Supplementary Information (SI) for Analyst.
This journal is © The Royal Society of Chemistry 2025

Supporting information

Boosting Living Bacillus Spore Identification: Kolmogorov-Arnold Network-Guided Convolutional Neural Network Combined with Laser Tweezers Raman Spectroscopy

Yifan Sun,[a,1] Xiao Peng, [a,1] Fusheng Du,[b] Lin He,[b] Yuan Lu, [c,*] Yufeng Yuan [b,*] and Junle Qu [a]

[a] State Key Laboratory of Radio Frequency Heterogeneous Integration (Shenzhen University), College of Physics and Optoelectronic Engineering, Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, Shenzhen University, Shenzhen, Guangdong 518060, China

[b] School of Electronic Engineering and Intelligentization, Dongguan University of Technology, Dongguan, Guangdong 523808, China

[c] The Sixth People's Hospital of Shenzhen University, Shenzhen University, Shenzhen, Guangdong, 518060, China

[1] Yifan Sun and Xiao Peng contributed equally to this work.

*Email: yufengyuan@dgut.edu.cn, chfsums@163.com

Supporting Figures

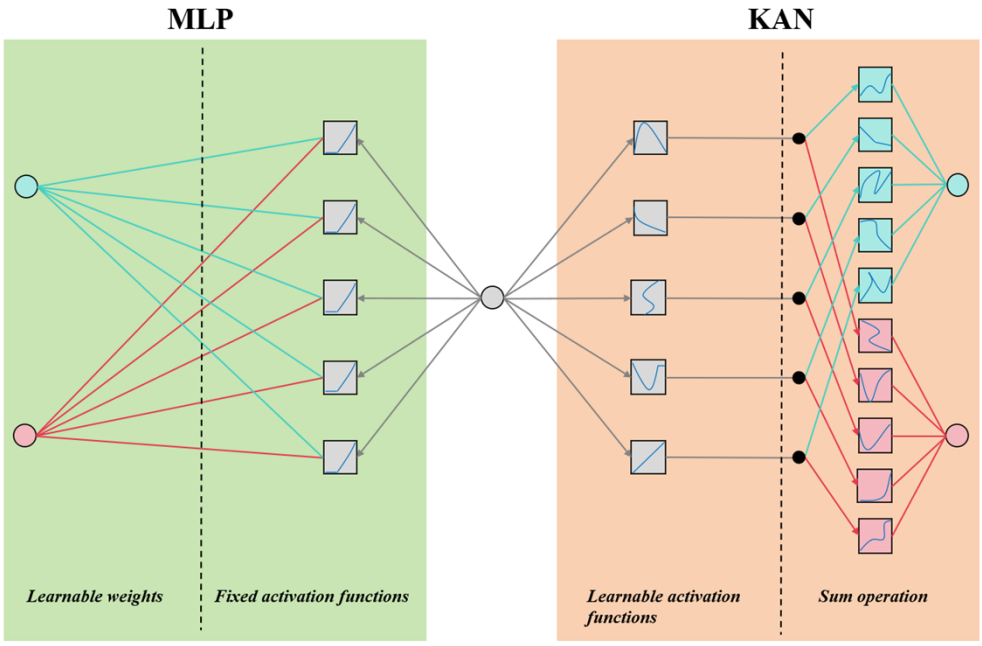

Figure S1. Comparison of MLP with KAN in algorithm.

