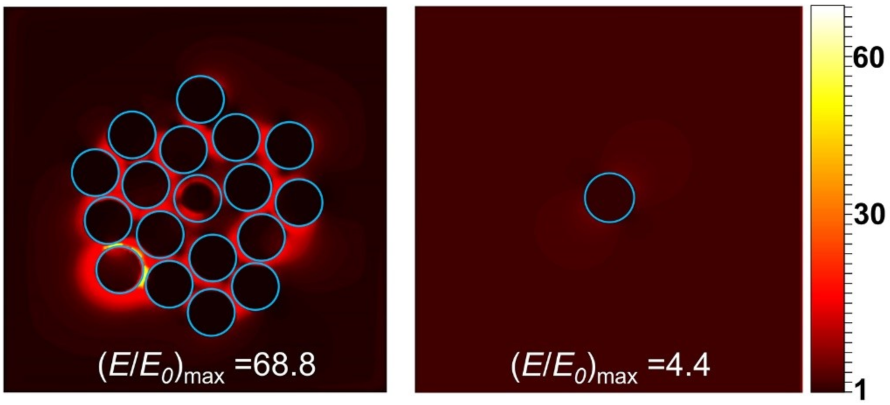


Figure S2. FDTD simulation results of electromagnetic field distribution of gold NP system (NP film) and an isolated NP.

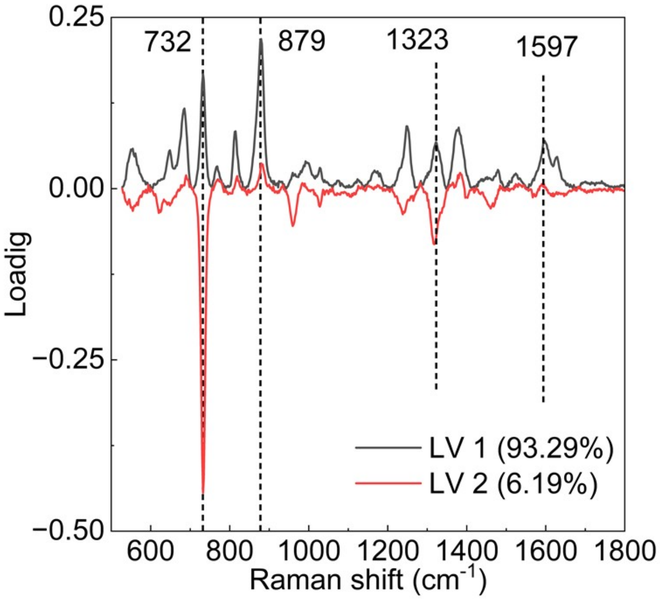


Figure S3. Loading plots highlighting key discriminatory spectral regions at 600-900 cm^{-1} and 1200-1600 cm^{-1}, corresponding to major biomolecular vibrations.

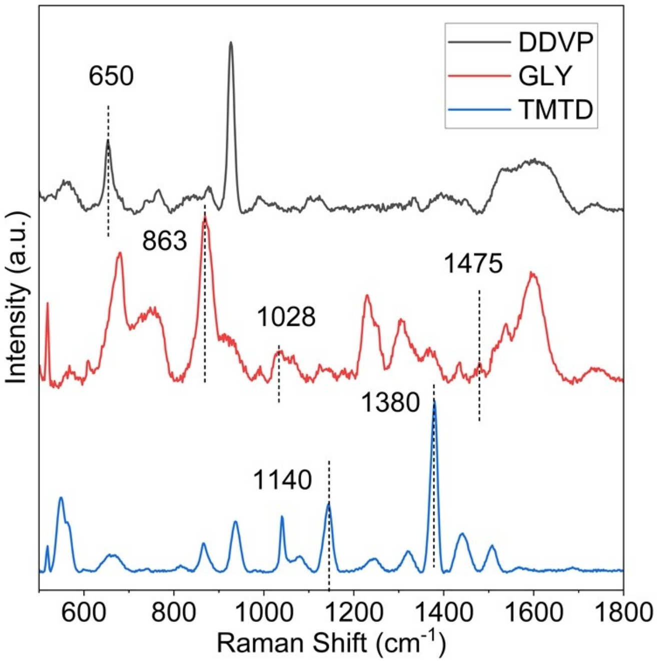


Figure S4.SERS spectra of three pesticides (DDVP, GLY and TMTD).

