Classification and prediction of *Klebsiella pneumoniae* strains with different MLST allelic profiles *via* SERS spectral analysis

Li-Yan Zhang[1,2,*], Benshun Tian[2,*], Yuan-Hong Huang[1], Bin Gu[2], Pei Ju[3], Yanfei Luo[2], Jiawei Tang[2] and Liang Wang[2]

[1] Laboratory Medicine, Ganzhou Municipal Hospital, Guangdong Provincial People's Hospital Ganzhou Hospital, Ganzhou, Guangdong Province, China
[2] Laboratory Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong Province, China
[3] School of Life Sciences, Xuzhou Medical University, Xuzhou, Jiangsu Province, China
* These authors contributed equally to this work.

ABSTRACT

The Gram-negative non-motile *Klebsiella pneuomoniae* is currently a major cause of hospital-acquired (HA) and community-acquired (CA) infections, leading to great public health concern globally, while rapid identification and accurate tracing of the pathogenic bacterium is essential in facilitating monitoring and controlling of *K. pneumoniae* outbreak and dissemination. Multi-locus sequence typing (MLST) is a commonly used typing approach with low cost that is able to distinguish bacterial isolates based on the allelic profiles of several housekeeping genes, despite low resolution and labor intensity of the method. Core-genome MLST scheme (cgMLST) is recently proposed to sub-type and monitor outbreaks of bacterial strains with high resolution and reliability, which uses hundreds or thousands of genes conserved in all or most members of the species. However, the method is complex and requires whole genome sequencing of bacterial strains with high costs. Therefore, it is urgently needed to develop novel methods with high resolution and low cost for bacterial typing. Surface enhanced Raman spectroscopy (SERS) is a rapid, sensitive and cheap method for bacterial identification. Previous studies confirmed that classification and prediction of bacterial strains *via* SERS spectral analysis correlated well with MLST typing results. However, there is currently no similar comparative analysis in *K. pneumoniae* strains. In this pilot study, 16 *K. pneumoniae* strains with different sequencing typings (STs) were selected and a phylogenetic tree was constructed based on core genome analysis. SERS spectra (N = 45/each strain) were generated for all the *K. pneumoniae* strains, which were then comparatively classified and predicted *via* six representative machine learning (ML) algorithms. According to the results, SERS technique coupled with the ML algorithm support vector machine (SVM) could achieve the highest accuracy (5-Fold Cross Validation = 100%) in terms of differentiating and predicting all the *K. pneumoniae* strains that were consistent to corresponding MLSTs. In sum, we show in this pilot study that the SERS-SVM based method is able to accurately predict *K. pneumoniae* MLST types, which has the application potential in clinical settings for tracing dissemination and controlling outbreak of *K. pneumoniae* in hospitals and communities with low costs and high rapidity.

Submitted 1 June 2023
Accepted 1 September 2023
Published 25 September 2023

Corresponding authors
Jiawei Tang, 15061183455@163.com
Liang Wang, healthscience@foxmail.com

Academic editor
Claus Wilke

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.16161**

Subjects Computational Biology, Microbiology, Computational Science, Data Mining and Machine Learning
Keywords *Klebsiella pneumoniae*, MLST, SERS, Hospital-acquired infection, Transmission

INTRODUCTION

The Gram-negative non-motile bacterium *Klebsiella pneumoniae* was first isolated from a pneumonia patient in 1875 by Edwin Klebs and further characterized by Carl Friedlander in 1882 (*Köhler & Mochmann, 1987*). Although *K. pneumoniae* is an opportunistic pathogen, the bacterium is able to cause infection in multiple sites in human beings such as lungs, bloodstream, and liver, *etc*., leading to pneumonia, sepsis, and liver abscess (*Ballén et al., 2021*; *Heiden et al., 2020*). Due to the increased antibiotic resistance of the pathogen, multidrug resistant *K. pneumoniae* could cause extremely difficult-to-treat infections due to limited therapeutic options (*Liu et al., 2022*). In addition, previous studies show that the bacterium is mainly responsible for hospital-acquired (HA) and community-acquired (CA) infections, primarily among immunocompromised patients, the elderly, and the newborns (*Meng et al., 2019*). Therefore, it is important to rapidly and accurately identify *K. pneumoniae* strains with genotyping methods so as to tracking the transmission routes of the pathogen in hospital and/or community settings, which will facilitate the prevention and control of the bacterial pathogen.

Bacterial genotyping is mainly used in microbiology for epidemiological surveillance, which is important to identify bacterial outbreaks and is able to track the origin and spreading of infectious agents (*Ochoa-Díaz, Daza-Giovannetty & Gómez-Camargo, 2018*). Currently, methods of bacterial genotyping include pulsed field gel electrophoresis (PFGE), multiple-locus variable number tandem repeat analysis (MLVA), multi-locus sequence typing (MLST), core-genome MLST (cgMLST), and core single nucleotide polymorphism (coreSNP), *etc*. (*Gona et al., 2020*; *Zhou et al., 2017*), among which PFGE, core-genome MLST (cgMLST), and core single nucleotide polymorphism (coreSNP) are the three most frequently used methods for the typing of *K. pneumoniae* strains (*Gona et al., 2020*). As a traditional but standard method, MLST was first proposed in 1998 and has been widely applied in characterizing bacterial strains *via* house-keeping genes, through which distinct alleles for each housekeeping gene are assigned and all alleles from all the chosen house-keeping genes are combined together to define the allelic profile that is also known as sequence type (ST) (*Maiden et al., 2013*). MLST only uses a short list of housekeeping genes and the analytical results have low resolution, while cgMLST is a novel molecular typing method with high-resolution that is based on whole genomic sequencing, which has high accuracy in bacterial typing and tracing and is gaining more acceptance in sequence typing analysis (*Yan et al., 2021*). However, these typing methods suffer their own limitations such as complex procedures, high costs, and low discrimination capacity, *etc*., which greatly hinders their practical use in real-world settings like clinical laboratory diagnosis. Therefore, novel methods are urgently needed to type and track the transmission of bacterial pathogens rapidly and accurately.

Surface enhanced Raman spectroscopy (SERS) is a highly sensitive and non-invasive method that has been employed for identifying bacterial species, antibiotic resistance, and

virulence phenotypes through the combination of computational methods such as clustering algorithms and machine learning methods (*Liu et al., 2022*; *Lyu et al., 2023*; *Usman et al., 2022*; *Wang et al., 2021*). Therefore, the technique holds the potential in discriminating bacterial strains with different sequence typing. A previous study has already showed that SERS spectral analysis had advantages over traditional genotyping methods for epidemiological surveillance of bacterial infections in terms of rapidity, automation and reliability (*Lu et al., 2013*). A pilot study also confirmed that the label-free SERS technique could identify antibiotic resistant isolates of three MLST-predefined living *Escherichia coli* groups (*Cheong et al., 2017*). However, there is, so far, limited studies focus on using label-free SERS technique for bacterial molecular typing in *Klebsiella* genus. A couple of studies provided preliminary but conflict results when comparing Raman spectroscopy with molecular typing for bacterial pathogens from the genus *Klebsiella*, which suggests that further studies are needed to compare the two methods (*Dieckmann et al., 2016*; *Overdevest et al., 2014*).

In order to elucidate the capacity of SERS technique in discriminating and predicting *K. pneumoniae* strains with different STs, we selected 16 *Klebsiella pneumoniae* strains with distinct STs that were isolated from clinical samples. SERS technique was then applied to these *K. pneumoniae* strains to generate average SERS spectrum for each ST type. Classification analysis *via* Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA) showed that SERS spectra belonging to different *K. pneumoniae* strains could cluster into different groups, while machine learning analysis confirmed the support vector machine (SVM) algorithm can achieve accurate prediction of *K. pneumoniae* strains with different STs, which is consistent with MLST analysis. Taken together, in this pilot study, we show that the SERS-SVM based method is able to accurately recognize *K. pneumoniae* MLST types for the first time, which has the application potential in clinical settings for tracing dissemination and controlling outbreak of *K. pneumoniae* strains in hospitals and communities with low costs, short time and high accuracy (Fig 1).

METHODS AND MATERIALS

Collection of *K. pneumoniae* strains

K. pneumoniae strains were obtained from the Clinical Microbiology Laboratory at Guangdong Provincial People's Hospital, Guangzhou, Guangdong Province, China. All the bacterial strains were grown overnight in commercial Luria Bertani (LB) liquid medium to the exponential growth phase and bacterial cells were collected by centrifugation at 4,500 rpm for 8 min followed by keeping the pellet and discarding the supernatant. The pellet was re-suspended in 2 mL of distilled deionized water (ddH_2O). Bacterial concentration was determined by plate-counting test performed on blood agar plates incubated at 37 °C for 24 h wherever needed. All experiments in contact with bacteria were sterilized *via* autoclaving at 121 °C for 30 min.