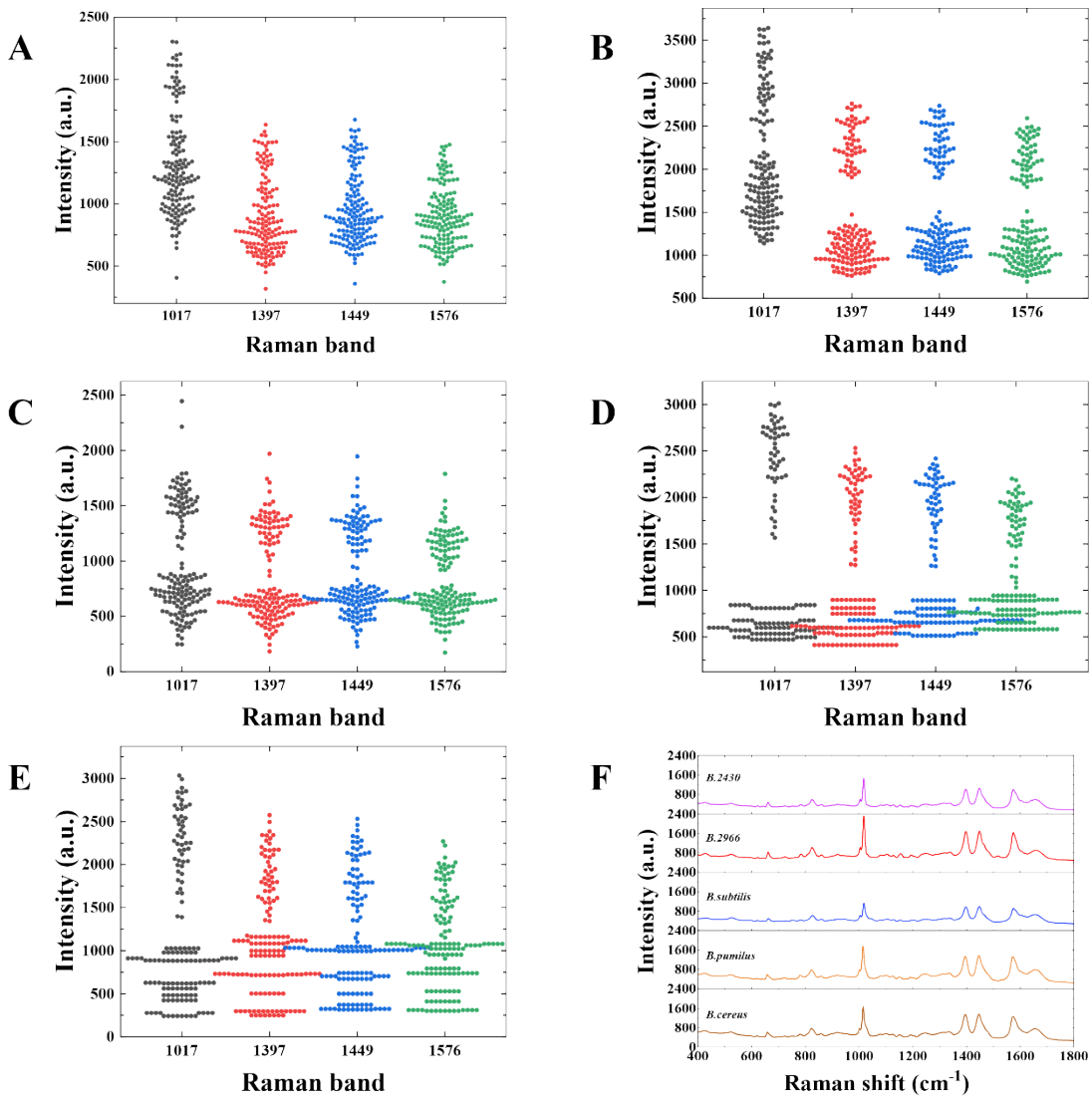

Figure S2. Beeswarm diagram of signal intensity based on four characteristic Raman bands (1017, 1397, 1449, and 1576 cm^{-1}) from five *Bacillus* species **(**A: *B. marisflavi* (MCCC1K02430); B: *B. aryabhata* (MCCC1K02966); C: *B. subtilis* (CICC63501); D: *B. pumilus* (CICC22276); E: *B. cereus* (CICC22369). (F) Averaged Raman spectra of captured spores originated from five *Bacillus* species.