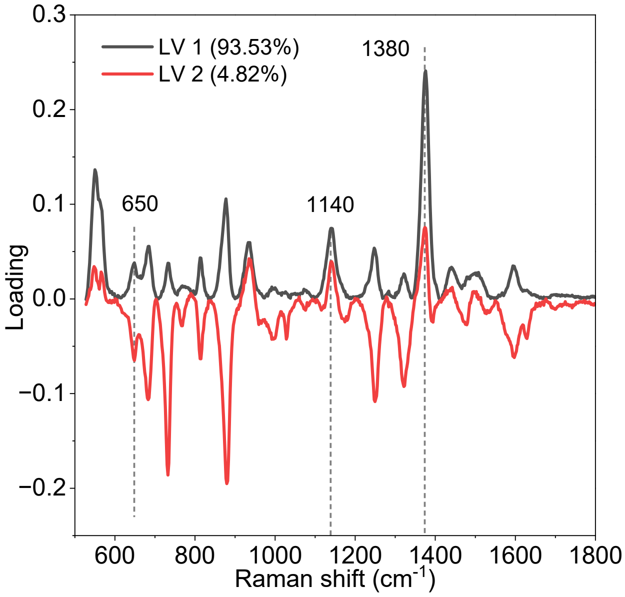


Figure S5. Loading plots highlighting key discriminatory spectral regions at 600-1000 cm^{-1} and 1100-1600 cm^{-1}, corresponding to major biomolecular vibrations.

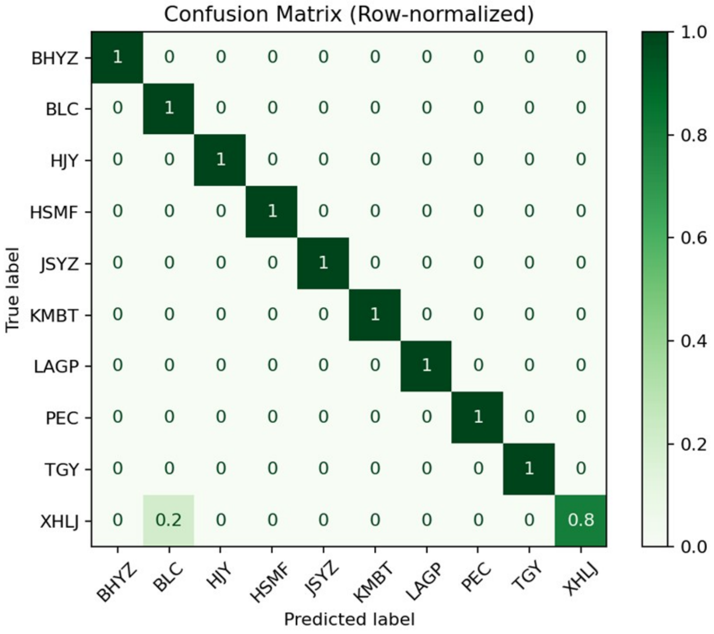


Figure S6. Training set confusion matrix of tea category classification.