Whole genome sequencing of *K. pneumoniae* strains

The Illumina MiSeq Instrument was used for paired-end (PE) sequencing, and the sequencing mode was set to be PE300. All the reads obtained by sequencing were evaluated

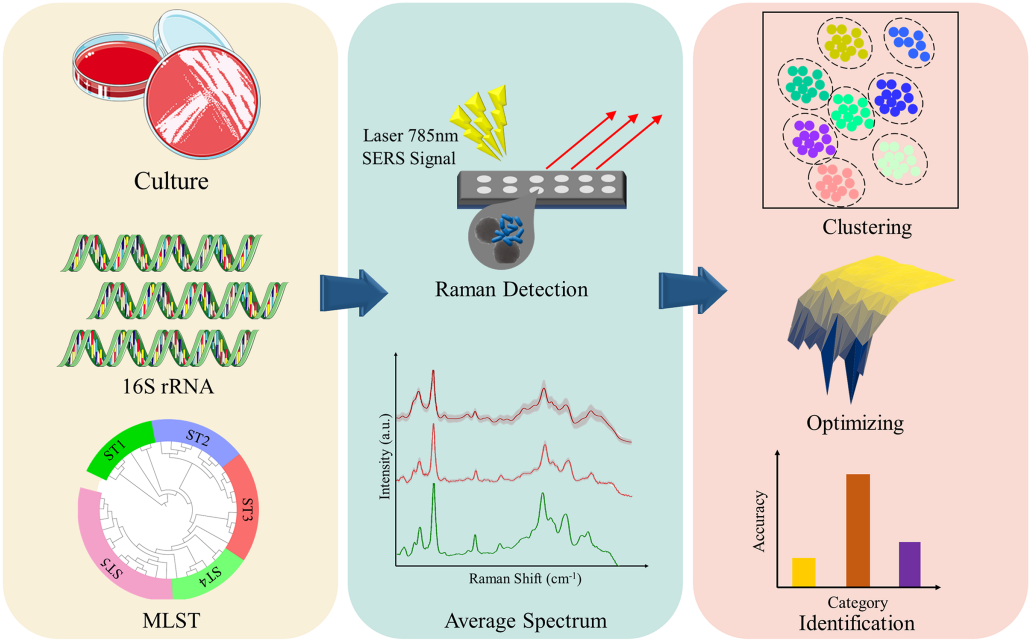

Figure 1 **Schematic illustration of the experimental and computational workflow of this study, which involves bacterial culture, genome sequencing, sequence typing, SERS spectral collection, and computational analysis of SERS spectra generated from *K. pneumoniae* strains with different STs.** Full-size DOI: 10.7717/peerj.16161/fig-1

via FastQC Software (version: 0.11.9; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for sequencing quality control, and then the software Trimmomatic (version: 0.39) was used to remove sequences with a probability of higher than 1% of wrong bases, sequences rich in adapters and sequences with too many N bases (*Bolger, Lohse & Usadel, 2014*). Finally, the SPAdes software (version: 3.1.2) was used for reads assembly, and all the contigs less than 200 bp were removed, and finally assembled whole genomes of *K. pneumoniae* strains with high-quality were obtained for further analysis (*Bankevich et al., 2012*). All the 16 *K. pneumoniae* genome assembled from raw reads have been submitted to NCBI server (BioProject ID: PRJNA960686).

Core-genome and phylogenetic analyses

The genome annotation software PROKKA (Version: 1.13) was used to annotate the whole genome sequences of 16 *K. pneumoniae* strains (*Seemann, 2014*). Then, Roary (Version: 3.13.0), a software enabling rapid large-scale prokaryotic pan-genome analysis, was used for core genome analysis with a minimum percentage identity of 95%, and a total of 3,364 core genes were generated by analyzing the 16 *K. pneumoniae* strains (*Page et al., 2015*). FastTree (Version: 2.1.12) was then run with "*-nt*" and "*-gtr*" settings to produce a phylogenetic tree in Newick format (*Price, Dehal & Arkin, 2009*). Finally, the Newick tree file produced by FastTree was imported into the webserver interactive Tree of Life (iTOL6) for phylogenetic tree visualization (*Letunic & Bork, 2021*).

Multi-locus sequence typing (MLST)

Based on the whole genome sequencing result, MLST software (version: 2.23.0) was used to scan the overlapping parts of the seven conserved housekeeping genes (*gapA, infB, mdh, pgi, phoE, rpoB, tonB*) within the assembled genome sequences of 16 *K. pneumoniae* strains. Through comparing the seven housekeeping genes in all the *K. pneumoniae* strains, sequence typing was determined. The seven conserved housekeeping genes are provided by the Public Databases for Molecular Typing and Microbial Genome Diversity (PubMLST).

Preparation of silver nanoparticles (AgNPs)

The preparation of silver nanoparticles (AgNPs) was based on a classical and facile chemical reduction method, which was recorded in details in previous studies (*Liu et al., 2023*; *Tang et al., 2022*). All the AgNPs used in this study were synthesized by ourselves in the lab at Guangzhou Provincial People's Hospital, Guangzhou, China. For the characterization of dimensions and morphology of the AgNPs, please refer to our recent publication (*Lyu et al., 2023*).

SERS spectral generation

SERS spectra were collected by using the InVia Reflex Confocal Raman Microscope (Renishaw, Wotton-under-Edge, UK). The Raman spectroscope was equipped with a 785 nm diode laser, achieving a spectral resolution of less 1 cm^{-1}. A bacteria sample (10 µl) was mixed with 10 µl of AgNPs and then incubated for 15 min to make sliver nanoparticles sufficiently interacted with the sample before dropping the mixture on silicon wafer. The wavelength of the instrument was calibrated automatically using an interior silicon wafer plus manual adjustment of external silicon wafer by setting the silicon peak at 520 cm^{-1}. Bacterial samples were excited with a near infrared 785 nm diode laser in a range of 500–1,800 cm^{-1}. The Raman excitation light was focused onto the sample using a 50× objective lens, with a laser power of 150 mW. To ensure the stability and reproducibility of the results, a fixed integration time of 20 s per spectrum was implemented. For each *K. pneumoniae* strain, a total of forty-five spectra were collected under controlled conditions of constant room temperature, guaranteeing the consistency of spectral acquisition for each sample (*Bashir et al., 2021*).

Average SERS spectra and deconvolution analysis

Average intensity of all replicated Raman spectra (N = 45) at each Raman shift was calculated to generate an average SERS spectrum for one ST typing strain of *K. pneumoniae*, and the spectral standard deviation (SD) was calculated and visualized in the average SERS spectrum for indicating the stability of the experimental data. The software Origin (Version 2019b; OriginLab, Northampton, MA, USA) was used to plot average Raman spectra, in which the shaded error band part represented SD. The wider the error band, the worse the reproducibility. Spectral characteristic peaks were generated by using LabSpec6 (HORIBA Scientific, Kyoto, Japan). In specificity, GaussLoren function was used for fitting peaks with parameters set to Level = 15%,

Size = 20. All the identified characteristic peaks were shown in the form of dot plot. Biological meanings of all the characteristic peaks were sourced from literature. Due to the high similarity of the different ST average Raman spectra, in order to explore the differences between different spectra, spectral deconvolution was conducted to process the average Raman spectrum for each ST classification. Specifically, the function of *fit peaks pro* in Origin software was used to fit characteristic peaks, and the function *Vogit* as the convolution form of Gaussian and Lorentzian functions was used to generate deconvolution sub-band for each average SERS spectrum.

SERS spectral clustering analysis

Raman spectral clustering aims to divide spectral dataset into different clusters according to a specific rule. Unsupervised learning algorithms like principal component analysis (PCA) are often used to analyze spectra by calling *PCA* function in sci-kit learn (version 0.21.3) data analysis library (*Ayala et al., 2018*; *Bashir et al., 2021*). In particular, *fit_transform* method was used to fit SERS spectra for different ST typing *K. pneumoniae* strains, and the top two principal components PC1 and PC2 with the largest contribution were selected to describe the overall characteristics of SERS spectra. However, due to the mild differences between SERS spectra of different ST types, the clustering effects were not good due to interfering factors that were not relevant to the grouping information. Therefore, Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) was used to avoid the influence of interference factors in SERS spectral data on the classification results. Specifically, the OPLS-DA function from multivariate statistical analysis software SIMCA (version 13.0, 32 bit) was applied to automatically fit all SERS data from different ST types, which separated the data that were not relevant to the classification information from the data matrix. The results of PCA and OPLS-DA clustering methods were shown in scatter plots, and SERS spectra from *K. pneumoniae* strains with ST types were marked with black dashed circles.