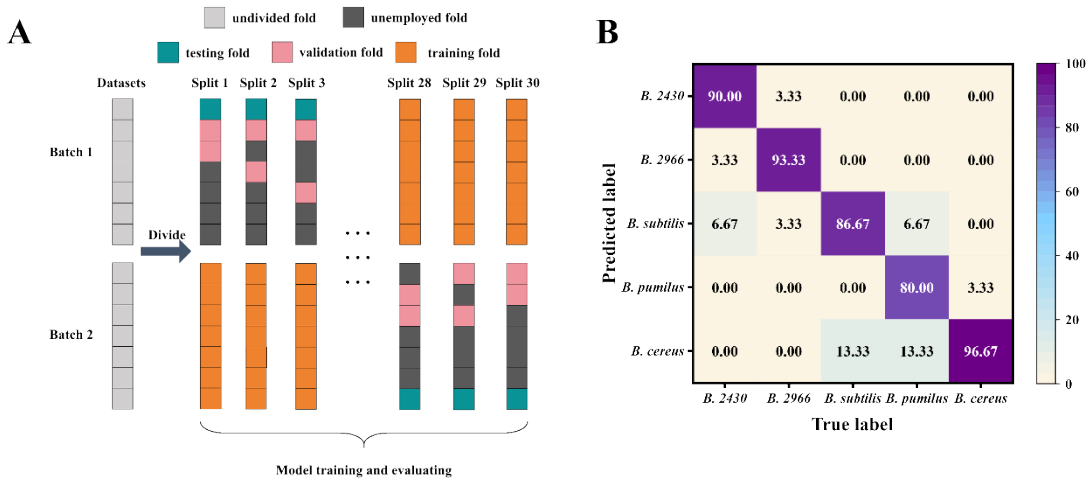

Figure S3. (A) Algorithm illustration of grouped cross-validation mechanism. In brief, the single-cell Raman spectra datasets were initially divided into 14 parts (7 parts for each batch), and each fold contain the same proportion spectra of five *Bacillus* species. Then, one entire batch were employed as training datasets. For the other batch, the five parts were selected two parts to act as validation datasets, and the left one was employed to test the performance of CNN model. When one batch were employed as training datasets, in the other batch, one confirmed fold was constantly employed to test the performance and two folds selected form rest of six were be used as validation data, in order to systematically quantify the performance. Consequently, it was divided into 15 splits, and each one stranded for a possibility of selecting two from five in two batch without repetition. (B) Confusion matrix of prediction accuracy obtained by optimal single CNN model.

