Received: 29 December 2025 | Revised: 20 February 2025 | Accepted: 28 February 2026 | Published online: 13 March 2026

Medinformatics
2026, Vol. 00(00) 1–16
DOI: 10.47852/bonviewMEDIN62028939

RESEARCH ARTICLE

Machine Learning-Enhanced SERS on Silicon-Gold Sensors for Ultra-High-Throughput Pathogen Detection

Bryan Guilcapi[1,*], Alessia Milano[1], Amalia D' Avino[1, 2], Domenico Sagnelli[1], Massimo Rippa[1], Valentina Marchesano[1], Ivan Salvatore Perrotta[3], Rosa Luisa Ambrosio[4], Giovanna Fusco[5], Maurizio Brigotti[1, 6], Stefano Morabito[7], and Lucia Petti[1,*]

[1] *Institute of Applied Sciences and Intelligent Systems, CNR, Italy*

[2] *Department of Engineering, University of Naples Parthenope, Italy*

[3] *OPUS automazione S.p.A., Italy*

[4] *Department of Veterinary Medicine and Animal Production, University of Naples Federico II, Italy*

[5] *Department of Animal Health Coordination, Istituto Zooprofilattico Sperimentale del Mezzogiorno, Italy*

[6] *Department of Medical and Surgical Sciences, University of Bologna, Italy*

[7] *Department of Food Safety, Istituto Superiore di Sanità, Italy*

Abstract: Surface-enhanced Raman spectroscopy (SERS) combined with machine learning (ML) has emerged as a powerful strategy for intelligent and automated biosensing. In this study, we present an in-house fabricated and data-driven SERS platform for automated discrimination of multiple experimental states within a nanostructured biosensor system. The sensing architecture was developed using reproducible gold nanoparticle-coated silicon substrates (Au NPs), synthesized and assembled in-house, followed by functionalization with a Raman reporter molecule (4-mercaptobenzoic acid, 4-MBA) or three biologically relevant targets: *Escherichia coli*, OC43 coronavirus, and Shiga toxin 2. The classification task was formulated as a six-class problem corresponding to distinct experimental sensor states: bare silicon substrate, Au NP-coated plasmonic substrate, 4-MBA-plasmonic functionalized surface, and plasmonic substrates functionalized with each biological target. Raman spectral data were processed through an automated analytical pipeline and evaluated using principal component analysis, linear discriminant analysis, support vector machines, random forest, and a one-dimensional convolutional neural network (1D-CNN). Among the evaluated models, the 1D-CNN achieved superior performance, providing the highest classification accuracy and robust discrimination across all six experimental classes. The results demonstrate that deep learning applied directly to normalized spectral vectors enhances feature extraction and class separability compared to conventional approaches. This work highlights the potential of integrating in-house engineered SERS nanoplatforms with automated ML frameworks for comprehensive sensor-state discrimination and next-generation intelligent biosensor development.

Keywords: surface-enhanced Raman spectroscopy, machine learning, pathogen detection, biosensing, automated diagnostics

1. Introduction

Surface-enhanced Raman spectroscopy (SERS) has emerged as a powerful analytical technique for molecular identification owing to its exceptional sensitivity, chemical specificity, and capability to generate distinctive vibrational fingerprints of analytes. By exploiting localized surface plasmon resonances on nanostructured metallic substrates, SERS enables the detection of biomolecules at extremely low concentrations, often down to the single-molecule level [1–4]. Compared with conventional analytical approaches such as microbial culture methods or molecular assays including polymerase chain reaction, SERS offers rapid analysis, minimal sample preparation, and label-free detection, making it highly attractive for applications in clinical diagnostics, environmental monitoring, and food safety [5, 6].

***Corresponding author:** Bryan Guilcapi & Lucia Petti, Institute of Applied Sciences and Intelligent Systems, CNR, Italy. Email: bryan.guilcapi@cnr.isasi.it, lucia.petti@cnr.isasi.it

Despite these advantages, the widespread adoption of SERS in real-world diagnostic settings has been hindered by several critical challenges, including signal variability, limited reproducibility across substrates, and the intrinsic complexity of Raman spectral data, particularly when dealing with biological samples [7, 8]. Variations in nanostructure morphology, analyte distribution, and measurement conditions can significantly affect spectral intensity and peak consistency, thereby complicating quantitative analysis and reliable classification. In this context, recent advances in automated SERS platforms have begun to address these limitations by integrating high-throughput data acquisition with advanced data processing and machine learning (ML) techniques [9–11]. Automation reduces operator-dependent variability and enables the collection of large spectral datasets with high spatial resolution, while ML algorithms facilitate the extraction of meaningful patterns from complex, high-dimensional spectral data. Collectively, these developments represent a critical step toward the realization of robust, scalable, and reproducible SERS-based diagnostic systems.

Multivariate statistical and machine learning methods have been extensively applied to Raman and SERS data to improve classification performance. Unsupervised techniques such as principal component analysis (PCA) are widely used for dimensionality reduction by transforming correlated spectral variables into a reduced set of orthogonal principal components (PCs), thereby preserving most of the variance while mitigating noise and redundancy and enabling more efficient and stable downstream model training [12–15]. In parallel, supervised learning approaches including support vector machines (SVMs) have demonstrated strong performance in Raman-based biological classification due to their robustness in high-dimensional feature spaces and effectiveness with limited training data [16–18], while linear discriminant analysis (LDA) projects data onto a discriminant subspace that maximizes class separability, providing interpretable linear decision boundaries when underlying distributional assumptions are satisfied [19]. Ensemble methods such as random forest (RF) further enhance classification robustness by aggregating multiple decision trees, thereby reducing overfitting and improving generalization in noisy datasets [20, 21]. More recently, deep learning approaches—particularly convolutional neural networks (CNNs)—have gained increasing attention in vibrational spectroscopy, as they can automatically learn hierarchical and nonlinear features directly from raw spectral data, eliminating the need for manual feature engineering [22, 23]. Their capacity to capture subtle spectral variations and handle complex class overlap has made CNNs especially promising for SERS-based classification of biological samples, including bacteria, viruses, and toxins [24, 25].

Alongside advances in data analytics, the development of reproducible and uniform SERS substrates remains a fundamental requirement for quantitative and high-throughput sensing. Silicon-based substrates coated with self-assembled gold nanoparticles have demonstrated excellent signal enhancement, stability, and scalability, providing dense and reproducible electromagnetic "hotspots" suitable for biological detection [26–29]. Building on these advances, in this work, we present a fully automated SERS-based diagnostic platform that integrates in-house fabricated, self-assembled gold nanoparticle sensors on silicon substrates with advanced ML and deep learning algorithms, enabling autonomous acquisition of thousands of Raman spectra with precise spatial mapping and supported by a robust database architecture for data storage, preprocessing, and model training. To assess platform performance in a clinically relevant multiclass classification scenario, the study was structured as a six-class classification framework encompassing distinct experimental sensor states rather than exclusively biological analytes. The investigated classes included bare silicon (Si) as a non-enhancing control, gold nanoparticle-coated silicon (Au NPs) as the active SERS substrate, and four analytes deposited on the SERS substrates—4-mercaptobenzoic acid (4-MBA), Shiga toxin 2 (Stx2), human coronavirus OC43 (OC43), and *Escherichia coli* (*E. coli*). Importantly, 4-MBA was incorporated as a Raman reporter molecule to validate SERS enhancement and surface functionalization efficiency and is not considered a biological target [30–32]. The biological targets—Stx2, OC43, and *E. coli*—were selected to represent a toxin, a virus, and a bacterium of significant public health relevance [33–35]. In particular, *E. coli* is a common indicator organism and a frequent cause of gastrointestinal infections [36–38], OC43 was chosen as a low-risk human coronavirus model relevant to recent coronavirus outbreaks [39–41], and Stx2 was selected as a model pathogenic toxin due to its extreme potency and critical role in severe human disease [42–44].

By integrating nanostructured SERS fabrication with classical ML and deep learning models, this work demonstrates robust discrimination across fabrication stages and biological functionalization conditions within an automated analytical pipeline.

2. Materials and Methods

2.1. Chemical reagents

Hydrogen tetrachloroaurate (III) trihydrate ($HAuCl_4 \cdot 3H_2O$), trisodium citrate ($Na_3C_6H_5O_7 \cdot 2H_2O$, 99%), absolute anhydrous ethanol, and toluene were purchased from Carlo Erba Reagents.

4-MBA (99%) and phosphate-buffered saline (PBS, pH 7.4) were obtained from Sigma-Aldrich. Silicon wafers were also supplied by Sigma-Aldrich and used as received.

2.2. Pathogens and toxins

Human coronavirus OC43 (HCoV-OC43, VR-1558™) was obtained from the American Type Culture Collection. Stx2 was produced from *Escherichia coli* C600 (933W) and purified by receptor analog affinity chromatography, followed by endotoxin removal using ActiCleanEtox columns (Sterogene Bioseparations, USA).

Pure cultures of pathogens and toxins were stored at −80 °C in aliquots and diluted in PBS prior to each experiment. Stx2 was provided by the Department of Medical and Surgical Sciences, University of Bologna, while *E. coli* samples were supplied by the Department of Veterinary Medicine and Animal Production, University of Naples Federico II.

2.3. Fabrication of SERS substrates

Gold nanoparticles (Au NPs) were synthesized via the citrate reduction method by adding 1% (w/v) sodium citrate to a boiling aqueous solution of $HAuCl_4$ (10^{-4} g/mL) under continuous stirring, followed by boiling for 30 min [45, 46]. The colloidal solution was allowed to cool naturally and subsequently centrifuged at 8000 rpm for 10 min. The resulting pellet was redispersed in a 1:1 (v/v) mixture of deionized water and ethanol [47].

SERS substrates were fabricated using a bottom-up, Langmuir–Blodgett-like self-assembly approach. A liquid interface was formed by layering 8 mL of toluene over 80 mL of water.