Supervised machine learning analysis of SERS spectra

Division of SERS spectra dataset

To achieve rapid and accurate identification of *K. pneumoniae* strains with different STs, we compared the prediction performance of six machine learning algorithms, that is, Adaptive boosting (AdaBoost), Decision Tree (DT), eXtreme Gradient Boosting (XGB), Quadratic Discriminant Analysis (QDA), Random Forest (RF), and SVM on SERS spectra. Before conducting data analysis, in order to optimize model training efficiency, we utilized the *train_test_split* function to reasonably divide the dataset into training, validation, and test sets with a ratio of 6:2:2. The training set was employed for constructing the machine learning model, while the validation set played a pivotal role in assessing the model's development process and providing the unbiased estimations. The test set, exclusively dedicated to evaluating the performance of the final trained model, remained separated from the model construction process (*Tang et al., 2022*).

Model parameter optimization

In order to determine the optimal performance of the final identification model among different models and within all parameter ranges of the model itself, GridSearch was used to determine the optimal combination of model parameters, all hyperparameter ranges for each model pre-defined in the program (Table S1). Specifically, *GridSearchCV* function was used to optimize the hyperparameter combination, and the *cv* parameter was set to 5, which means that five times of cross validation would be performed. The hyperparameter combination with the highest average score was taken as the best for the final model training. We recorded all the parameter combinations for each model and visualized the gradient model scores.

Model performance evaluation

Quantitative evaluation of model effectiveness is key to determine model performance. In this study, seven evaluation indexes including Accuracy, Precision, Recall, F1-score, fitting time, area under the curve (AUC) and five-fold cross validation were used to evaluate the model performance. For evaluating the predication capacity of machine learning models, there are four main categories: (1) True Positive (TP); (2) False Negative (FN); (3) False Positive (FP); and (4) True Negative (TN). The accuracy score describes the proportion of the predicted correct samples in the total number of samples by calling the *accuracy_score* function. The calculation formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

In order to avoid the unbalance in dataset splitting and due to the unmeasurable of real predictive ability of the model, Precision and Recall were used for evaluation. Precision was calculated by calling *precision_score* (average = 'micro') function to indicate how many of the predicted true samples are true. Recall was calculated by calling *recall_score* (average = 'macro') to indicate how many true values are recognized by the model in the actual dataset. The formula for calculating these two indexes is as follows:

$$Precision = \frac{TP}{TP + FN}, \ Recall = \frac{TP}{TP + FP}$$

Precision and Recall are a pair of mutually restrictive metrics, in order to comprehensively consider the factors of the two metrics, F1-score is used as the weighted harmonic average of Precision and Recall. The *f1_score* (average = 'weighted') function was called to measure the model's ability to find true value. The formula is as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The AUC value is the area under the curve of the operating characteristic curve. Different from the above metrics, this metrics does not depend on the selection of threshold. The larger the area under the curve is, the better the model effect will be. In this study *roc_auc_score* function was used to calculate the value of AUC. The calculation formula is:

$$AUC = \frac{1}{2}\sum_{i=1}^{n-1}(TPR_{i+1} + TPR_i)(FPR_{i+1} - FPR_i)$$

n represents the total number of points on the ROC curve, each point on the curve represents the classification result of a particular classifier, the coordinates of the last point on the ROC curve are denoted as (FPR_n, TPR_n), where FPR_n is the False Positive Rate and TPR_n is the True Positive Rate at that point (*Wang et al., 2022*).

Considering the similar prediction performance of different machine learning models, in order to improve the efficiency of model identification and reduce computing costs, we compared the model fitting time on the same dataset, and *time* function was used to record the start time and end time of the model training. The less the time, the lower the computing resources consumed by the model. The calculation formula is:

$$Times = Time_{start} - Time_{end}$$

RESULTS

Whole genome sequencing and Core-/Pan-genome analysis

Whole genome sequencing

General features of the 16 *K. pneumoniae* genomes are presented in Table 1, which were obtained by integrating genome assembly and annotation results. Genome sizes range from 5.32 to 6.23 Mbps. The number of predicted protein coding sequences (CDSs) in the 16 isolated varied from 4,977 (Strain ID: 2470) to 5,900 (Strain ID: 2497). The overall GC content in these strains ranges from 56.70% to 57.66% and remains relatively consistent among different isolates. All strains have a single tmRNA coding gene. There is a slight variation in the number of ribosome RNA (rRNA) and transfer RNA (tRNA) coding genes among the strains varies, but without significant differences (*Wang et al., 2019*).

Core-/Pan-genome analysis

The total pan-genome analysis for the 16 *K. pneumoniae* strains contain 12,523 coding sequences (CDSs), among which 3,364 (26.86% of total CDSs) were considered as core genes across all 16 strains while 9,159 (73.14% of total CDSs) constituted the accessory fractions, which were unique to each *K. pneumoniae* genome, respectively.

The *K. pneumoniae* strain (strain ID: 2500) has the lowest number of the unique genes (1,522 CDSs) while the *K. pneumoniae* strain (strain ID: 2497) has the highest number of unique genes (2,536 CDSs). A Venn diagram was plotted to show the core-genome and pan-genome analysis for all the studied *K. pneumoniae* strains (Fig. 2A) while a phylogenetics tree based on core-genome sequences were generated and visualized (Fig. 2B). MLST was computationally performed based on whole genome sequences, which were labelled on the exterior circle of the phylogenetic tree, indicating the phylogenetic relationship between sequence types.

Table 1 Basic information of genome sequencing, assembly, and annotation data for the 16 *K. pneumoniae* strains.

Strain ID	Contigs	Bases	CDS	tRNA	rRNA	tmRNA	GC content (%)
1476	202	5,429,699	5,004	83	11	1	57.66
2109	690	5,951,243	5,431	85	11	1	57.47
2130	179	5,634,669	5,255	84	10	1	57.09
2260	98	5,447,909	5,029	85	12	1	57.25
2369	149	5,690,453	5,379	82	12	1	57.01
2424	507	5,745,670	5,254	82	14	1	56.71
2443	245	5,501,232	5,083	84	12	1	57.25
2467	135	5,607,898	5,217	85	16	1	57.05
2468	150	5,534,833	5,100	79	11	1	57.27
2470	57	5,391,662	4,977	79	10	1	57.28
2478	201	5,962,258	5,505	85	12	1	56.67
2481	223	5,921,972	5,531	83	19	1	56.92
2497	222	6,227,838	5,900	89	12	1	56.70
2500	74	5,321,704	4,886	82	13	1	57.45
2501	105	5,364,821	5,021	83	13	1	57.38
2503	226	5,749,244	5,437	80	14	1	57.10

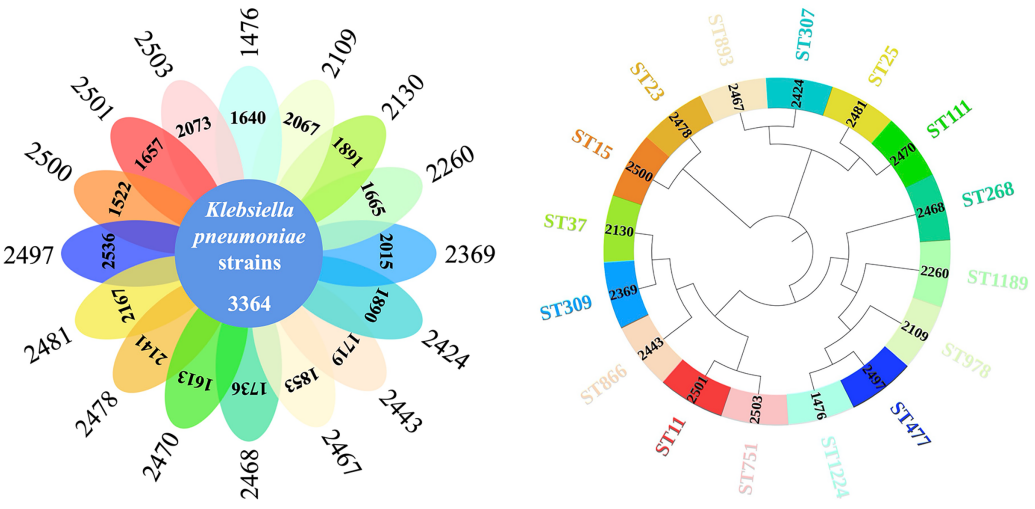

Figure 2 **Core-genome and pan-genome analyses of *K. pneumoniae* strains with different STs.** (A) Venn diagram of shared and unique CDSs among 16 *K. pneumoniae* strains. (B) Phylogenetic tree constructed *via* core genome analysis of 16 *K. pneumoniae* strains. Computational MLST typing results were labelled in the exterior circle of the phylogenetic tree adjacent to each *K. pneumoniae* strain ID in the interior circle, accordingly. Full-size DOI: 10.7717/peerj.16161/fig-2