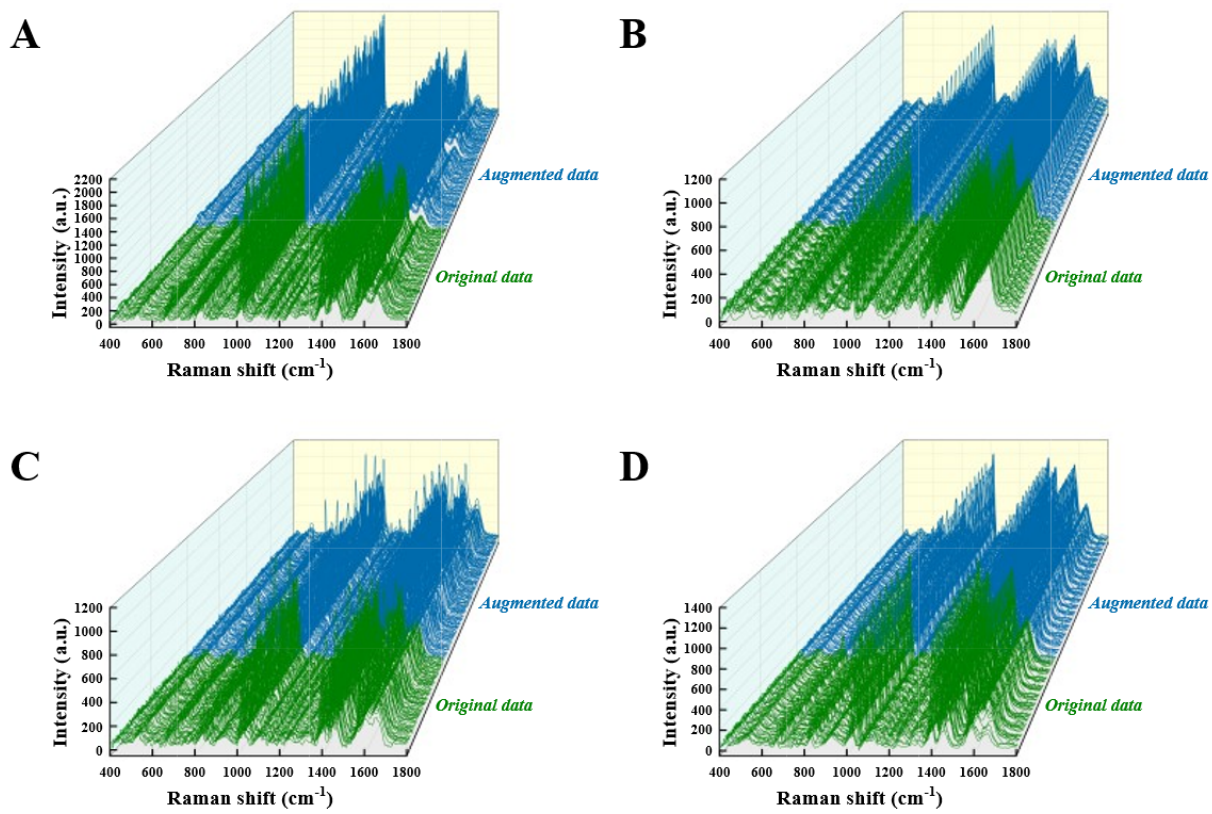

Figure S4. With a data augmentation by 300, obtained distribution of spectra datasets including augmented and original data based on four *Bacillus* spore species (A) *Bacillus aryabhata* (MCCC1K02966), (B) *Bacillus subtilis* (CICC63501), (C) *Bacillus pumilus* (CICC22276), and (D) *Bacillus cereus* (CICC22369), respectively.