The Au NP dispersion (4 mL) was injected at a controlled rate of 3 mL/h using a mechanical syringe pump. As the toluene gradually evaporated, Au NPs self-assembled into a dense monolayer at the interface, which was subsequently transferred onto silicon wafers to obtain uniform and reproducible SERS-active substrates [47–50].

2.4. Sample deposition and SERS fingerprinting

For SERS fingerprint analysis, aliquots of each analyte were deposited directly onto the Au NP-coated silicon substrates. The following concentrations were used: Stx2 (154 ng/mL), OC43 (10^4 TCID$_{50}$/mL), *E. coli* (10^5 CFU/mL), and 4-MBA (100 µg/mL), all prepared in PBS.

After deposition, substrates were incubated overnight at 4 °C in a humidified chamber to facilitate analyte adsorption. Samples were then rinsed seven times with 1 mL of bi-distilled water and dried under a gentle nitrogen stream prior to Raman measurements.

2.5. Automated Raman spectral acquisition and workflow overview

Raman spectra were acquired using a 785 nm excitation laser (Laser-785nm-LAB-FC, Ocean Insight) with a beam diameter of 1 mm. The scattered signal was collected using an Enhanced QEPro Raman spectrometer (Ocean Insight) equipped with a 10 µm slit, covering a spectral range of 200–2500 cm^{-1}. The SERS substrates were mounted on an automated XY translation stage driven by two precision motors (M-403, PI Instruments) and controlled by C-863 controllers. Automated spectral mapping and data acquisition were performed using custom MATLAB routines (detailed in **Supplementary Materials**). The laser power was set to 20 mW, with an integration time of 10 s per spectrum and a sampling interval of 1 s. For silicon and Au NP substrates, three spectra were collected per point. For biological samples deposited on SERS substrates, two spectra (measurement and control) were acquired at each position to minimize potential laser-induced damage.

To provide an overview of the experimental and computational workflow implemented in this study, Figure 1 summarizes the fully automated SERS-based data acquisition and analysis pipeline. The platform integrates automated Raman spectral acquisition with precise spatial mapping, centralized data storage, and database construction, followed by multivariate preprocessing and feature extraction using techniques such as PCA and LDA. Supervised ML models are subsequently trained and optimized through systematic hyperparameter tuning, while deep learning architectures enable end-to-end spectral classification. Finally, the trained models are applied for independent validation and prediction, allowing assessment of both classification performance and overall system robustness. This integrated workflow enables automated, high-throughput SERS data generation, model training, and final validation within a unified framework.

2.6. Data preprocessing and outlier removal

Raw spectral data were visually inspected and preprocessed prior to analysis. Outlier detection was performed using the Mahalanobis distance, a multivariate statistical metric that accounts for covariance between variables and is well suited for high-dimensional spectral datasets [51–54].

The Mahalanobis distance is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \tag{1}$$

where x is the data vector (the point you want to measure), μ is the average vector (the average of the multivariate distribution), S is the covariance matrix of the data, S^{-1} is the inverse of the covariance matrix, and $(x - \mu)^T$ is the transposed vector of $(x - \mu)$.

For each sensor and analyte class, spectra falling outside predefined upper and lower Mahalanobis distance thresholds were excluded. This procedure enabled the construction of a cleaned and statistically robust dataset for subsequent analysis. Details of the preprocessing workflow and MATLAB implementation are provided in **Supplementary Materials**. Outlier removal was implemented as a quality control step to exclude spectra acquired from regions with insufficient nanoparticle coverage, excessive hotspot aggregation, or dominant silicon background, based on combined Mahalanobis distance thresholds and user-guided intensity range selection to retain reproducible and representative SERS signals.

2.7. Machine learning and deep learning analysis

2.7.1. Input data representation and preprocessing

All ML and deep learning models were trained using Raman spectra acquired over the full spectral range. Each spectrum was represented as a one-dimensional vector, where each element corresponds to an intensity value at a specific Raman shift. Spectral matrices were organized with samples as rows and spectral variables as columns. Prior to model training, spectra were standardized using z-score normalization. For classical ML models, normalization was applied to the entire dataset after data splitting. For deep learning models, normalization parameters (mean and standard deviation) were computed exclusively on the training set and subsequently applied to the validation data to prevent data leakage [55, 56].

2.7.2. PCA-based feature extraction

PCA was employed as an unsupervised dimensionality reduction technique to extract relevant spectral features. PCA was performed on standardized spectra using singular value decomposition. For visualization purposes, the first three PCs (PC1–PC3) were analyzed. For classification tasks, different PCA configurations were adopted depending on the classifier. In the PCA–SVM model, the first three PCs were retained as input features. In contrast, for the PCA–RF model, the number of retained components was selected to explain at least 95% of the cumulative variance of the dataset.

2.7.3. LDA-based feature extraction

LDA was used as a supervised dimensionality reduction technique to maximize class separability. LDA was implemented using a linear discriminant function and applied following a strict training–testing separation strategy. For classification tasks, the dataset was first divided into training (70%) and test (30%) subsets. The LDA transformation was fitted exclusively on the training data and subsequently applied to both training and test sets. The maximum number of LDA components was limited to the number of classes minus one. The resulting LDA features were used as input for SVM and RF classifiers.

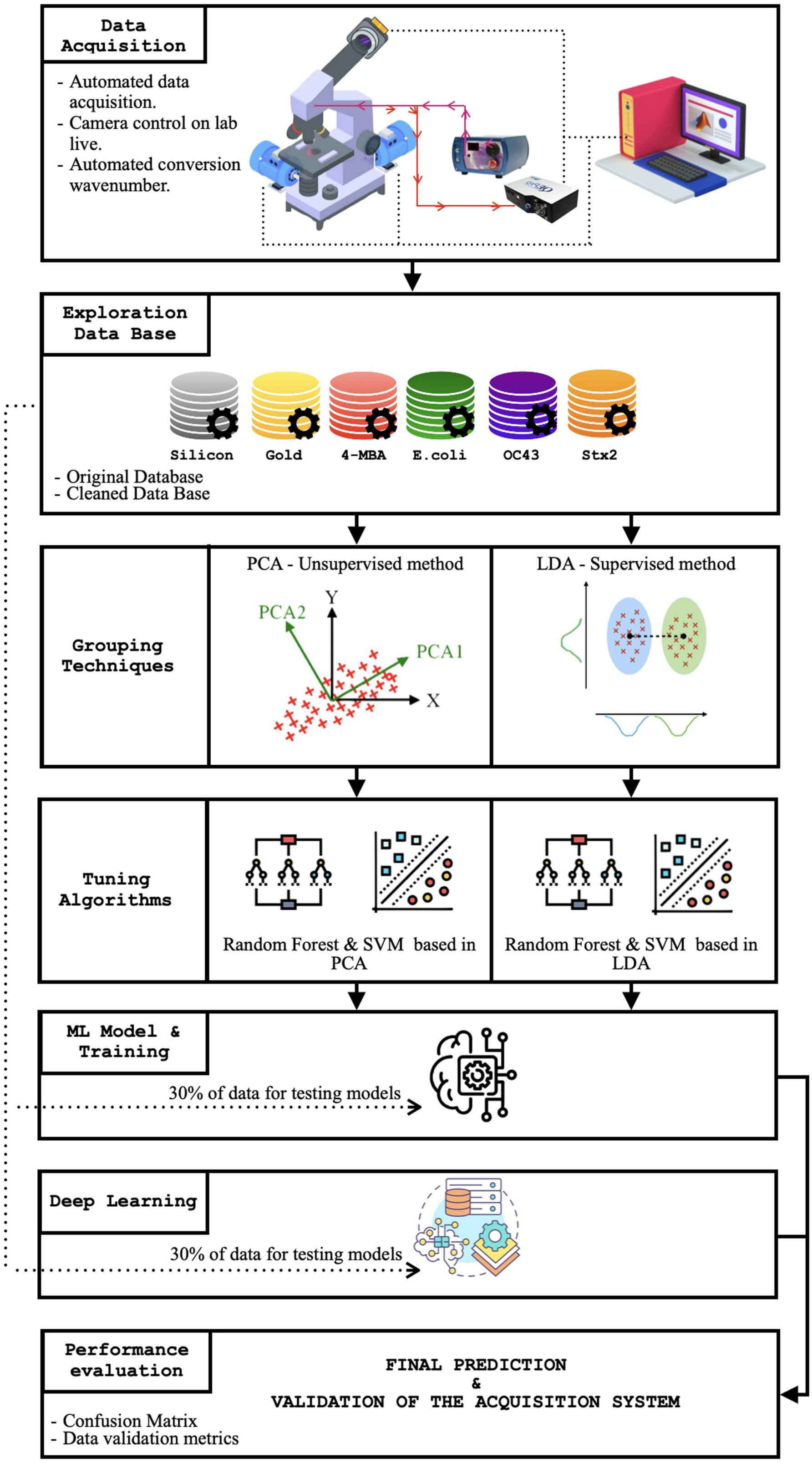

Data Acquisition
- Automated data acquisition.
- Camera control on lab live.
- Automated conversion wavenumber.
Exploration Data Base
Silicon
Gold
4-MBA
E.coli
OC43
Stx2
- Original Database
- Cleaned Data Base
Grouping Techniques
PCA - Unsupervised method
PCA2
Y
PCA1
X
LDA - Supervised method
Tuning Algorithms
Random Forest & SVM based in PCA
Random Forest & SVM based in LDA
ML Model & Training
30% of data for testing models
Deep Learning
30% of data for testing models
Performance evaluation
FINAL PREDICTION
&
VALIDATION OF THE ACQUISITION SYSTEM
- Confusion Matrix
- Data validation metrics

Figure 1. Schematic overview of the fully automated SERS-based experimental and computational workflow. The pipeline includes automated Raman spectral acquisition, database construction, multivariate preprocessing (PCA/LDA), machine learning training and hyperparameter tuning, deep learning-based classification, and final validation and prediction of the automated acquisition system

2.7.4. SVM classifiers (PCA–SVM and LDA–SVM)

SVM classifiers were implemented using an error-correcting output codes (ECOC) framework for multiclass classification. A linear kernel was employed in all cases. For PCA–SVM, the input consisted of the first three PCs derived from PCA. For LDA–SVM, the classifier was trained on the LDA-projected features obtained exclusively from the training set. Model performance was evaluated using a hold-out validation strategy (70% training, 30% testing), and cross-validation was performed on the training data to estimate generalization error.