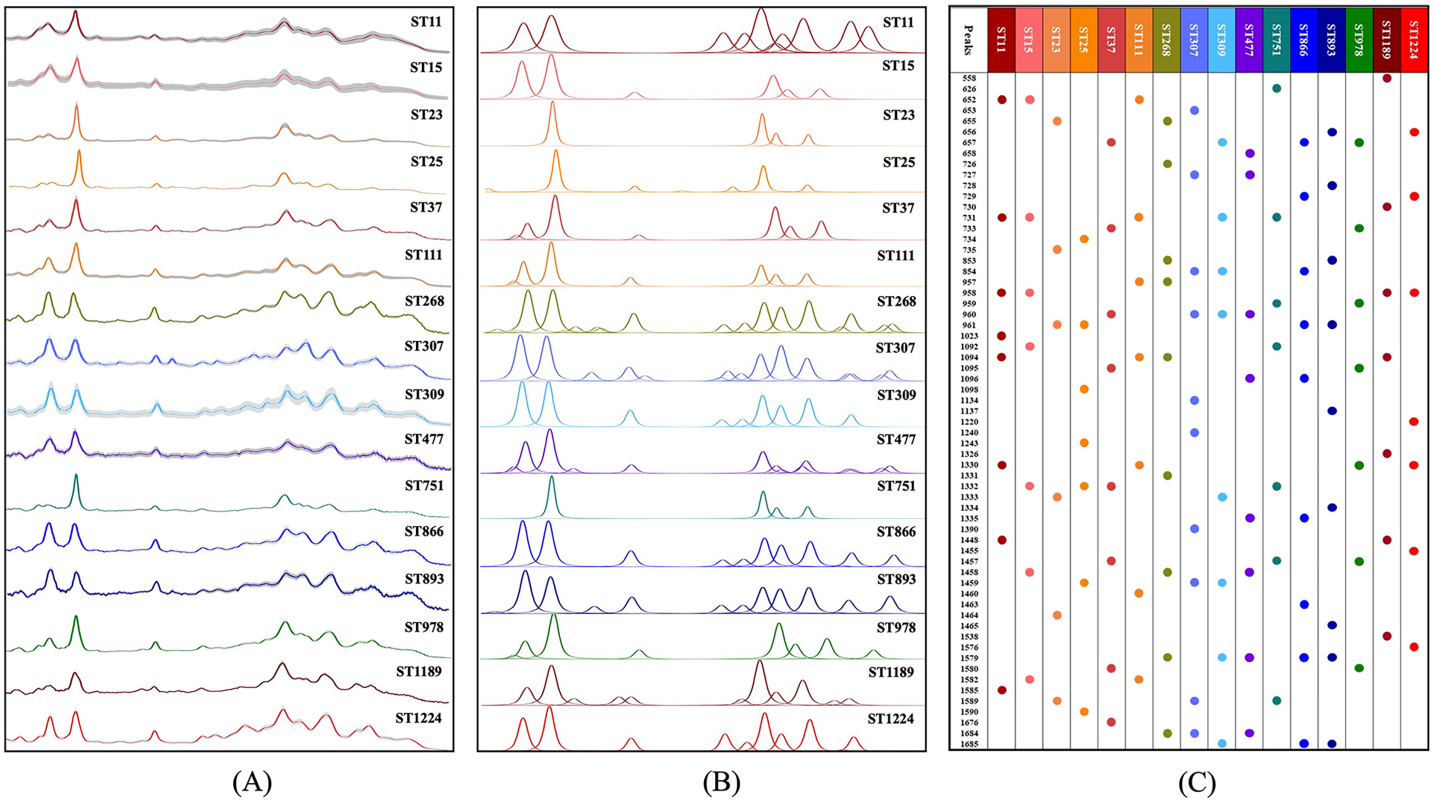

Figure 3 Average and deconvoluted SERS spectra together with characteristic spectral speaks for each *K. pneumoniae* strain with a unique sequence typing. (A) Average SERS spectra of 16 *K. pneumoniae* strains. (B) Deconvoluted SERS spectral bands. X-axis represents Raman shifts in the range of 530–1,800 cm^{-1}, while Y-axis represents the relative Raman intensity. (C) Dot matrix indicating the distribution of characteristic peaks for SERS spectra of *K. pneumoniae* strains. Full-size DOI: 10.7717/peerj.16161/fig-3

Analysis of average and deconvoluted SERS spectra

An average Raman spectrum was used to reflect the overall distribution trend of SERS signal intensities for a single *K. pneumoniae* ST type. We calculated the average signal intensities of specific ST type at each Raman shift to generate the average Raman spectrum (Fig. 3A). The Raman signal standard error of each ST type was calculated to describe the degree of reproducibility of the SERS signal. The narrower the error band, the better the spectral reproducibility. Due to the high similarity of partial average spectra, such as ST268 and ST866, it is difficult to see the difference between the two spectra by naked eyes only. Therefore, spectral deconvolution was conducted to fit each spectral characteristic peaks (Fig. 3B). The distribution and intensity changes of spectral characteristic peaks were comprehensively considered. It can be seen that the SERS spectral deconvolution curves of ST268 and ST866 differ obviously in different Raman shift ranges. For example, in the range of 800–900 cm^{-1} and 1,650–1,750 cm^{-1}, ST268 has four difference deconvolution peaks than ST866 (790 cm^{-1}, 850 cm^{-1}, 1,675 cm^{-1}, and 1,695 cm^{-1}). In addition, we present the fitted spectral peaks in the form of dot matrix (Fig. 3C), and the metabolites corresponding to the peaks of 16 *K. pneumoniae* ST types are sourced from literature and summarized in Table S2.

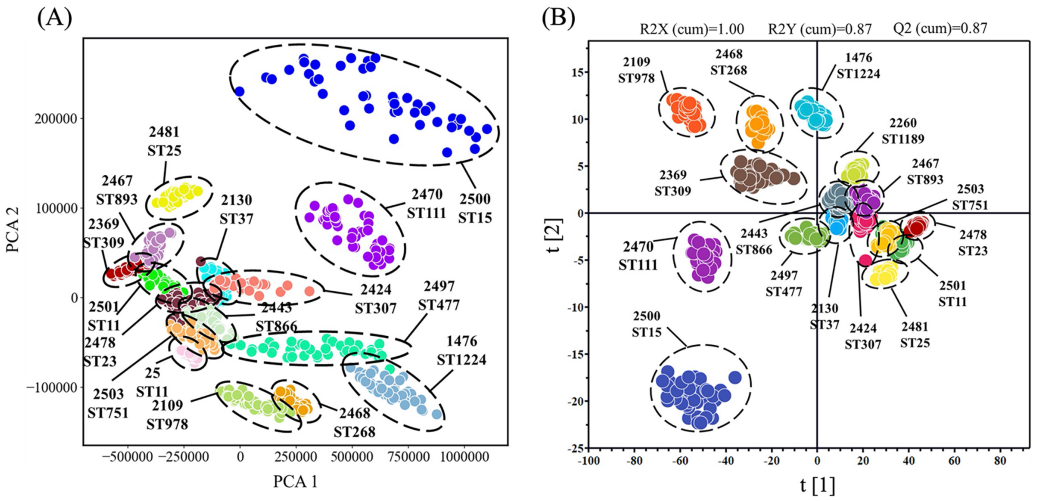

Figure 4 Clustering analysis of SERS spectra of *K. pneumoniae* strains with different STs through the (A) PCA algorithm and the (B) OPLS-DA algorithm. The results were visualized in scatterplot. All the *K. pneumoniae* strains were labelled with both unique Strain ID and sequencing typing.

Full-size DOI: 10.7717/peerj.16161/fig-4

Clustering analysis of K. pneumoniae SERS spectra

To provide better insight into the SERS spectral analysis of *K. pneumoniae* strains with different ST types, we firstly used the unsupervised machine learning algorithm PCA to observe the natural clustering trends between different ST types. The results showed that the clustering of SERS spectra for the same ST type was discrete with overlapping among different ST types. In addition, the PCA method could not quantitatively evaluate the clustering results. Therefore, we used the supervised learning algorithm OPLS-DA as an alternative to analyze the SERS spectral of different ST types of *K. pneumoniae* strains. This method was used to weaken the spectral fluctuation of the same ST classification and maximize the difference among the 16 ST types. According the clustering result as shown in Fig. 4, spectral sample points of different ST types were clustered into different clusters, and the score of three performance evaluation indices were R2X (cum) = 1.00, R2Y (cum) = 0.87 and Q2 (cum) = 0.87, indicating that the OPLS-DA algorithm could better distinguish SERS spectral data of different ST types into separated groups. Through the clustering analysis *via* OPLS-DA, it was also revealed that SERS spectra of *K. pneumoniae* with different ST types were separable *via* computational methods, suggesting the intrinsic spectral differences among these strains. However, clustering analysis cannot provide prediction results. Therefore, when clustering new unknown SERS spectral data, the clustering algorithm needs to re-calculate the distribution of each sample point, which cannot quickly classify new samples.