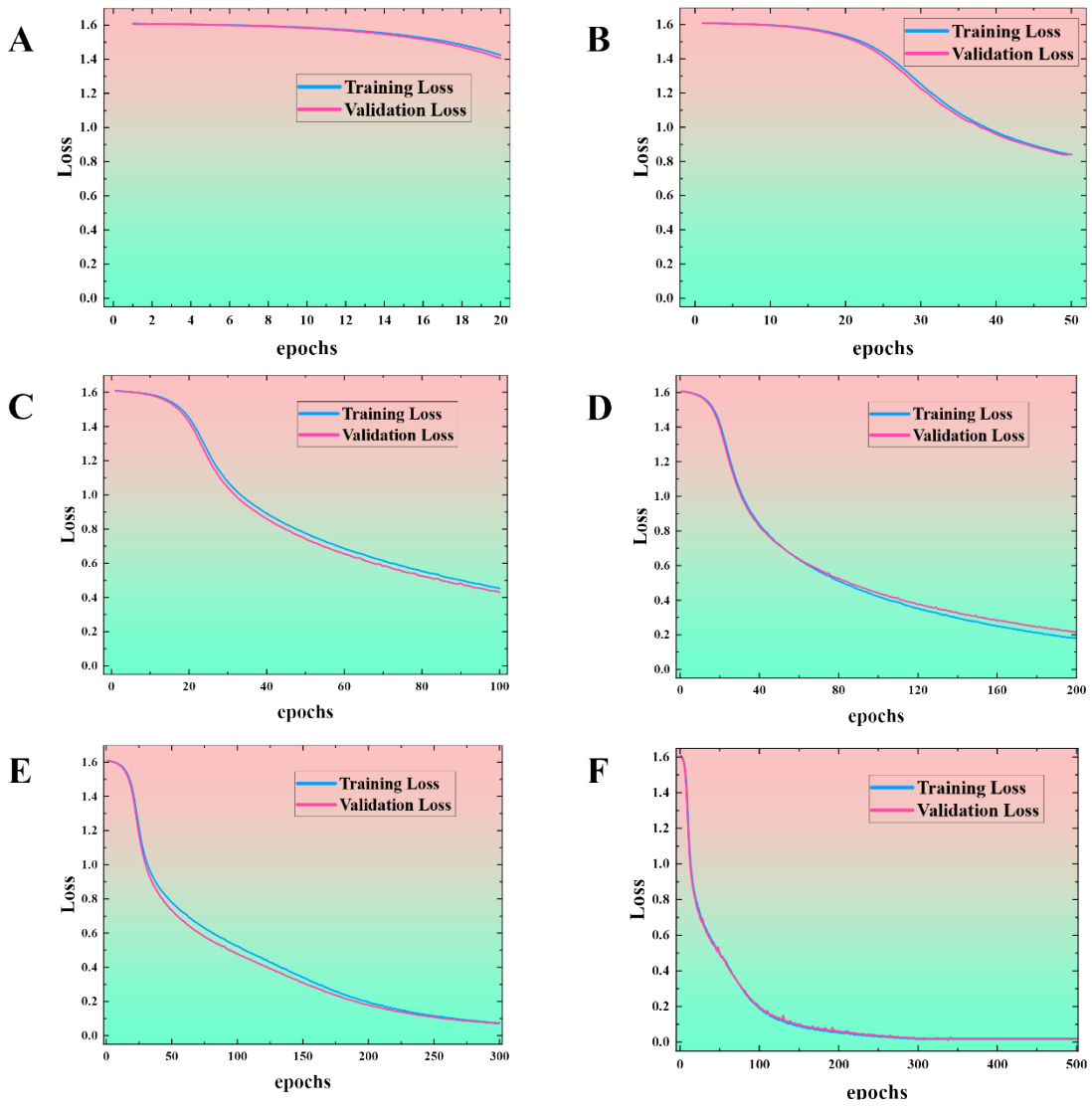

Figure S5. Monitoring of two loss function curves from the training (blue curve) and validation datasets (pink curve) by running various epochs ((A) 20 epochs, (B) 50 epochs, (C) 100 epochs, (D) 200 epochs, (E) 300 epochs, (F) 500 epochs).