2.7.5. Random forest classifiers (PCA–RF and LDA–RF)

RF classifiers were trained using an ensemble of 100 decision trees. For PCA–RF, the input features consisted of PCA components explaining at least 95% of the total variance. For LDA–RF, LDA-transformed features obtained from the training data were used. Out-of-bag (OOB) prediction and feature importance analysis were enabled to assess model robustness and feature relevance. Classification performance was evaluated using accuracy, confusion matrices, and class-wise precision, recall, and F1-score metrics.

2.7.6. One-dimensional convolutional neural network (1D-CNN)

A one-dimensional convolutional neural network (1D-CNN) was developed to perform end-to-end classification directly from normalized Raman spectra. Input spectra were reshaped into one-dimensional tensors and processed through two convolutional layers with kernel sizes of 5 and 3, respectively, followed by batch normalization, ReLU activation, and max-pooling operations. The convolutional layers were followed by a fully connected layer with 128 neurons, a dropout layer (dropout rate = 0.3), and a final softmax classification layer. The network was trained using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 64, and a maximum of 50 epochs. A validation split of 30% was used to monitor training performance.

A summary of all ML and deep learning models, including input features and key hyperparameters, is provided in Table 1 to facilitate reproducibility.

3. Results and Discussion

All reported performance metrics correspond to the predefined six-class sensor-state classification framework.

3.1. Dataset overview and spectral acquisition

Using the automated SERS acquisition platform, a total of 10,430 Raman spectra were collected across all experimental conditions. After applying the preprocessing and outlier removal procedure based on the Mahalanobis distance (Section 2.6), the dataset was refined to 8489 high-quality spectra, corresponding to approximately 54 h of total acquisition time.

The final dataset comprised six distinct classes: Si, gold nanoparticle-coated silicon (Au NPs), and four analytes deposited on SERS substrates (4-MBA, *E. coli*, OC43, and Stx2). The detailed distribution of sensors, sensing areas, number of measurements, and cleaned spectra for each class is reported in Table 2. This large and balanced dataset provided a robust basis for both classical ML and deep learning analyses.

3.2. Average SERS spectra and vibrational fingerprinting

Figure 2 shows the mean SERS spectra for each class, highlighting distinct spectral features associated with substrates and biological analytes. The silicon substrate exhibited a dominant phonon mode at 521 cm^{-1}, which was consistently observed across all spectra and served as an internal reference. In contrast,

Table 1. Summary of machine learning and deep learning models, architectures, and hyperparameters used in this study

Model	Feature input	Dimensionality	Main hyperparameters	Training details
PCA–SVM	PCA scores	3 PCs	Linear kernel, ECOC (one-vs-all)	70/30 hold-out, CV on training
PCA–RF	PCA scores	PCs ≥ 95% variance	100 trees, Gini split, OOB enabled	70/30 hold-out
LDA–SVM	LDA scores	≤5 LDs	Linear kernel, ECOC	70/30 hold-out, 5-fold CV
LDA–RF	LDA scores	≤5 LDs	100 trees, OOB enabled	70/30 hold-out
1D–CNN	Full spectra	N spectral points	Conv (32,5), Conv (64,3), FC (128), Dropout 0.3	Adam, LR = 1e-4, batch = 64, 50 epochs

Table 2. Experimental classes and SERS data acquisition parameters for each sample type

Classes	Number of sensors [*u*]	Sensing area [mm^2]	Measures per sensor [*u*]	Full database [*u*]	Clean database [*u*]	Measurement time per sensor [*mins*]
Silicon	10	16	243	2430	2302	50
Au NPs	10	16	243	2430	2339	50
4-MBA	2	9	650	1300	975	150
E. coli	2	9	650	1300	884	150
OC43	2	9	650	1300	976	150
Stx2	2	9	650	1300	1013	150

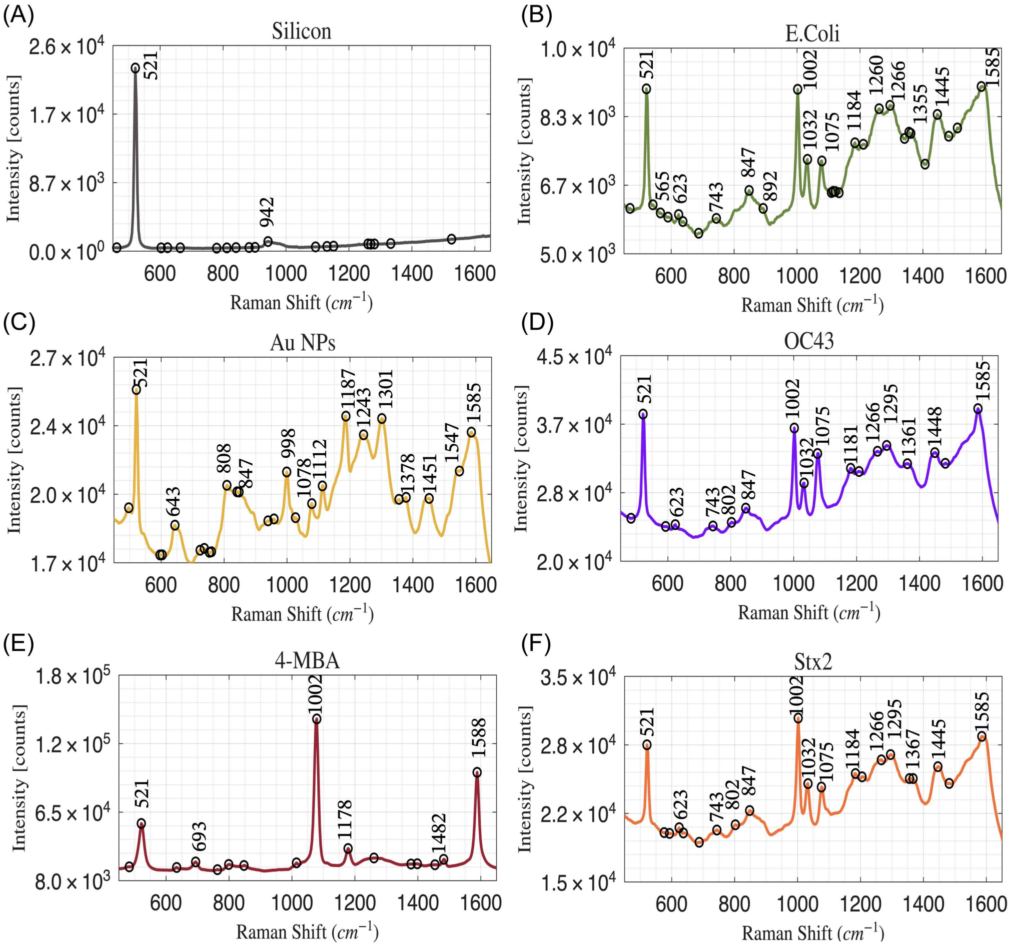


Figure 2. Spectral average for each pathogen, SERS detection, and peak identification: (A) Si, (B) *E. coli*, (C) Au NPs, (D) OC43, (E) 4-MBA, and (F) Stx2

Au NP-coated substrates displayed an increased number of vibrational modes, reflecting enhanced electromagnetic field localization.

The Raman reporter molecule 4-MBA exhibited sharp and well-defined peaks characteristic of aromatic ring vibrations and C–C stretching modes, confirming its suitability as a SERS benchmark analyte. Biological samples showed more complex spectral signatures: *E. coli* spectra presented contributions from nucleic acids, proteins, and lipids; OC43 spectra were dominated by vibrational modes associated with nucleic acids and protein backbones; and Stx2 spectra exhibited characteristic amide and protein-related vibrations.

Across all analytes, the number and distribution of vibrational modes were consistent with assignments reported in the literature (Table 3), confirming the reliability and reproducibility of the fabricated SERS substrates.

Table 3. Relevant vibrational modes

Raman SERS [cm^{-1}]	Au NPs	4-MBA	*E. coli*	OC43	Stx2	Refs.
469	NaN	NaN	C–C skeletal stretching	NaN	NaN	[57]
483	NaN	Ring deformation	NaN	RNA backbone	NaN	[58]
521	TO phonon mode of silicon substrate	TO phonon mode of silicon substrate	TO phonon mode of silicon substrate	TO phonon mode of silicon substrate	TO phonon mode of silicon substrate	[59]
541–565	NaN	NaN	POz–symmetric stretch	NaN	NaN	[60]

(*Continued*)

Table 3. (*Continued*)