Machine learning analysis of *K. pneumoniae* SERS spectra

Parameter optimization

Unlike clustering analysis that cannot provide specific labels for clustered samples, supervised machine learning analysis is able to generate prediction results with specific

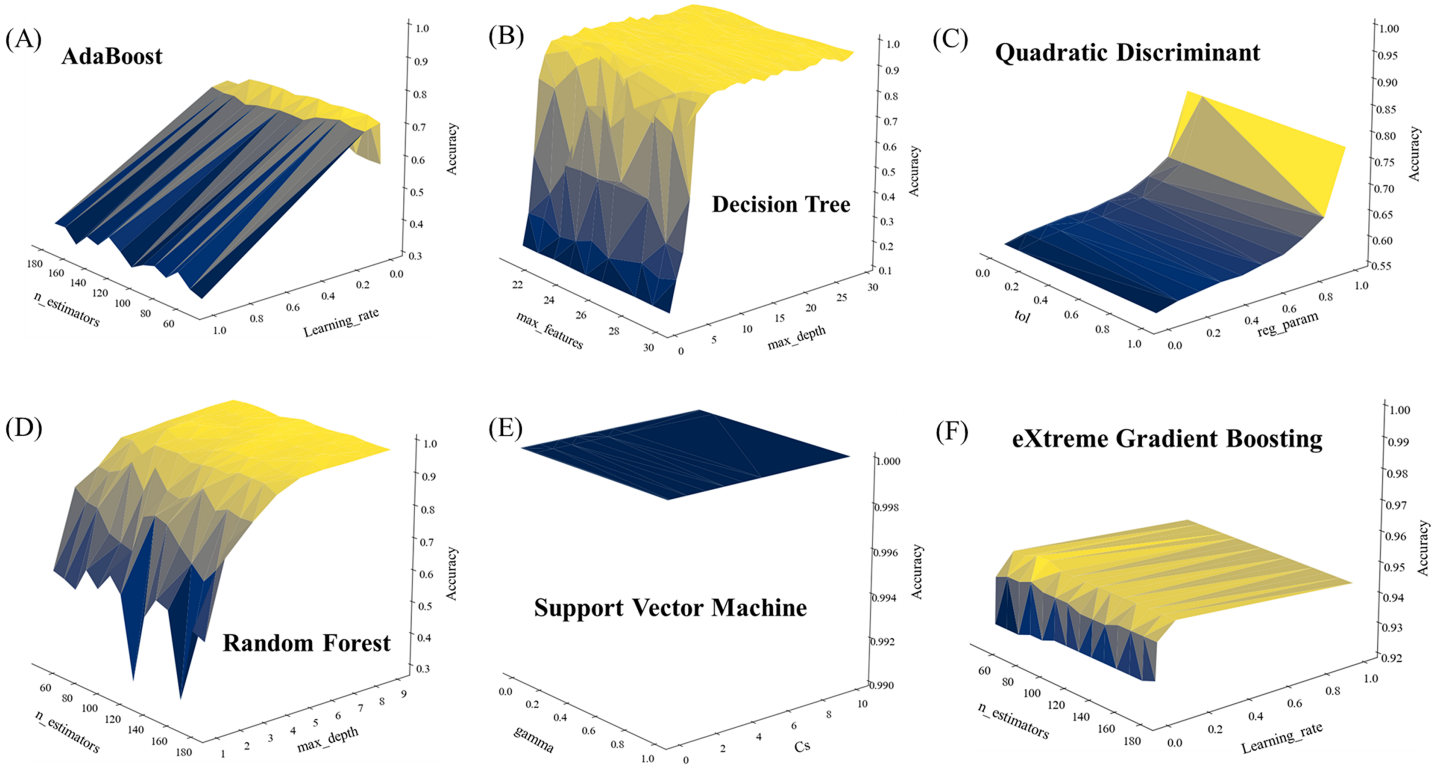

Figure 5 Parameter optimization of six supervised machine learning algorithms used in this study. (A) AdaBoost. (B) Decision tree. (C) Quadratic discriminant. (D) Random forest. (E) Support vector machine. (F) Extreme gradient boosting.

Full-size DOI: 10.7717/peerj.16161/fig-5

labels based on trained models (*Cunningham, Cord & Delany, 2008*). However, different prediction results could be generated with different combinations of hyperparameters, which emphasizes the importance of parameter optimization during modeling training process (*Lameski et al., 2015*). In this study, the *GridSearch* function was used to obtain the best combination of hyperparameters. According to the *GridSearch* gradient plots (Fig. 5), the accuracy score of SVM in all parameter combinations is 1 (Fig. 5E), indicating that SVM algorithm has excellent analytical ability in small samples of high dimensional data, which is consistent with previous studies (*Cheng et al., 2020*). DT, RF and XGB algorithms also show good analytical ability (Figs. 5B, 5D and 5F), all quickly reaching to an accuracy score of more than 0.95 within a few parameter ranges, and remaining stable in the majority of parameter combinations. As for the AdaBoost and QDA algorithms, neither of the two algorithms scored above 0.8 for all parameter combinations, indicating that these algorithms need more computing resources.

Comparison of supervised machine learning algorithms

In this study, we compared the performance of six supervised machine learning algorithms, and explored their ability in identifying ST types by analyzing SERS signal data from 16 *K. pneumoniae*. Seven machine learning evaluation metrics were used to evaluate different models. Computational results were shown in Table 2, according to which the

Table 2 **Performance comparison of six supervised machine learning algorithms on the prediction of *K. pneumoniae* strains with distinct STs based on SERS spectral analysis.**

Algorithm	ACC	Precision	Recall	F1	5-Fold CV	AUC	Time (s)
SVM	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0.50
RF	97.92%	97.92%	98.30%	97.88%	97.04%	98.08%	2.60
DT	96.53%	96.53%	96.16%	96.46%	94.98%	96.96%	0.01
XGB	93.75%	93.75%	94.16%	93.73%	94.30%	93.67%	21.40
QDA	76.39%	76.39%	76.96%	74.47%	76.81%	75.54%	0.09
AdaB	59.03%	59.03%	64.61%	52.56%	70.90%	59.00%	5.38

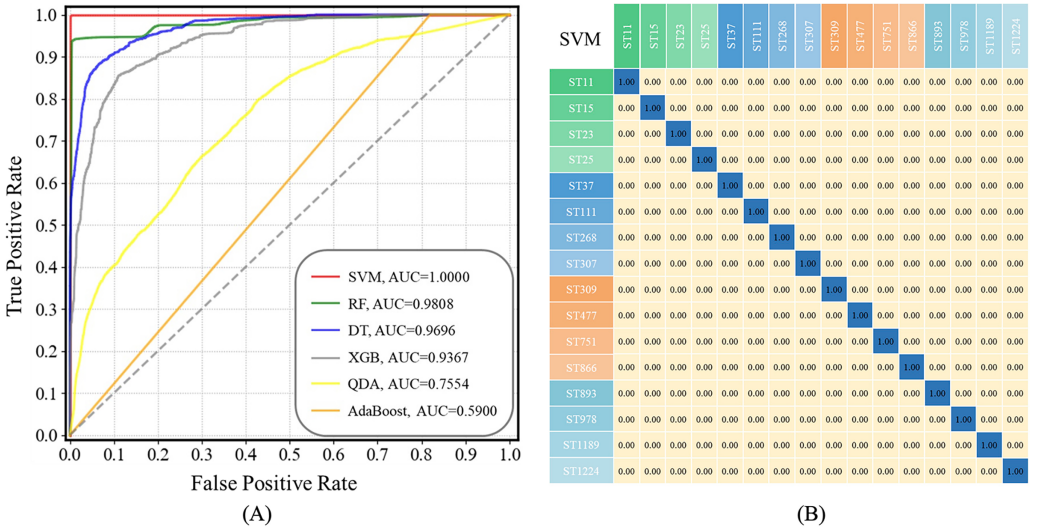

Figure 6 **ROC curves for the six machine learning algorithms and the confusion matrix of SVM algorithm when applied to the SERS spectra of *K. pneumoniae* strains with different STs.** (A) ROC curve. According to the comparison, SVM achieved the best performance with area under curve (AUC) = 1.00 than all other algorithms. (B) Confusion matrix. The percentages in the confusion matrix stand for the correctly classified (diagonal) or mis-classified (off-diagonal) spectra, respectively.
Full-size DOI: 10.7717/peerj.16161/fig-6

SVM model achieves the best performance among all algorithms. All score indices for SVM are 100%, and the training time of the model is relatively short (Times = 0.50 s), indicating that the SVM model can accurately and efficiently identify different ST types. RF, DT and XGB also achieved greater than 90% identification accuracy. The fitting time of XGB model was 21.40 s, which consumed the highest amount of computational resource among all algorithms. It is worth noting that the fitting time of the DT model is only 0.01 s, and the identification accuracy was 96.53%, while the five-fold cross validation score was 94.84% with a slight overfitting of the model, which indicated that DT algorithm was a fast identification method of *K. pneumoniae* ST types. In contract, the QDA and AdaBoost algorithms did not achieve good results in the data of this study. The accuracy value of AdaBoost was only 59.03%, and the five-fold CV score was 70.90%, indicating that the model was underfitting and the parameter range should be further expanded.