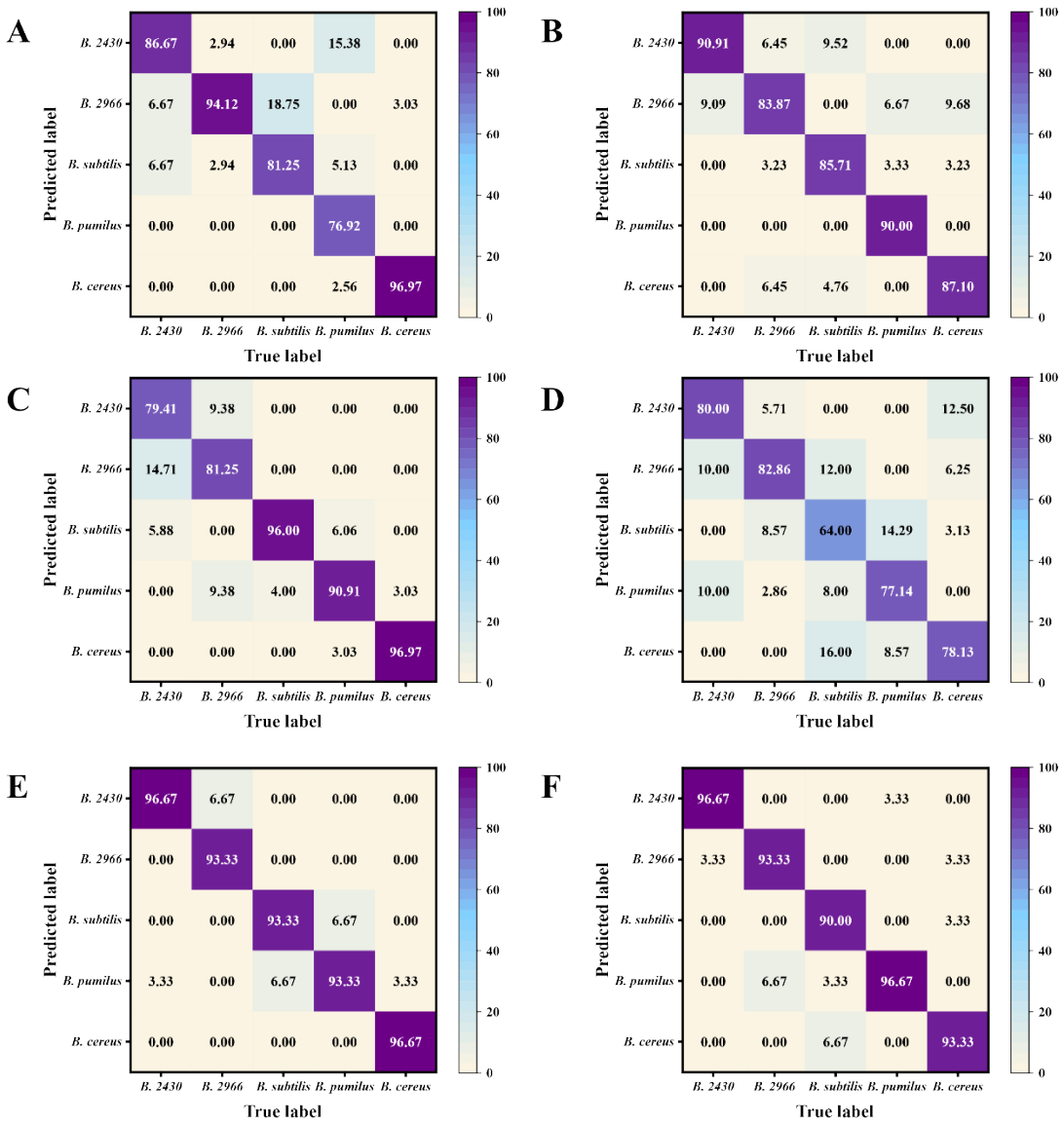

Figure S6. Confusion matrix of prediction accuracy based on various classifiers such as (A) Random Forest, (B) SVM, (C) Xgboost, (D) KNN, (E) ResNet, and (F) Transformer, respectively.