Raman SERS [cm^{-1}]	Au NPs	4-MBA	*E. coli*	OC43	Stx2	Refs.
576	NaN	NaN	NaN	NaN	C–S bending	[61]
589	NaN	NaN	Tyrosine or adenine ring breathing	NaN	NaN	[62]
593	NaN	NaN	NaN	C–S stretching	C–S stretching	[63]
623	NaN	NaN	Carotenoid ring vibrations	POz–symmetric stretching of RNA	Phosphate backbone	[64]
633	NaN	C–H in–lane bending	NaN	NaN	NaN	[65]
637	NaN	NaN	Ring breathing mode	NaN	Ring breathing mode	[64]
687	NaN	NaN	Cytosine or guanine	NaN	Protein secondary	[66]
693	NaN	C–C bending	NaN	NaN	NaN	[67]
743	NaN	NaN	Adenine	Adenine	DNA/RNA	[57]
763–799	NaN	C–H bending	NaN	NaN	NaN	[67]
802	NaN	NaN	NaN	C–O–C bending	C–O–C bending	[66, 68]
847	X	C–H deformation	C–H deformation	Tyrosine	Tyrosine ring breathing	[61, 69]
892	NaN	NaN	Protein backbone vibrations	NaN	NaN	[62]
1002	NaN	NaN	Phenylalanine symmetric	Phenylalanine symmetric	Phenylalanine symmetric	[70]
1014	NaN	Benzene ring breathing	NaN	NaN	NaN	[69]
1032	NaN	NaN	X	C–N stretch	C–N stretch	[66]
1075	NaN	NaN	NaN	RNA backbone phosphate	RNA backbone phosphate	[68]
1078	C=C stretching	C=C stretching	C=C stretching	NaN	NaN	[66, 69, 71]
1109	NaN	NaN	Phospholipid C–C/C–N stretching	NaN	NaN	[70]
1178	NaN	C–H bending	NaN	NaN	NaN	[65]
1181	NaN	NaN	NaN	C–H bending	NaN	[68]
1184	NaN	NaN	C–H bending	NaN	Amide III region	[62, 64]
1187	C–H or C–C	NaN	NaN	NaN	NaN	[72]
1243	C–N or C–O	NaN	NaN	NaN	NaN	[72]
1260	NaN	C–N or C–O	Amide III	NaN	NaN	[65]
1266–1295	NaN	NaN	NaN	Amide III	Amide III	[68]
1301	C=C stretching	NaN	NaN	NaN	NaN	[73]
1341	NaN	NaN	Amide III	NaN	NaN	[68]
1355	D-band of carbon	NaN	X	NaN	C–H bending	[72]
1361	NaN	NaN	X	Nucleic acid	NaN	[68]
1378–1400	X	Symmetric COO	NaN	NaN	NaN	[65]
1406	NaN	NaN	COO symmetric stretching	NaN	NaN	[70]
1445	NaN	NaN	CH2/CH3 deformation	NaN	CH2/CH3 deformation	[62]

(*Continued*)

Table 3. (*Continued*)

Raman SERS [cm^{-1}]	Au NPs	4-MBA	*E. coli*	OC43	Stx2	Refs.
1448	NaN	NaN	NaN	CH2/CH3 deformation	NaN	[68]
1451	CH2 bending	NaN	NaN	NaN	NaN	[73]
1454	NaN	C=C stretching	NaN	NaN	NaN	[65]
1482	NaN	C=C stretching	C=C stretching	C=C stretching	C=C stretching	[65, 66]
1509	NaN	NaN	C=C ring vibrations	NaN	NaN	[62]
1547	C=C stretching	NaN	NaN	NaN	NaN	[73]
1585–1588	C=C stretching	C=C stretching	C=C stretching	C=C stretching	C=C stretching	[73]

3.3. Unsupervised analysis: principal component analysis (PCA)

PCA was applied to the cleaned spectral dataset to explore intrinsic clustering behavior without using class labels. The first two PCs accounted for 96.9% of the total variance, with PC1 explaining 66.74% and PC2 30.16% of the variance.

Figure 3(A) shows the three-dimensional projection of the first three PCs, revealing partial clustering among the six classes. Substrate-related classes (Si and Au NPs) formed distinct clusters, while biological analytes exhibited partial overlap, particularly between OC43 and Stx2.

Loading plots (Figure 3B) indicate that specific Raman shift regions contribute disproportionately to class separation, suggesting that these vibrational modes play a key role in discriminating between different pathogens and substrates. While PCA enabled a meaningful reduction of data dimensionality, the observed overlap highlighted the need for supervised classification methods to improve discrimination.

3.4. PCA-based supervised classification

3.4.1. Support vector machine (PCA-SVM)

An SVM classifier with ECOC was trained using PCA-reduced features. The confusion matrix (Figure 3C) demonstrates high classification accuracy for several classes, including Si and Au NPs (≈100%) and 4-MBA (≈99%).

However, misclassification was observed among biologically similar classes. OC43 and Stx2 exhibited increased confusion, with classification accuracies of approximately 92% and 95%, respectively. Figure 3(D) compares the distribution of real data and predicted class labels in PCA space, showing good overall agreement while highlighting regions of class overlap. These results indicate that PCA–SVM performs well for substrates and simple analytes but encounters limitations when separating complex biological spectra.

3.4.2. Random forest classifiers (PCA-RF)

RF classification was also applied to PCA-reduced data. As shown in Figure 3(E), the PCA–RF model achieved near-perfect accuracy for Si and Au NP substrates, with strong performance for 4-MBA (≈97–98%).

Lower accuracy was observed for *E. coli*, OC43, and Stx2, with classification rates ranging between 79% and 83%, reflecting spectral overlap and biological complexity. The real versus predicted data distribution (Figure 3F) confirms the robustness of the RF classifier while revealing misclassification trends similar to those observed in PCA–SVM. Overall, PCA–RF demonstrated improved tolerance to noise but limited discriminatory power for closely related biological classes.

3.5. Supervised dimensionality reduction: linear discriminant analysis (LDA)

LDA was applied to enhance class separability using labeled data. Figure 4A presents the one-dimensional LDA projection with fitted Gaussian distributions, where improved separation among classes is evident. The three-dimensional LDA projection (Figure 4B) further highlights enhanced clustering compared to PCA, particularly for substrate-related classes.

Despite improved separability, partial overlap persisted among biological analytes, indicating that linear discriminant boundaries alone may be insufficient for complete discrimination in complex SERS datasets.

3.6. LDA-based classification

3.6.1. Support vector machine (LDA–SVM)

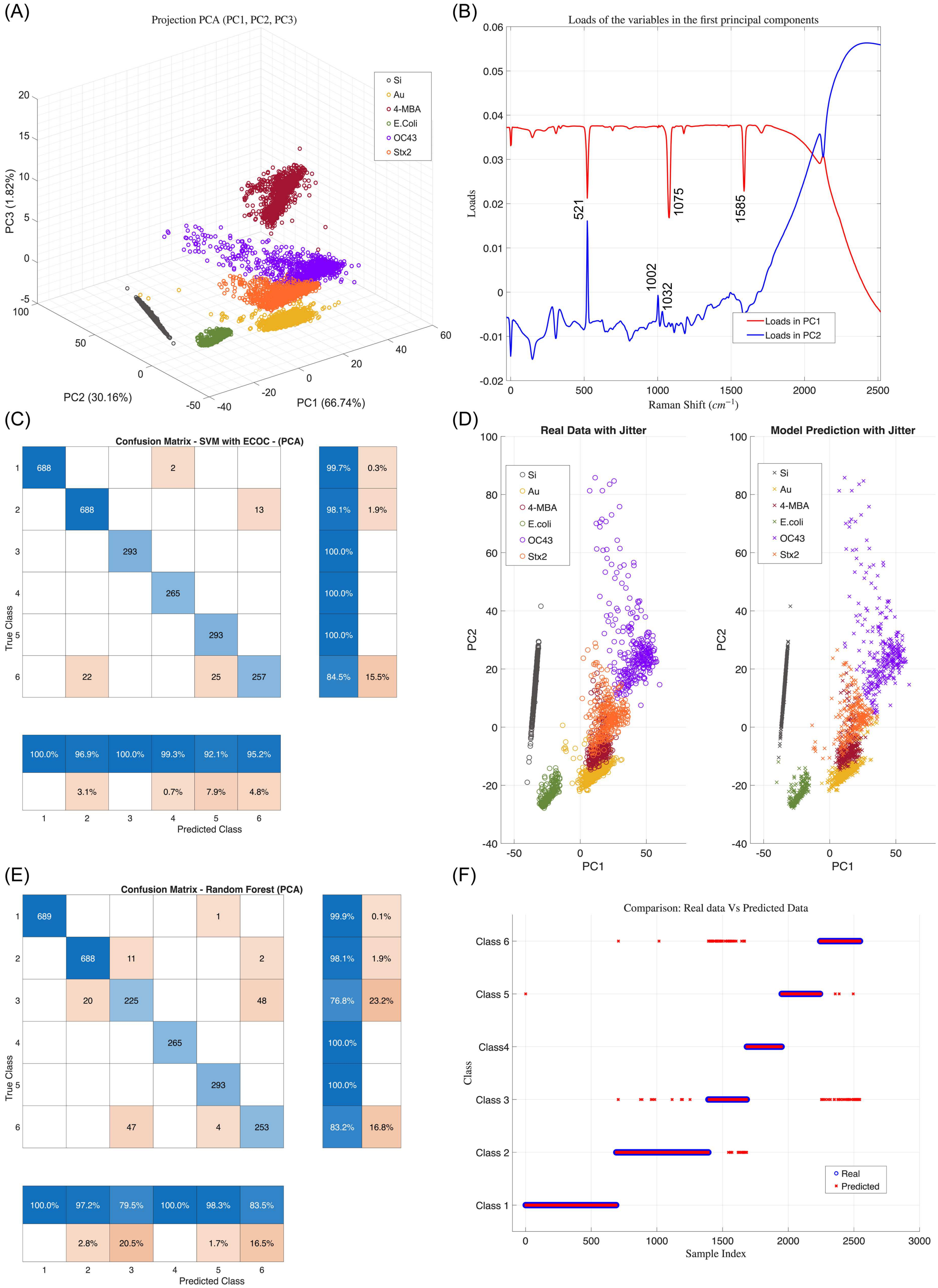
The LDA–SVM model achieved high classification accuracy for Si (93%), Au NPs (99%), and 4-MBA (98%) (Figure 4C). In contrast, classification performance decreased for *E. coli*, OC43, and Stx2, with accuracies dropping below 70% in some cases. Significant confusion between OC43 and Stx2 was observed, reflecting similarities in their vibrational signatures.

The comparison between real and predicted data distributions (Figure 4D) confirms the model's strong performance for substrates and simple analytes, while highlighting its limitations for complex biological samples.