ROC curves compare the sensitivity and specificity of supervised machine learning methods across a range of values for their predictive capacities, while AUC means overall accuracies in distinguishing data samples (*Liu et al., 2023*). As for confusion matrix, it is a table summarizing classification results of a supervised machine learning algorithm based on the true class and predicted class (*Liu et al., 2022*; *Wang et al., 2022*). In this study, both ROC curves for all the prediction models and a confusion matrix for the optimal prediction model SVM were present in Fig. 6. The x-axis represents specificity (false positive rate, FPR) and the y-axis represents sensitivity (true positive rate, TPR) in ROC curves. It could be seen that the AUC value of the SVM model is equal to 1.00, suggesting that the SVM model had highest specificity and sensitivity during strain prediction (Fig. 6A). As for the confusion matrix, it showed specific performance of the SVM model on the test dataset, according to which, the identification accuracy of the SVM model for *K. pneumoniae* STs was very high, indicating that the optimized SVM model could accurately recognize *K. pneumoniae* STs with very low error rates based on SERS spectral analysis.

DISCUSSION

K. pneumoniae is a common cause of nosocomial and community-acquired infections (*Dieckmann et al., 2016*). Classification and prediction of *K. pneumoniae* strains is crucial to determine the source and route of contamination. Currently, the classical bacterial typing methods such as lysozyme typing and serotyping are gradually being replaced by molecular biological methods (*Dieckmann et al., 2016*). The strategies of pulsed field gel electrophoresis (PFGE) and multilocus sequence typing (MLST) are contributing to global epidemiological and evolutionary studies (*Gona et al., 2020*). In particular, MLST is an unambiguous procedure for effectively determining bacterial population structure and genealogical assignment based on sequence data of standardized fragments of housekeeping genes (*Mammina et al., 2009*). However, the procedure of data analysis in large-scale studies requires high cost and time-consuming, limiting the use of MLST. Therefore, novel methods which are amenable for achieving rapid and accurate identification of *K. pneumoniae* typing need to be developed. In this study, based on the MLST typing results, we used SERS spectra for verification and analyzed the fingerprints of different ST types, which showed that the chemometric analysis method was able to distinguish closely related *K. pneumoniae* strains with different ST types based on SERS spectra.

Previous studies have shown that the SERS spectra of different bacteria contain all the information of all molecules in the bacteria. For bacteria, different morphological or physiological characteristics have different molecular basis (*Lu et al., 2020*). Therefore, it is reasonable to assume that the average SERS spectra can be generated according to the difference in the distribution of characteristic peaks of different bacterial spectra, and bacteria of different species and genera can be easily and rapidly distinguished (*Wang et al., 2022*). As a proof of concept, we collected the SERS spectra of 16 *K. pneumoniae* strains with unique ST types and measured their SERS spectra. Reproducible SERS spectra (N = 45) were collected for each ST to obtain enough spectra for covering the different

morphological and physiological characteristics, simultaneously the average SERS spectra of the different ST types were calculated to avoid the variability of any single spectrum (Fig. 3A). However, the ST types of *K. pneumoniae* were less related, which made it difficult to distinguish the differences in average spectra between different ST types.

Considering the similarity between the average Raman spectra of multiple ST types, *Pezzotti et al. (2022)* used the linear polynomial expression of the Gauss-Lorentzian function to match the experimental spectra of the minimum scattering based on the average Raman spectra to generate the deconvolution curve of a series of bacteria in the experiment. The experimental results showed that although there were some morphological similarities among seven *Candida auris* significant differences between the deconvoluted spectra could be readily identified. Similarly, we fitted peaks to the SERS spectrum *via* the *Vogit* linear function, which show significant vibrational differences between the different *K. pneumoniae* ST types (Fig. 3B). For example, in the comparison between two bacterial types ST268 and ST866, ST268 has two characteristic peaks at 800–900 cm^{-1}, that is, 790 cm^{-1} (cytosine, uracil) (*Witkowska et al., 2017*) and 850 cm^{-1} (DNA/RNA) (*Bandeliuk et al., 2022*), respectively. while the two characteristic peaks in the range of 1,650–1,750 cm^{-1} are 1,675 cm^{-1} (C=C and C=O stretching vibrations) (*Pezzotti, 2021*) and 1,695 cm^{-1} (-C=CA-stretching) (*Heidari Torkabadi et al., 2014*). The existence of these four spectral deconvolution peaks could be exploited distinguishing ST268 from ST866. Although this method is able to show the "phenotypic difference" among different ST types, it is strongly influenced by the relative intensity of the characteristic peaks of the SERS spectra, which makes it much less useful in practice. Therefore, machine learning algorithms based on advanced statistical methods have been used for subsequent spectral data analysis that greatly improves time efficiency and application potential.

As a supervised clustering analysis algorithm, OPLS-DA is widely applied in the task of distinguishing SERS spectral data (*Liu et al., 2023*; *Tang et al., 2022*). Due to the high dimensionality of SERS data, *Cheng et al. (2022)* found that different leukemia cells could be identified by their intrinsic phenotypic Raman spectra identified *via* the analysis of OPLS-DA algorithm. In order to show the differences between different *K. pneumoniae* ST types and the internal relationship of the same ST type, we used the OPLS-DA algorithm to perform cluster analysis on the spectral data of 16 *K. pneumoniae* ST types (Fig. 4), the result showed that the spectral sample points of different ST types were clustered into different clusters, and the model evaluation indices were R2X = 1.00, R2Y = 0.87, and Q2 = 0.87, indicating that OPLS-DA had a strong ability to distinguish different ST types of SERS spectra. In order to improve the application of machine learning methods in different SERS spectra, and realize the "end-to-end" rapid classification and prediction of spectral data, this study aims to build a spectral identification model suitable for different ST types, and to achieve rapid diagnosis of bacterial typing.

In a clinical diagnostic study on multidrug-resistant *K. pneumoniae*, *Lyu et al. (2023)* collected 121 strains of *K. pneumoniae* with different resistance profiles, which achieved a predictive accuracy of 99.46% by utilizing convolutional neural network (CNN) combined with attention mechanism. This study confirmed the accuracy and feasibility of SERS spectroscopy for distinguishing *K. pneumoniae* with the assistance of machine learning

algorithms. The high sensitivity of the SERS signal and the interference of factors such as the coffee ring effect during sample preparation could lead to large differences between Raman spectra of isotypes (*Koya et al., 2019*), which will further affect the performance of the model. Therefore, in order to attain a classification model that exhibits similar high accuracy and stable performance, the parameter search of the machine learning algorithm is very important. The *GridSearchCV* method originated from scikit-learn library is able to rapidly search for the optimal hyperparameters (*Gao et al., 2022*; *Lei et al., 2022*). In a recent study performed by *Wang et al. (2022)* the *GridSearchCV* method was utilized to optimize the parameters of three machine learning models when analyzing the SERS spectral data of 30 bacteria strains from 9 different genera isolated from clinical samples. In another study of Raman fingerprint of spoilage fungi, *Guo et al. (2021)* used a grid search to optimize the hyperparameters of the model and showed the process with a grid gradient and the optimized values. In this study, for the six different machine learning algorithm models used in this study, we set the parameter range of each model separately, and used the grid search gradient map (Fig. 5) to show the fitting process of each machine learning model. From the model fitting results, it can be found that the SVM algorithm (Fig. 5E) maintains 100% identification accuracy in all parameter combinations, indicating that the SVM model can be well applied to the spectral data analysis of different ST types. In contrast to previous studies, *Ciloglu et al. (2022)* employed a non-linear autoencoder algorithm to extract spectral features when using the SVM algorithm to differentiate colistin-resistant and susceptible strains of *K. pneumoniae*. Their autoencoder-SVM model achieved an accuracy of 94%. However, in the course of this study, it was discovered that the feature extraction process was unnecessary, as the SVM model alone yielded satisfactory results. The best combination of parameters (Table S1) fitted to each model is fed into the algorithm for model training, and the test set samples are used to test the real application performance of each algorithm. Different evaluation metrics are often used to measure the performance of machine learning (*Ma et al., 2023*; *Tang et al., 2022*). In this study, we comprehensively considered the performance of different algorithms in all indicators (Table 2), and found that the SVM algorithm scored 100% in all indicators, and the model fitting time was relatively short. In sum, our results show that SVM is an efficient and stable algorithm suitable for ST typing of *K. pneumoniae*, and has potential application for rapid tracing of the spread and control of *K. pneumoniae* in hospitals and communities.