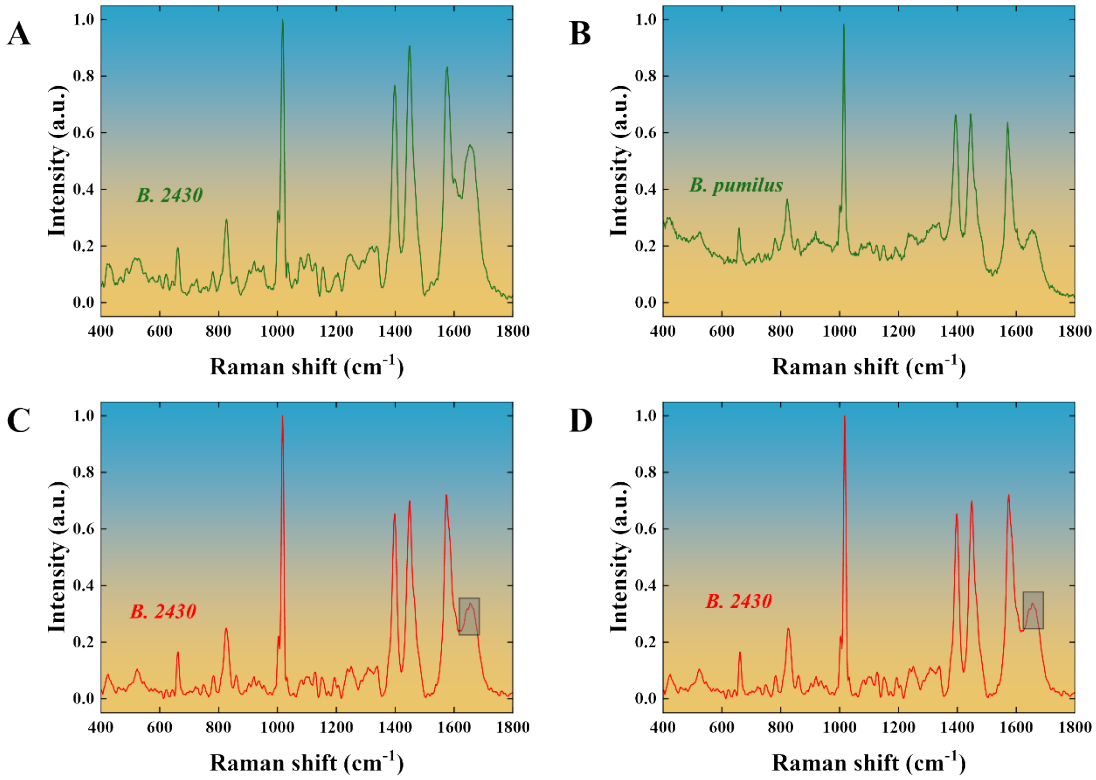

Figure S7. Examples of misjudged Raman spectra (*B. 2430* was misjudged as *B. pumilus*): (A) Right judgment of *B. 2430*, (B) Right judgment of *B. pumilus*, (C) Wrong judgment of *B. 2430*, (D) Wrong judgment of simulated *B. 2430*. Noting that, the Raman spectra (D) was the simulated spectra obtained from the Raman spectra (C), and the Raman bands at 1657 cm^{-1} was highlighted.

Supporting Tables

Table S1. Obtained Raman spectra datasets via various data augmentation

Datasets	Original datasets	Augmented by 200	Augmented by 300	Augmented by 400	Augmented by 500
Training dataset	503	700	1050	1400	1750
Validating dataset	180	200	300	400	500
Testing dataset	70	100	150	200	250
Total number	753	1000	1500	2000	2500

Table S2. Prediction accuracy and classification contribution based on blocking Raman bands

Occluded Raman band (cm^{-1})	Prediction accuracy	Classification contribution
400-648	100%	0%
687-797	99.83%	0.17%
849-982	99.70%	0.30%
1716-1800	99.78%	0.22%

Table S3. Comparison of prediction accuracy using various noise levels of Gaussian noise

Noise level	Prediction accuracy
5%	96.00±3.46%
10%	97.80±1.79%
15%	96.67±2.33%
20%	94.87±4.89%

Table S4. Comparison of prediction accuracy using various augmentation strategies

Data volume	Gaussian noise	baseline variations	Poisson noise	GAN spectra
With data augmentation by 200	93.70±3.94%	94.00±3.32%	93.18±4.32%	93.50±3.04%
With data augmentation by 300	97.80±1.79%	97.40±1.56%	97.67±1.72%	97.47±2.76%
With data augmentation by 400	97.91±1.92%	97.55±1.77%	97.85±1.70%	97.55±1.21%
With data augmentation by 500	98.15±2.21%	98.00±1.72%	98.08±1.46%	98.12±1.90%

Table S5. Comparison of prediction accuracy using various spline order in KAN

Spline **order**	Prediction accuracy
5	95.13±2.62%
10	97.80 ± 1.79%
15	94.80±1.29%
20	93.04±2.29%

Table S6. Prediction accuracy of KAN-guided CNN platform through running 10-time independent operation at 300 epochs

Time	Accuracy
1	96.67%
2	99.33%
3	94.00%
4	100.00%
5	98.67%
6	97.33%
7	98.00%
8	96.00%
9	98.00%
10	100.00%
Average	97.80±1.79%

Table S7. Performance reporting on KAN-guided CNN platform via 300 epochs

Bacillus class	Precision	95% CI on Precision	Recall	95% CI on Recall	F1 scores
B. 2430	90.3%	(75.1%, 96.6%)	93.3%	(78.7%, 98.2%)	91.8%
B. 2966	100.0%	(87.1%, 100.0%)	86.7%	(70.3%, 94.7%)	92.9%
B. subtilis	93.8%	(79.9%, 98.3%)	100.0%	(88.7%, 100.0%)	96.8%
B. pumilus	87.5%	(71.9%, 95.0%)	93.3%	(78.7%, 98.2%)	90.3%
B. cereus	96.6%	(83.3%, 99.4%)	93.3%	(78.7%, 98.2%)	94.9%

Table S8. For an independent validation spectra dataset, obtained prediction accuracy of KAN-guided CNN platform through running 10-time independent operation.

Time	Accuracy
1	95.5
2	96.5
3	97.0
4	95.5
5	96.5
6	95.5
7	95

8	96.5
9	95.5
10	96.5
Average	96.0±0.63%