3.6.2. Random forest classifier (LDA–RF)

The LDA–RF model demonstrated moderate classification performance across all classes (Figure 4E). Substrate and reporter molecule classes achieved accuracies above 90%, whereas biological analytes showed reduced accuracy, particularly OC43 (32%) and *E. coli* (55%).

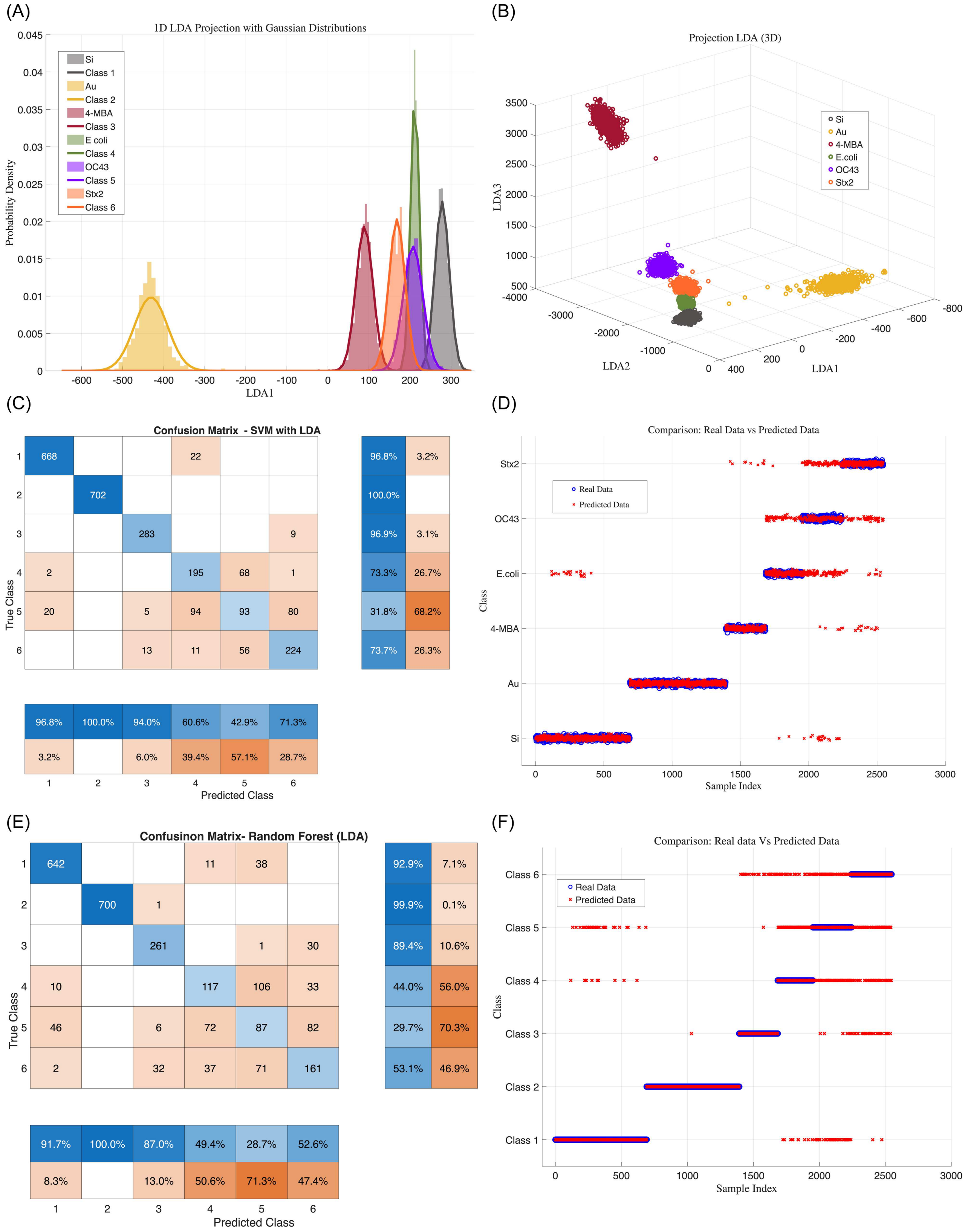
Although RF benefitted from the supervised dimensionality reduction provided by LDA, spectral overlap among biological



True Class	1	2	3	4	5	6		
1	688			2			99.7%	0.3%
2		688				13	98.1%	1.9%
3			293				100.0%	
4				265			100.0%	
5					293		100.0%	
6		22			25	257	84.5%	15.5%
	100.0%	96.9%	100.0%	99.3%	92.1%	95.2%		
		3.1%		0.7%	7.9%	4.8%		

True Class	1	2	3	4	5	6		
1	689				1		99.9%	0.1%
2		688	11			2	98.1%	1.9%
3		20	225			48	76.8%	23.2%
4				265			100.0%	
5					293		100.0%	
6			47		4	253	83.2%	16.8%
	100.0%	97.2%	79.5%	100.0%	98.3%	83.5%		
		2.8%	20.5%		1.7%	16.5%		

Figure 3. The results of the principal component analysis: (A) PCA projection in 3D and (B) loads of the variables in the 2-PC. The results of support vector machine based on PCA: (C) Confusion matrix and (D) data distribution in PCA space (left) vs the predicted classification results (right). The results of random forest based on PCA: (E) Confusion matrix and (F) real data distribution (blue circle) vs predicted data (red x).



Confusion Matrix - SVM with LDA

	1	2	3	4	5	6		
1	668			22			96.8%	3.2%
2		702					100.0%	
3			283			9	96.9%	3.1%
4	2			195	68	1	73.3%	26.7%
5	20		5	94	93	80	31.8%	68.2%
6			13	11	56	224	73.7%	26.3%
	96.8%	100.0%	94.0%	60.6%	42.9%	71.3%		
	3.2%		6.0%	39.4%	57.1%	28.7%		

Confusinon Matrix- Random Forest (LDA)

	1	2	3	4	5	6		
1	642			11	38		92.9%	7.1%
2		700	1				99.9%	0.1%
3			261		1	30	89.4%	10.6%
4	10			117	106	33	44.0%	56.0%
5	46		6	72	87	82	29.7%	70.3%
6	2		32	37	71	161	53.1%	46.9%
	91.7%	100.0%	87.0%	49.4%	28.7%	52.6%		
	8.3%		13.0%	50.6%	71.3%	47.4%		

Figure 4. The results of LDA: (A) Gaussian distribution with class separation in 1-D and (B) LDA projection in 3D. The results of support vector machine based on LDA: (C) Confusion matrix SVM and (D) real data distribution (blue circle) vs the predicted data (red x). The results of random forest based on LDA: (E) Confusion matrix and (F) real data distribution (blue circle) vs predicted data (red x).

samples remained a limiting factor. The real versus predicted data distribution (Figure 4F) further illustrates these trends.

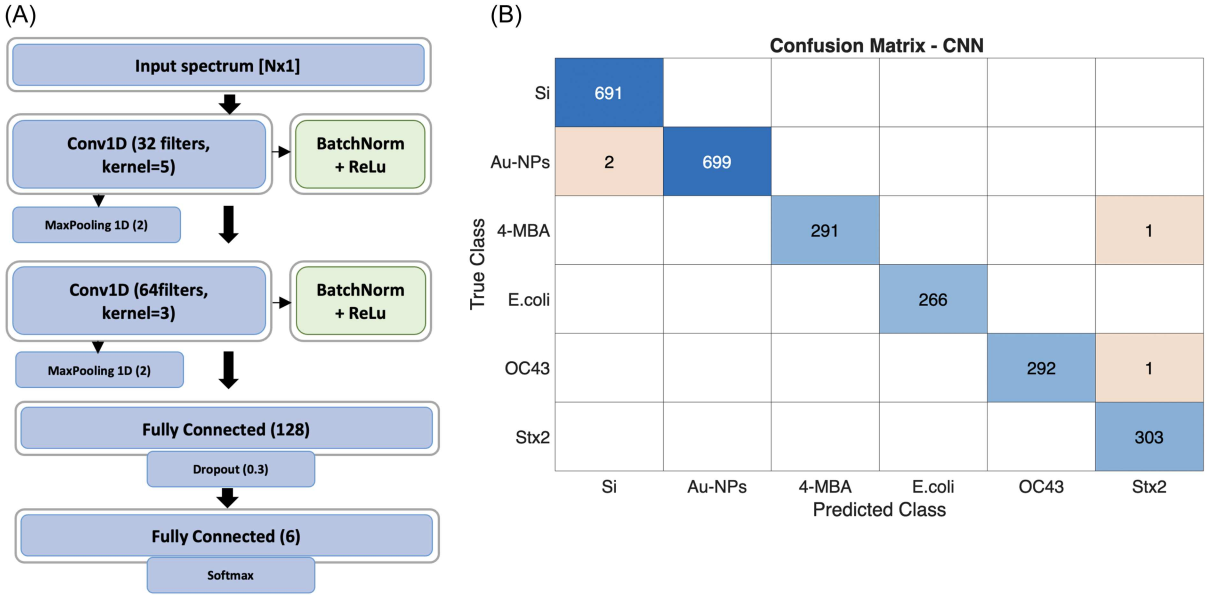
3.7. Deep learning approach: convolutional neural network (CNN)

A custom 1D-CNN was trained directly on the preprocessed spectral data without prior dimensionality reduction, enabling end-to-end learning of discriminative spectral features. The CNN architecture (Figure 5A) effectively captured both local and global spectral patterns relevant for multiclass classification. The confusion matrix obtained from the independent test dataset (Figure 5B) demonstrates an overall classification accuracy of 99.84%, indicating excellent discriminative performance across all six classes. Class-wise performance metrics further confirm the robustness of the model, with precision, recall, and F1-scores approaching unity for all classes (Table 4). In particular, perfect classification was achieved for *E. coli* (precision = 1.000, recall = 1.000, F1-score = 1.000), while Si, Au NPs, and 4-MBA also exhibited near-perfect performance (F1-scores ≥ 0.998). Only minimal misclassification was observed between OC43 and Stx2, consistent with their partially overlapping spectral features, yet both classes maintained F1-scores above 0.996. Cross-validation analysis further demonstrated the stability of the deep learning model, yielding an average accuracy of 99.84% ± 0.07% and consistently high class-wise F1-scores. Overall, the CNN substantially outperformed all classical ML approaches evaluated in this study, highlighting the advantage of deep learning for automated, high-accuracy SERS-based biological classification. Despite moderate differences in class sample sizes, the CNN achieved consistently high and balanced class-wise performance, indicating that classification accuracy was not driven by majority classes.