CONCLUSIONS

K. pneumoniae is a major public health concern worldwide due to its high mortality rate in clinical settings. Rapid and accurate identification and discrimination of different ST types of *K. pneumoniae* is crucial for monitoring and controlling the spread of *K. pneumoniae*. However, the complexity and high cost of traditional methods make efficient and cheap bacterial typing methods urgently needed. This study explored the performance of SERS technology, combining it with multiple advanced machine learning algorithms, for the identification of 16 different ST-typed *K. pneumoniae*. Experimental results show that Raman spectroscopy is sufficient to obtain high-quality bacterial SERS spectra in clinical

laboratories, and that intrinsic differences between different ST typings are revealed by averaging SERS spectra and spectral deconvolution. Through OPLS-DA analysis, it is found that different types of bacterial spectral sample points can be automatically divided into different clusters. Comparing the performance of different machine learning models, the SVM algorithm can accurately classify and predict each type of *K. pneumoniae*, which is consistent with the MLST results. In summary, this study confirms that SERS technology combined with machine learning algorithm can accurately predict the ST types of different *K. pneumoniae*, and has the potential for clinical application with low cost, high speed and high accuracy, and lays the foundation for SERS technology in hospital and community infection detection.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their thoughtful comments that greatly improve the quality of the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant Nos.: 2021A1515220022, 2022A1515220023), the Research Foundation for Advanced Talents of Guandong Provincial People's Hospital (Grant No. KY012023293), the Guangdong Provincial Medical Science and Technology Research Fund (Grant No.: 20201124122643844), and the Ganzhou Science and Technology Bureau Project (Grant No.: GZ2022ZSF252). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Guangdong Basic and Applied Basic Research Foundation: 2021A1515220022 and 2022A1515220023.
Guandong Provincial People's Hospital: KY012023293.
Guangdong Provincial Medical Science and Technology Research Fund: 20201124122643844.
Ganzhou Science and Technology Bureau Project: GZ2022ZSF252.

Competing Interests

Liang Wang is an Academic Editor for PeerJ.

Author Contributions

- Li-Yan Zhang conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Benshun Tian performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Yuan-Hong Huang performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Bin Gu performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Pei Ju performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Yanfei Luo performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Jiawei Tang performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Liang Wang conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The genome sequencing and assembly is available at NCBI: PRJNA960686.

Data Availability

The following information was supplied regarding data availability:

The raw measurements are available in the Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.16161#supplemental-information.

REFERENCES

Ayala OD, Wakeman CA, Pence IJ, Gaddy JA, Slaughter JC, Skaar EP, Mahadevan-Jansen AJ. 2018. Drug-resistant Staphylococcus aureus strains reveal distinct biochemical features with Raman microspectroscopy. *ACS Infectious Diseases* **4(8)**:1197–1210 DOI 10.1021/acsinfecdis.8b00029.

Ballén V, Gabasa Y, Ratia C, Ortega R, Tejero M, Soto S. 2021. Antibiotic resistance and virulence profiles of klebsiella pneumoniae strains isolated from different clinical sources. *Frontiers in Cellular and Infection Microbiology* **11**:1 DOI 10.3389/fcimb.2021.738223.

Bandeliuk O, Assaf A, Bittel M, Durand M-J, Thouand GJ. 2022. Development and automation of a bacterial biosensor to the targeting of the pollutants toxic effects by Portable Raman Spectrometer. *Sensors* **22(12)**:4352 DOI 10.3390/s22124352.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19(5)**:455–477 DOI 10.1089/cmb.2012.0021.

Bashir S, Nawaz H, Majeed MI, Mohsin M, Abdullah S, Ali S, Rashid N, Kashif M, Batool F, Abubakar MJ. 2021. Rapid and sensitive discrimination among carbapenem resistant and susceptible E. coli strains using surface enhanced Raman spectroscopy combined with chemometric tools. *Photodiagnosis and Photodynamic Therapy* **34**:102280 DOI 10.1016/j.pdpdt.2021.102280.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30(15)**:2114–2120 DOI 10.1093/bioinformatics/btu170.

Cheng X, Liang H, Li Q, Wang J, Liu J, Zhang Y, Ru Y, Zhou YJ. 2022. Raman spectroscopy differ leukemic cells from their healthy counterparts and screen biomarkers in acute leukemia. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy* **281**:121558 DOI 10.1016/j.saa.2022.12155.

Cheng H, Xu C, Zhang D, Zhang Z, Liu J, Lv XJ. 2020. Multiclass identification of hepatitis C based on serum Raman spectroscopy. *Photodiagnosis and Photodynamic Therapy* **30**:101735 DOI 10.1016/j.pdpdt.2020.101735.

Cheong Y, Jin Kim Y, Kang H, Choi S, Joo Lee H. 2017. Label-free identification of antibiotic resistant isolates of living Escherichia coli: pilot study. *Microscopy Research and Technique* **80(2)**:177–182 DOI 10.1002/jemt.22785.

Ciloglu FU, Hora M, Gundogdu A, Kahraman M, Tokmakci M, Aydin OJ. 2022. SERS-based sensor with a machine learning based effective feature extraction technique for fast detection of colistin-resistant Klebsiella pneumoniae. *Analytica Chimica Acta* **1221**:340094 DOI 10.1016/j.aca.2022.340094.

Cunningham P, Cord M, Delany SJ. 2008. Supervised learning. In: *Machine Learning Techniques for Multimedia*. Berlin: Springer, 21–49.

Dieckmann R, Hammerl JA, Hahmann H, Wicke A, Kleta S, Dabrowski PW, Nitsche A, Stämmler M, Al Dahouk S, Lasch PJ. 2016. Rapid characterisation of Klebsiella oxytoca isolates from contaminated liquid hand soap using mass spectrometry, FTIR and Raman spectroscopy. *Faraday Discussions* **187**:353–375 DOI 10.1039/C5FD00165J.

Gao W, Zhou L, Liu S, Guan Y, Gao H, Hui BJ. 2022. Machine learning prediction of lignin content in poplar with Raman spectroscopy. *Bioresource Technology* **348**:126812 DOI 10.1016/j.biortech.2022.126812.

Gona F, Comandatore F, Battaglia S, Piazza A, Trovato A, Lorenzin G, Cichero P, Biancardi A, Nizzero P, Moro MJ. 2020. Comparison of core-genome MLST, coreSNP and PFGE methods for Klebsiella pneumoniae cluster analysis. *Microbial Genomics* **6(4)**:e000347 DOI 10.1099/mgen.0.000347.

Guo Z, Wang M, Barimah AO, Chen Q, Li H, Shi J, El-Seedi HR, Zou XJ. 2021. Label-free surface enhanced Raman scattering spectroscopy for discrimination and detection of dominant apple spoilage fungus. *International Journal of Food Microbiology* **338**:108990 DOI 10.1016/j.ijfoodmicro.2020.108990.

Heidari Torkabadi H, Bethel CR, Papp-Wallace KM, De Boer PA, Bonomo RA, Carey PR. 2014. Following drug uptake and reactions inside Escherichia coli cells by Raman microspectroscopy. *Biochemistry* **53**:4113–4121 DOI 10.1021/bi500529c.

Heiden SE, Hübner N-O, Bohnert JA, Heidecke C-D, Kramer A, Balau V, Gierer W, Schaefer S, Eckmanns T, Gatermann S, Eger E, Guenther S, Becker K, Schaufler K. 2020. A Klebsiella pneumoniae ST307 outbreak clone from Germany demonstrates features of extensive drug resistance, hypermucoviscosity, and enhanced iron acquisition. *Genome Medicine* **12(1)**:589 DOI 10.1186/s13073-020-00814-6.