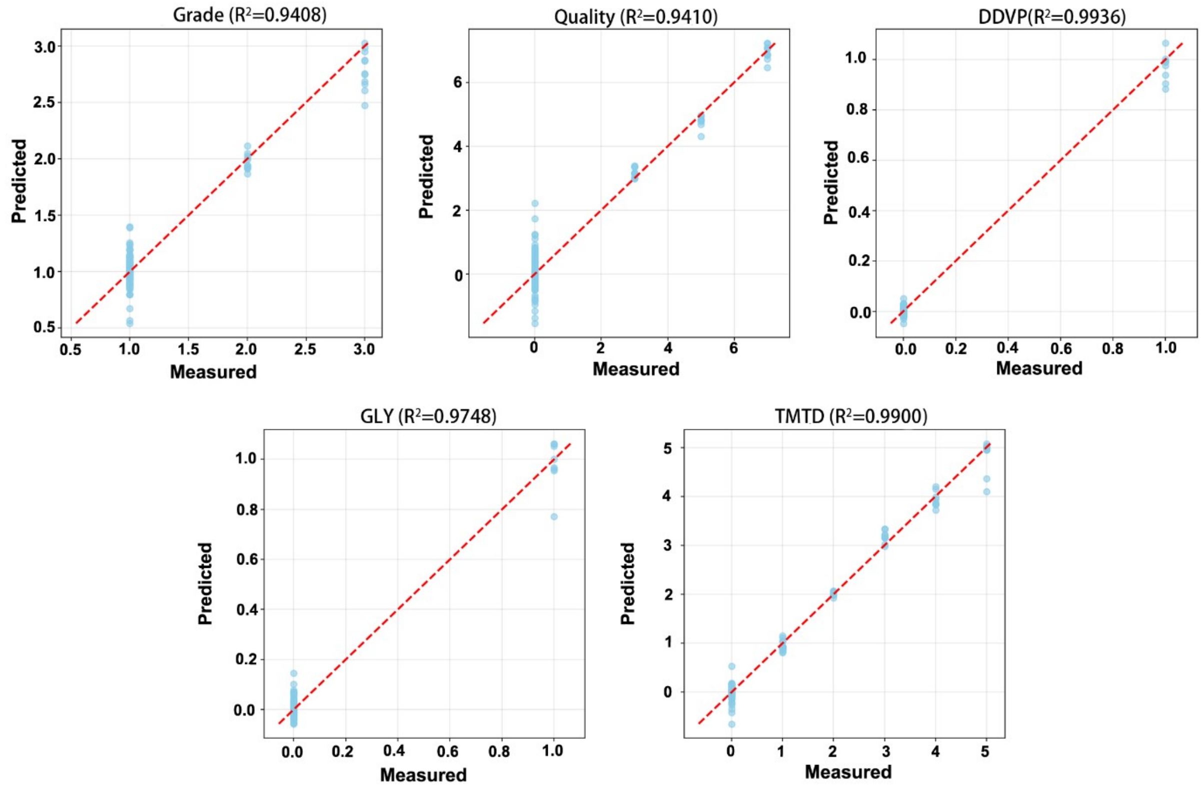
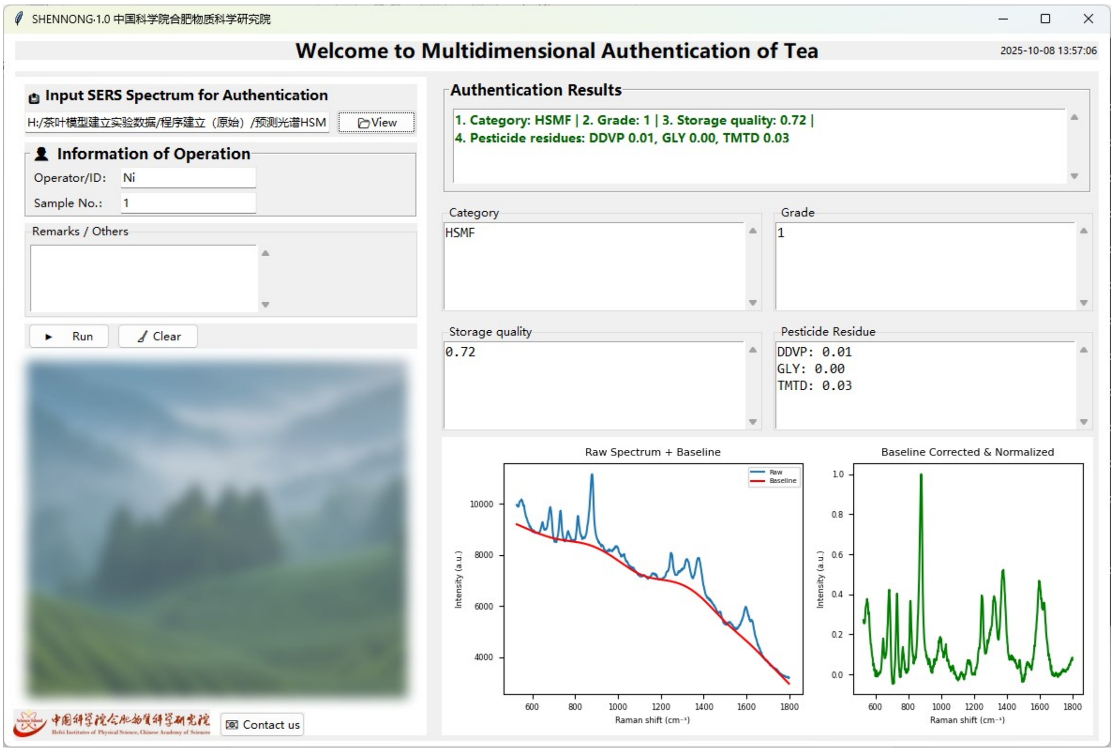


Figure S7. SVM regression for quantitative prediction of tea properties (grade, quality, and residues of DDVP, GLY, and TMTD). Scatter plots of predicted versus measured values exhibit strong correlations across all endpoints.



SHENNONG-1.0 中国科学院合肥物质科学研究院

Welcome to Multidimensional Authentication of Tea

2025-10-08 13:57:06

Input SERS Spectrum for Authentication

H:/茶叶模型建立实验数据/程序建立（原始）/预测光谱HSM

Information of Operation

Operator/ID: Ni

Sample No.: 1

Remarks / Others

Authentication Results

1. Category: HSMF | 2. Grade: 1 | 3. Storage quality: 0.72 |
4. Pesticide residues: DDVP 0.01, GLY 0.00, TMTD 0.03

Category: HSMF

Grade: 1

Storage quality: 0.72

Pesticide Residue:

```
DDVP: 0.01
GLY: 0.00
TMTD: 0.03
```

Figure S8. User interface of SHENNONG 1.0 application.

Table S1. Robustness of model performance across 10 cross-validation splits (C (5,2) =10).

Model	**Mean overall accuracy**	**SD (across 10 splits)**	**Best**	**Worst**
Grade classification	0.99833	0.00500	1.00000	0.98333
Tea variety classification	0.99050	0.01106	1.00000	0.97500
Pesticide regression	0.98875	0.01420	1.00000	0.96250
Storage quality	0.99200	0.00748	1.00000	0.98000