3.8. Discussion

This study demonstrates the effectiveness of integrating a fully automated SERS acquisition platform with advanced ML and deep learning methods for robust multiclass classification of chemical and biological targets. The combination of reproducible gold nanoparticle-based substrates, high-throughput Raman mapping, and centralized data handling enabled the generation of a large and statistically robust spectral dataset, providing a strong foundation for reliable model training and evaluation.

Unsupervised PCA analysis revealed clear separation between substrate-related classes while highlighting partial spectral overlap among biologically complex samples. This behavior is consistent with previous reports showing that bacteria, viruses, and toxins share common molecular constituents—such as proteins, nucleic acids, and lipids—leading to overlapping vibrational



True Class \ Predicted Class	Si	Au-NPs	4-MBA	E.coli	OC43	Stx2
Si	691					
Au-NPs	2	699				
4-MBA			291			1
E.coli				266		
OC43					292	1
Stx2						303

Figure 5. The proposed (A) one-dimensional convolutional neural network (1D-CNN) architecture for Raman spectral classification. The network operates directly on normalized one-dimensional spectral vectors and (B) confusion matrix using CNN

Table 4. Class-wise precision, recall, and F1-score of the CNN model obtained from stratified cross-validation on the SERS spectral dataset

Classes	Precision	Recall	F1-score
Silicon	0.99711	1	0.99855
Au NPs	1	0.99715	0.99857
4-MBA	1	0.99658	0.99828
E. coli	1	1	1
OC43	1	0.99659	0.99829
Stx2	0.99344	1	0.99671

Table 5. Comparison of the proposed fully automated SERS–CNN platform with representative state-of-the-art SERS-based biosensing and spectroscopic classification studies

Study	Target	Method	Classes	Automation	Reported accuracy %
Acquarelli et al. [22]	Various chemicals	CNN	Binary/few	No	~95–98
Li et al. [24]	Bacteria	CNN + SERS	Binary	No	~97–99
Gao et al. [25]	Toxins	Deep learning	Binary	Partial	~96–98
Chen et al. [17]	Pathogens	SVM + PCA	Binary	No	~90–96
Hu et al. [18]	Bacteria	RF/SVM	3 classes	No	~93–97
This work	Substrate + molecule + toxin + virus + bacterium	CNN + automated SERS	6 classes	Yes (fully automated)	≈99.8–99

signatures. Although PCA was useful for exploratory analysis, its limited ability to fully separate closely related biological classes underscores the intrinsic challenges of unsupervised linear methods for complex SERS datasets.

Supervised classical ML models, including SVM and RF classifiers combined with PCA or LDA, achieved strong performance for simple analytes and substrates but showed reduced accuracy for spectrally similar biological hazards. These findings indicate that classical approaches may struggle to capture subtle nonlinear relationships embedded in complex biological SERS spectra.

In contrast, the CNN achieved near-perfect classification across all six classes by learning hierarchical and nonlinear spectral features directly from standardized raw spectra. This end-to-end learning capability enables the CNN to exploit subtle intensity variations distributed across multiple spectral regions, providing a clear advantage for complex SERS-based biosensing tasks.

As summarized in Table 5, the proposed framework achieves performance that is competitive with or superior to representative state-of-the-art SERS-based biosensing and spectroscopic deep learning studies, while addressing a more challenging multiclass scenario and incorporating fully automated high-throughput acquisition. Beyond algorithmic accuracy, automation reduces operator-dependent variability and supports scalable data generation. Although this study was conducted under controlled laboratory conditions, future work will extend the platform to more complex matrices and independent datasets to further assess generalization and real-world applicability.

4. Conclusion

In conclusion, this study presents an integrated and fully automated SERS-based diagnostic platform that combines reproducible nanostructured substrates with advanced ML and deep learning algorithms. Through a systematic comparison of classical multivariate methods and CNNs, we demonstrate that deep learning significantly outperforms traditional approaches in the classification of complex biological SERS spectra.

The proposed platform enables rapid, label-free, and highly accurate discrimination of substrates, chemical reporters, bacteria, viruses, and toxins, highlighting its potential for scalable and user-independent biosensing applications. By bridging automated spectral acquisition with intelligent data analysis, this work contributes to the advancement of SERS from a laboratory-based analytical technique toward a practical tool for real-world diagnostics. Future developments aimed at validating the system with clinical and environmental samples will further support its translation into routine diagnostic workflows.

Acknowledgment

The authors gratefully acknowledge the support for this work from OPUS automazione S.p.A. Financial support from the project (CUP) B25H23002780007 EMDAS—“Environment Mobile Data Acquisition System” PR Campania FESR 21/27, PR CAMPANIA FSE + 21/27, and POC 14 /20.

The authors gratefully acknowledge the support for this work from the European Union—NextGenerationEU Call PRIN2022 Development of a plasmonic nanobiosensor for the rapid diagnosis of Shiga toxin-producing *E. coli* human infections at the point of care—SENSOSTEC (CUP B53D23019990006).

Funding Support

This work is sponsored by 2019 Project of Humanities and Social Sciences of Henan Provincial Department of Education: “Research on Translation for Overseas Publicity from the Perspective of the Persuasion Theory in Western Rhetoric” (2019-ZZJH-643); 2019 Teaching Reform Project of School of Foreign Studies, North China University of Water Resources and Electric Power: “The Reforming Design and Practice of Mixed Teaching of Online and Offline Course for English Rhetoric”; and 2019 Teaching Reform Project of Henan Province: A Research on the “Golden Lesson” of College English from the Perspective of Telling Chinese Stories (2019SJGLX284).

Ethical Statement

This study did not involve human participants or animals. Therefore, ethical approval and informed consent were not required.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The clean data that support the findings of this study are openly available in public repositories: https://github.com/BryanGuilcapi/Machine-Learning-Enhanced-SERS-on-Silicon-Gold-Sensors-for-Ultra-High-Throughput-Pathogen-Detection.

The original database is accessible by requesting the corresponding author.

Author Contribution Statement

Bryan Guilcapi1: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Alessia Milano:** Software, Validation, Investigation, Data curation, Writing – review & editing. **Amalia D' Avino:** Investigation, Writing – review & editing. **Domenico Sagnelli:** Validation, Investigation. **Massimo Rippa:** Investigation. **Valentina Marchesano:** Investigation. **Ivan Salvatore Perrotta:** Investigation. **Rosa Luisa Ambrosio:** Investigation. **Giovanna Fusco:** Investigation. **Maurizio Brigotti:** Investigation, Writing – review & editing. **Stefano Morabito:** Resources. **Lucia Petti:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Leong, S. X., Leong, Y. X., Tan, E. X., Sim, H. Y. F., Koh, C. S. L., Lee, Y. H., . . . , & Ling, X. Y. (2022). Noninvasive and point-of-care surface-enhanced Raman scattering (SERS)-based breathalyzer for mass screening of coronavirus disease 2019 (COVID-19) under 5 min. *ACS Nano*, *16*(2), 2629–2639. https://doi.org/10.1021/acsnano.1c09371

[2] Rippa, M., Marchesano, V., Vestri, A., Sagnelli, D., Fusco, G., Zyss, J., . . . , & Petti, L. (2024). Fractal plasmonic molecule for multi-sensing: SERS platform for SARS-CoV-2 detection. *ACS Applied Nano Materials*, *7*(7), 6958–6968. https://doi.org/10.1021/acsanm.3c06006

[3] Lin, L. L., Alvarez-Puebla, R., Liz-Marzan, L. M., Trau, M., Wang, J., Fabris, L., . . . , & Ye, J. (2025). Surface-enhanced Raman spectroscopy for biomedical applications: Recent advances and future challenges. *ACS Applied Materials & Interfaces*, *17*(11), 16287–16379. https://doi.org/10.1021/acsami.4c17502

[4] Usman, M., Tang, J.-W., Li, F., Lai, J.-X., Liu, Q.-H., Liu, W., & Wang, L. (2023). Recent advances in surface enhanced Raman spectroscopy for bacterial pathogen identifications. *Journal of Advanced Research*, *51*, 91–107. https://doi.org/10.1016/j.jare.2022.11.010

[5] Rippa, M., Sagnelli, D., Vestri, A., Marchesano, V., Munari, B., Carnicelli, D., . . . , & Petti, L. (2022). Plasmonic metasurfaces for specific SERS detection of Shiga toxins. *ACS Applied Materials & Interfaces*, *14*(4), 4969–4979. https://doi.org/10.1021/acsami.1c21553

[6] Xu, Y., Chen, R., Jiang, S., Zhou, L., Jiang, T., Gu, C., . . . , & Zhou, J. (2023). Insights into the semiconductor SERS activity: The impact of the defect-induced energy band offset and electron lifetime change. *ACS Applied Materials & Interfaces*, *15*(35), 42026–42036. https://doi.org/10.1021/acsami.3c06363

[7] McAlpine, J. B., Chen, S. N., Kutateladze, A., MacMillan, J. B., Appendino, G., Barison, A., . . . , & Pauli, G. F. (2019). The value of universally available raw NMR data for transparency, reproducibility, and integrity in natural product research. *Natural Product Reports*, *36*(1), 35–107. https://doi.org/10.1039/C7NP00064B