Koya SK, Yurgelevic S, Brusatori M, Huang C, Diebel LN, Auner GWJ. 2019. Rapid detection of Clostridium difficile toxins in stool by Raman spectroscopy. *The Journal of Surgical* **244**:111–116 DOI 10.1016/j.jss.2019.06.039.

Köhler W, Mochmann HJ. 1987. Carl Friedländer (1847–1887) and the discovery of the Pneumococcus–in memory of the centenary of his death. *Zeitschrift für Ärztliche Fortbildung* **81**:615–618.

Lameski P, Zdravevski E, Mingov R, Kulakov A. 2015. SVM parameter tuning with grid search and its impact on reduction of model over-fitting. In: *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Cham: Springer, 464–474 DOI 10.1007/978-3-319-25783-9_41.

Lei B, Bissonnette JR, Hogan ÚE, Bec AE, Feng X, Smith RDJ. 2022. Customizable machine-learning models for rapid microplastic identification using Raman microscopy. *Analytical Chemistry* **94**:17011–17019 DOI 10.1021/acs.analchem.2c02451.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49(W1)**:W293–W296 DOI 10.1093/nar/gkab301.

Liu W, Tang J-W, Lyu J-W, Wang J-J, Pan Y-C, Shi X-Y, Liu Q-H, Zhang X, Gu B, Wang L, Carroll KC, Rebrošová K. 2022. Discrimination between carbapenem-resistant and carbapenem-sensitive Klebsiella pneumoniae Strains through computational analysis of surface-enhanced Raman spectra: a pilot study. *Microbiology Spectrum* **10(1)**:277 DOI 10.1128/spectrum.02409-21.

Liu W, Tang J-W, Mou J-Y, Lyu J-W, Di Y-W, Liao Y-L, Luo Y-F, Li Z-K, Wu X, Wang LJ. 2023. Rapid discrimination of Shigella spp. and Escherichia coli via label-free surface enhanced Raman spectroscopy coupled with machine learning algorithms. *Frontiers in Microbiology* **14**:1101357 DOI 10.3389/fmicb.2023.1101357.

Lu W, Chen X, Wang L, Li H, Fu YVJ. 2020. Combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification. *Analytical Chemistry* **92**:6288–6296 DOI 10.1021/acs.analchem.9b04946.

Lu X, Samuelson DR, Xu Y, Zhang H, Wang S, Rasco BA, Xu J, Konkel ME. 2013. Detecting and tracking nosocomial methicillin-resistant Staphylococcus aureus using a microfluidic SERS biosensor. *Analytical Chemistry* **85(4)**:2320–2327 DOI 10.1021/ac303279u.

Lyu J-W, Zhang XD, Tang J-W, Zhao Y-H, Liu S-L, Zhao Y, Zhang N, Wang D, Ye L, Chen X-LJ. 2023. Rapid prediction of multidrug-resistant klebsiella pneumoniae through deep learning analysis of sers spectra. *Microbiology Spectrum* **11**:e0412622 DOI 10.1128/spectrum.04126-22.

Ma Z-W, Tang J-W, Liu Q-H, Mou J-Y, Qiao R, Du Y, Wu C-Y, Tang D-Q, Wang LJ. 2023. Identification of geographic origins of Morus alba Linn. *Journal of Biomolecular Structure & Dynamics* Epub ahead of print 20 February 2023 1–14 DOI 10.1080/07391102.2023.2180433.

Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* **11(10)**:728–736 DOI 10.1038/nrmicro3093.

Mammina C, Aleo A, Romani C, Pellissier N, Nicoletti P, Pecile P, Nastasi A, Pontello MMJ. 2009. Characterization of Listeria monocytogenes isolates from human listeriosis cases in Italy. *Journal of Clinical Microbiology* **47(9)**:2925–2930 DOI 10.1128/JCM.00102-09.

Meng X, Yang J, Duan J, Liu S, Huang X, Wen X, Huang X, Fu C, Li J, Dou Q, Liu Y, Wang J, Yan Q, Zou M, Liu W, Peng Z, Chen L, Li C, Wu A. 2019. Assessing molecular epidemiology of carbapenem-resistant Klebsiella pneumoniae (CR-KP) with MLST and MALDI-TOF in Central China. *Scientific Reports* **9(1)**:e33 DOI 10.1038/s41598-018-38295-8.

Ochoa-Díaz MM, Daza-Giovannetty S, Gómez-Camargo D. 2018. Bacterial genotyping methods: from the basics to modern. *Host-Pathogen Interactions* **1734**:13–20 DOI 10.1007/978-1-4939-7604-1_2.

Overdevest ITMA, Heck M, van der Zwaluw K, Huijsdens X, van Santen M, Rijnsburger M, Eustace A, Xu L, Hawkey P, Savelkoul P, Vandenbroucke-Grauls C, Willemsen I, van der Ven J, Verhulst C, Kluytmans JAJW. 2014. Extended-spectrum β-lactamase producing

Klebsiella spp. in chicken meat and humans: a comparison of typing methods. *Clinical Microbiology and Infection* **20(3)**:251–255 DOI 10.1111/1469-0691.12277.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31(22)**:3691–3693 DOI 10.1093/bioinformatics/btv421.

Pezzotti GJ. 2021. Raman spectroscopy in cell biology and microbiology. *Journal of Raman Spectroscopy* **52(12)**:2348–2443 DOI 10.1002/jrs.6204.

Pezzotti G, Kobara M, Nakaya T, Imamura H, Fujii T, Miyamoto N, Adachi T, Yamamoto T, Kanamura N, Ohgitani KJ. 2022. Raman metabolomics of Candida auris clades: profiling and barcode identification. *International Journal of Molecular Sciences* **23(19)**:11736 DOI 10.3390/ijms231911736.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26(7)**:1641–1650 DOI 10.1093/molbev/msp077.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30(14)**:2068–2069 DOI 10.1093/bioinformatics/btu153.

Tang J-W, Qiao R, Xiong X-S, Tang B-X, He Y-W, Yang Y-Y, Ju P, Wen P-B, Zhang X, Wang LJ. 2022. Rapid discrimination of glycogen particles originated from different eukaryotic organisms. *International Journal of Biological Macromolecules* **222(Pt A)**:1027–1036 DOI 10.1016/j.ijbiomac.2022.09.233.

Usman M, Tang J-W, Li F, Lai J-X, Liu Q-H, Liu W, Wang L. 2022. Recent advances in surface enhanced Raman spectroscopy for bacterial pathogen identifications. *Journal of Advanced Research* **51**:91–107 DOI 10.1016/j.jare.2022.11.010.

Wang L, Liu W, Tang J-W, Wang J-J, Liu Q-H, Wen P-B, Wang M-M, Pan Y-C, Gu B, Zhang X. 2021. Applications of Raman spectroscopy in bacterial infections: principles, advantages, and shortcomings. *Frontiers in Microbiology* **12**:991 DOI 10.3389/fmicb.2021.683580.

Wang L, Tang J-W, Li F, Usman M, Wu C-Y, Liu Q-H, Kang H-Q, Liu W, Gu BJ. 2022. Identification of bacterial pathogens at genus and species levels through combination of Raman spectrometry and deep-learning algorithms. *Microbiology Spectrum* **10(6)**:e0258022 DOI 10.1128/spectrum.02580-22.

Wang L, Zhu Z, Qian H, Li Y, Chen Y, Ma P, Gu BJ. 2019. Comparative genome analysis of 15 clinical Shigella flexneri strains regarding virulence and antibiotic resistance. *AIMS Microbiology* **5(3)**:205–222 DOI 10.3934/microbiol.2019.3.205.

Witkowska E, Korsak D, Kowalska A, Księżopolska-Gocalska M, Niedziółka-Jönsson J, Roźniecka E, Michałowicz W, Albrycht P, Podrażka M, Hołyst RJ. 2017. Surface-enhanced Raman spectroscopy introduced into the international standard organization (ISO) regulations as an alternative method for detection and identification of pathogens in the food industry. *Analytical and Bioanalytical Chemistry* **409**:1555–1567 DOI 10.1007/s00216-016-0090-z.

Yan S, Zhang W, Li C, Liu X, Zhu L, Chen L, Yang B. 2021. Serotyping, MLST, and core genome MLST analysis of Salmonella enterica from different sources in China during 2004–2019. *Frontiers in Microbiology* **12**:e1002776 DOI 10.3389/fmicb.2021.688614.

Zhou H, Liu W, Qin T, Liu C, Ren H. 2017. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of Klebsiella pneumoniae. *Frontiers in Microbiology* **8**:85 DOI 10.3389/fmicb.2017.00371.