[8] Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

[9] Khan, S., Ullah, R., Khan, A., Ashraf, R., Ali, H., Bilal, M., & Saleem, M. (2018). Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. *Photodiagnosis and Photodynamic Therapy*, *23*, 89–93. https://doi.org/10.1016/j.pdpdt.2018.05.010

[10] Guleken, Z., Jakubczyk, P., Wiesław, P., Krzysztof, P., Bulut, H., Öten, E., . . . , & Tarhan, N. (2022). Characterization of Covid-19 infected pregnant women sera using laboratory indexes, vibrational spectroscopy, and machine learning classifications. *Talanta*, *237*, 122916. https://doi.org/10.1016/j.talanta.2021.122916

[11] Luo, R., Popp, J., & Bocklitz, T. (2022). Deep learning for Raman spectroscopy: A review. *Analytica*, *3*(3), 287–301. https://doi.org/10.3390/analytica3030020

[12] Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

[13] He, X., Liu, Y., Huang, S., Liu, Y., Pu, X., & Xu, T. (2018). Raman spectroscopy coupled with principal component analysis to quantitatively analyze four crystallographic phases of explosive CL-20. *RSC Advances*, *8*(41), 23348–23352. https://doi.org/10.1039/C8RA02189A

[14] Senger, R. S., & Scherr, D. (2020). Resolving complex phenotypes with Raman spectroscopy and chemometrics. *Current Opinion in Biotechnology*, *66*, 277–282. https://doi.org/10.1016/j.copbio.2020.09.007

[15] Laghmati, S., Hamida, S., Hicham, K., Cherradi, B., & Tmiri, A. (2023). An improved breast cancer disease prediction system using ML and PCA. *Multimedia Tools and Applications*, *83*(11), 33785–33821. https://doi.org/10.1007/s11042-023-16874-w

[16] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, *15*(1), 41–51. https://doi.org/10.21873/cgp.20063

[17] Chen, C., Yang, L., Zhao, J., Yuan, Y., Chen, C., Tang, J., . . . , & Lv, X. (2020). Urine Raman spectroscopy for rapid and inexpensive diagnosis of chronic renal failure (Crf) using multiple classification algorithms. *Optik*, *203*, 164043. https://doi.org/10.1016/j.ijleo.2019.164043

[18] Hu, S., Li, H., Chen, C., Chen, C., Zhao, D., Dong, B., . . . , & Xie, Y. (2022). Raman spectroscopy combined with machine learning algorithms to detect adulterated Suichang native honey. *Scientific Reports*, *12*(1), 3456. https://doi.org/10.1038/s41598-022-07222-3

[19] Gatto, B. B., de Souza, L. S., & Dos Santos, E. M. (2017). A deep network model based on subspaces: A novel approach for image classification. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, 436–439.

[20] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[21] Liaw, A., & Wiener, M. (2002). Classification and regression by RandomForest. *R news*, *2*(3), 18–22.

[22] Acquarelli, J., van Laarhoven, T., Gerretzen, J., Tran, T. N., Buydens, L. M., & Marchiori, E. (2017). Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica Chimica Acta*, *954*, 22–31. https://doi.org/10.1016/j.aca.2016.12.010

[23] Desai, M., & Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer

perceptron neural network (MLP) and convolutional neural network (CNN). *Clinical eHealth*, *4*, 1–11. https://doi.org/10.1016/j.ceh.2020.11.002

[24] Li, Y., Chen, C., Chen, F., Chen, C., Gao, R., Yang, B., . . . , & Lv, X. (2021). Serum Raman spectroscopy combined with Deep Neural Network for analysis and rapid screening of hyperthyroidism and hypothyroidism. *Photodiagnosis and Photodynamic Therapy*, *35*, 102382. https://doi.org/10.1016/j.pdpdt.2021.102382

[25] Gao, R., Yang, B., Chen, C., Chen, F., Chen, C., Zhao, D., & Lv, X. (2021). Recognition of chronic renal failure based on Raman spectroscopy and convolutional neural network. *Photodiagnosis and Photodynamic Therapy*, *34*, 102313. https://doi.org/10.1016/j.pdpdt.2021.102313

[26] Lai, W., Zhou, J., Jia, Z., Petti, L., & Mormile, P. (2015). Ag@Au hexagonal nanorings: Synthesis, mechanistic analysis and structure-dependent optical characteristics. *Journal of Materials Chemistry C*, *3*(37), 9726–9733. https://doi.org/10.1039/C5TC02017D

[27] Liu, Y., Zhou, J., Yuan, X., Jiang, T., Petti, L., Zhou, L., & Mormile, P. (2015). Hydrothermal synthesis of gold polyhedral nanocrystals by varying surfactant concentration and their LSPR and SERS properties. *RSC Advances*, *5*(84), 68668–68675. https://doi.org/10.1039/C5RA10781D

[28] Galeotti, F., Pisco, M., & Cusano, A. (2018). Self-assembly on optical fibers: A powerful nanofabrication tool for next generation "lab-on-fiber" optrodes. *Nanoscale*, *10*(48), 22673–22700. https://doi.org/10.1039/C8NR06002A

[29] Pisco, M., & Galeotti, F. (2021). Nano- and micropatterning on optical fibers by bottom-up approach: The importance of being ordered. *Applied Sciences*, *11*(7), 3254. https://doi.org/10.3390/app11073254

[30] Jaitpal, S., Chavva, S. R., & Mabbott, S. (2022). 3d printed SERS-active thin-film substrates used to quantify levels of the genotoxic isothiazolinone. *ACS Omega*, *7*(3), 2850–2860. https://doi.org/10.1021/acsomega.1c05707

[31] Cardellini, J., Dallari, C., De Santis, I., Riccio, L., Ceni, C., Morrone, A., . . . , & Berti, D. (2024). Hybrid lipid-AuNP clusters as highly efficient SERS substrates for biomedical applications. *Nature Communications*, *15*(1), 7975. https://doi.org/10.1038/s41467-024-52205-9

[32] Vu, N. N., Ng, K. W., Jaitpal, S., Negahdary, M., Nguyen, T., Kodam, R. S., & Mabbott, S. (2025). High precision automated synthesis of surface-enhanced (Resonance) Raman nanotags. *ACS Sensors*, *10*(5), 3515–3529. https://doi.org/10.1021/acssensors.5c00068

[33] D'Avino, A., Milano, A., Marchesano, V., Guilcapi, B., Sagnelli, D., Rippa, M., . . . , & Petti, L. (2025). Flexible gold nanoparticle SERS tape for rapid, label-free and ultrasensitive detection and differentiation of Shiga toxin variants. *Biosensors and Bioelectronics: X*, *27*, 100696. https://doi.org/10.1016/j.biosx.2025.100696

[34] Milano, A., D'Avino, A., Marchesano, V., Sagnelli, D., Rippa, M., Guilcapi, B., . . . , & Petti, L. (2025). Advancing medical diagnostics: Rapid, label-free detection and differentiation of Shiga toxin variants in human serum using a cost-effective PCA-assisted SERS platform. *ACS Applied Materials & Interfaces*, *17*(46), 63237–63252. https://doi.org/10.1021/acsami.5c18171

[35] Rippa, M., Milano, A., Marchesano, V., Sagnelli, D., Guilcapi, B., D'Avino, A., . . . , & Petti, L. (2025). Diagnostic oriented discrimination of different Shiga toxins via PCA-assisted SERS-based plasmonic metasurface. *Nanophotonics*, *14*(23), 4005–4018. https://doi.org/10.1515/nanoph-2024-0696

[36] Hamdy, A. M., AbdelMageed, R. I., Mohammed, B. A., Mohammedy, D. F., Fahmy, B. S., & Mahmoud, A. G. (2024). Environmental risk factors for relapses in pediatric onset inflammatory bowel disease. *QJM: An International Journal of Medicine*, *117*(Supplement_2), hcae175.731. https://doi.org/10.1093/qjmed/hcae175.731

[37] Rana, Md. L., Ullah, Md. A., Hoque, M. N., Hassan, J., Siddique, M. P., & Rahman, Md. T. (2025). Preliminary survey of biofilm forming, antibiotic resistant Escherichia coli in fishes from land based aquaculture systems and open water bodies in Bangladesh. *Scientific Reports*, *15*(1), 7811. https://doi.org/10.1038/s41598-024-80536-6

[38] Bitew, G., Dagnew, M., Dereje, M., Birhanu, A., Gashaw, Y., Ambachew, A., & Tessema, B. (2025). Burden of multidrug resistant bacterial isolates and its associated risk factors among UTI-confirmed geriatrics in Gondar town. *Scientific Reports*, *15*(1), 14270. https://doi.org/10.1038/s41598-025-89123-9

[39] Yang, P., & Wang, X. (2020). COVID-19: A new challenge for human beings. *Cellular & Molecular Immunology*, *17*(5), 555–557. https://doi.org/10.1038/s41423-020-0407-x

[40] Veiga, A. B. G. D., Martins, L. G., Riediger, I., Mazetto, A., Debur, M. D. C., & Gregianini, T. S. (2021). More than just a common cold: Endemic coronaviruses OC43, HKU1, NL63, and 229E associated with severe acute respiratory infection and fatality cases among healthy adults. *Journal of Medical Virology*, *93*(2), 1002–1007. https://doi.org/10.1002/jmv.26362

[41] Kim, M. I., & Lee, C. (2023). Human coronavirus OC43 as a low-risk model to study covid-19. *Viruses*, *15*(2), 578. https://doi.org/10.3390/v15020578

[42] Etcheverría, A. I., & Padola, N. L. (2013). Shiga toxin-producing Escherichia coli: Factors involved in virulence and cattle colonization. *Virulence*, *4*(5), 366–372. https://doi.org/10.4161/viru.24642

[43] Byrne, L., Adams, N., & Jenkins, C. (2020). Association between Shiga toxin–producing Escherichia coli O157: H7 stx gene subtype and disease severity, England, 2009–2019. *Emerging Infectious Diseases*, *26*(10), 2394–2400. https://doi.org/10.3201/eid2610.200319

[44] Gill, A., Dussault, F., McMahon, T., Petronella, N., Wang, X., Cebelinski, E., . . . , & Carrillo, C. (2022). Characterization of atypical Shiga toxin gene sequences and description of stx2j, a new subtype. *Journal of Clinical Microbiology*, *60*(3), e02229–21. https://doi.org/10.1128/jcm.02229-21

[45] Grzelczak, M., Pérez-Juste, J., Mulvaney, P., & Liz-Marzán, L. M. (2008). Shape control in gold nanoparticle synthesis. *Chemical Society Reviews*, *37*(9), 1783. https://doi.org/10.1039/b711490g

[46] Sharma, S., Chaurasia, S., Dinday, S., Srivastava, G., Singh, A., Chanotiya, C. S., & Ghosh, S. (2024). High-level biosynthesis of enantiopure germacrene D in yeast. *Applied Microbiology and Biotechnology*, *108*(1), 50. https://doi.org/10.1007/s00253-023-12885-7

[47] Lin, G., Zhou, X., & Lijie, L. (2024). Mechanistic understanding of nanoparticle interactions to achieve highly-ordered arrays through self-assembly for sensitive surface-enhanced

Raman scattering detection of trace thiram. *Food Chemistry*, *455*, 139852. https://doi.org/10.1016/j.foodchem.2024.139852

[48] Zhou, H., Wan, F., Jian, Y., Guo, F., Zhang, M., Shi, S., . . . , & Ding, W. (2023). Chitosan/dsRNA polyplex nanoparticles advance environmental RNA interference efficiency through activating clathrin-dependent endocytosis. *International Journal of Biological Macromolecules*, *253*, 127021. https://doi.org/10.1016/j.ijbiomac.2023.127021

[49] Peck, K. A., Lien, J., Su, M., Stacy, A. D., & Guo, T. (2023). Bottom-up then top-down synthesis of gold nanostructures using mesoporous silica-coated gold nanorods. *ACS Omega*, *8*(45), 42667–42677. https://doi.org/10.1021/acsomega.3c05444

[50] Wang, Y., Zhang, Z., Sun, Y., Wu, H., Luo, L., & Song, Y. (2025). Recent advances in surface-enhanced Raman scattering for pathogenic bacteria detection: A review. *Sensors*, *25*(5), 1370. https://doi.org/10.3390/s25051370

[51] Barhen, A., & Daudin, J. J. (1995). Generalization of the Mahalanobis distance in the mixed case. *Journal of Multivariate Analysis*, *53*(2), 332–342. https://doi.org/10.1006/jmva.1995.1040

[52] Vandeginste, B. G., Massart, D. L., Buydens, L. M., De Jong, S., Lewi, P. J., & Smeyers-Verbeke, J. (1998). Supervised pattern recognition. In B. G. Vandeginste, D. L. Massart, L. M. Buydens, S. De Jong, P. J. Lewi, & J. Smeyers-Verbeke (Eds.), *Data handling in science and technology* (*20*, pp. 207–241). Elsevier. https://doi.org/10.1016/S0922-3487(98)80043-9

[53] Olvera Astivia, O. L. (2024). A method to simulate multivariate outliers with known Mahalanobis distances for normal and non-normal data. *Methods in Psychology*, *11*, 100157. https://doi.org/10.1016/j.metip.2024.100157

[54] Martos, G., Muñoz, A., & González, J. (2013). On the generalization of the Mahalanobis distance. In J. Ruiz-Shulcloper & G. Sanniti di Baja (Eds.), *Progress in pattern recognition, image analysis, computer vision, and applications Iberoamerican Congress on Pattern Recognition*, 125–132. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41822-8_16

[55] Srivastava, S., Wang, W., Zhou, W., Jin, M., & Vikesland, P. J. (2024). Machine learning-assisted surface-enhanced Raman spectroscopy detection for environmental applications: A review. *Environmental Science & Technology*, *58*(47), 20830–20848. https://doi.org/10.1021/acs.est.4c06737

[56] Wang, H., Chen, C., Tong, D., Chen, C., Gao, R., Han, H., & Lv, X. (2021). Serum Raman spectroscopy combined with multiple algorithms for diagnosing thyroid dysfunction and chronic renal failure. *Photodiagnosis and Photodynamic Therapy*, *34*, 102241. https://doi.org/10.1016/j.pdpdt.2021.102241

[57] Shanmukh, S., Jones, L., Driskell, J., Zhao, Y., Dluhy, R., & Tripp, R. A. (2006). Rapid and sensitive detection of respiratory virus molecular signatures using a silver nanorod array SERS substrate. *Nano Letters*, *6*(11), 2630–2636. https://doi.org/10.1021/nl061666f

[58] Correa-Duarte, M. A., Pazos Perez, N., Guerrini, L., Giannini, V., & Alvarez-Puebla, R. A. (2015). Boosting the quantitative inorganic surface-enhanced Raman scattering sensing to the limit: The case of nitrite/nitrate detection. *The Journal of Physical Chemistry Letters*, *6*(5), 868–874. https://doi.org/10.1021/acs.jpclett.5b00115

[59] Lellinger, D., Thomson, J., Coca-Lopez, N., Ntziouni, A., Nikoloudakis, N., Fernández-Álvarez, M., . . . , & Diz, E. L. (2025). Interlaboratory study to minimize wavelength calibration uncertainty due to peak fitting of reference material spectra in Raman spectroscopy. *Applied Spectroscopy*, *79*(12), 1669–1679. https://doi.org/10.1177/00037028251330654

[60] Ewing, A. V., & Kazarian, S. G. (2017). Infrared spectroscopy and spectroscopic imaging in forensic science. *The Analyst*, *142*(2), 257–272. https://doi.org/10.1039/C6AN02244H

[61] Movasaghi, Z., Rehman, S., & Rehman, I. U. (2007). Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, *42*(5), 493–541. https://doi.org/10.1080/05704920701551530

[62] Jarvis, R. M., & Goodacre, R. (2004). Discrimination of bacteria using surface-enhanced Raman spectroscopy. *Analytical Chemistry*, *76*(1), 40–47. https://doi.org/10.1021/ac034689c

[63] Zheng, X., Wu, G., Wang, J., Yin, L., & Lv, X. (2022). Rapid detection of hysteromyoma and cervical cancer based on serum surface-enhanced Raman spectroscopy and a support vector machine. *Biomedical Optics Express*, *13*(4), 1912. https://doi.org/10.1364/BOE.448121

[64] Al-Saadi, S., Raman, R. K. S., & Panter, C. (2021). A two-step silane coating incorporated with quaternary ammonium silane for mitigation of microbial corrosion of mild steel. *ACS Omega*, *6*(26), 16913–16923. https://doi.org/10.1021/acsomega.1c01567

[65] Guerrini, L., Sanchez-Cortes, S., Cruz, V. L., Martinez, S., Ristori, S., & Feis, A. (2011). Surface-enhanced Raman spectra of dimethoate and omethoate. *Journal of Raman Spectroscopy*, *42*(5), 980–985. https://doi.org/10.1002/jrs.2823

[66] Wang, X., Fan, L., Wang, S., Zhang, Y., Li, F., Zan, Q., . . . , & Dong, C. (2021). Real-time monitoring mitochondrial viscosity during mitophagy using a mitochondria-immobilized near-infrared aggregation-induced emission probe. *Analytical Chemistry*, *93*(6), 3241–3249. https://doi.org/10.1021/acs.analchem.0c04826

[67] Köck, E.-M., Kogler, M., Bielz, T., Klötzer, B., & Penner, S. (2013). In Situ FT-IR spectroscopic study of CO2 and CO adsorption on Y2 O3, ZrO2, and Yttria-Stabilized ZrO2. *The Journal of Physical Chemistry C*, *117*(34), 17666–17673. https://doi.org/10.1021/jp405625x

[68] Singh, V., Lalitha, K., Maheswari, C. U., Sridharan, V., Pradhan, D., Batra, S., & Nagarajan, S. (2024). Remediation of dyes using supramolecular material derived from carbohydrate based π-gelator using the bottom-up assembly approach. *ACS Omega*, *9*(5), 5695–5704. https://doi.org/10.1021/acsomega.3c08179

[69] López-Lorente, Á. I., & Valcárcel, M. (2016). The third way in analytical nanoscience and nanotechnology: Involvement of nanotools and nanoanalytes in the same analytical process. *TrAC Trends in Analytical Chemistry*, *75*, 1–9. https://doi.org/10.1016/j.trac.2015.06.011

[70] Sa, Y., Feng, X., Lei, C., Yu, Y., Jiang, T., & Wang, Y. (2017). Evaluation of the effectiveness of micro-Raman spectroscopy in monitoring the mineral contents change of human enamel in vitro. *Lasers in Medical Science*, *32*(5), 985–991. https://doi.org/10.1007/s10103-017-2197-7

[71] Bodelón, G., & Pastoriza-Santos, I. (2020). Recent progress in surface-enhanced Raman scattering for the detection of chemical contaminants in water. *Frontiers in Chemistry*, *8*, 478. https://doi.org/10.3389/fchem.2020.00478

[72] Ly, N. H., Son, S. J., Jang, S., Lee, C., Lee, J. I., & Joo, S.-W. (2021). Surface-enhanced Raman sensing of semi-volatile organic compounds by plasmonic

nanostructures. *Nanomaterials*, *11*(10), 2619. https://doi.org/10.3390/nano11102619

[73] Szekeres, G. P., & Kneipp, J. (2019). SERS probing of proteins in gold nanoparticle agglomerates. *Frontiers in Chemistry*, *7*, 30. https://doi.org/10.3389/fchem.2019.00030

How to Cite: Guilcapi, B., Milano, A., Avino, A. D., Sagnelli, D., Rippa, M., Marchesano, V., …, & Petti, L. (2026). Machine Learning-Enhanced SERS on Silicon-Gold Sensors for Ultra-High-Throughput Pathogen Detection. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN62028939