Research Article | Vol. 15, No. 3 / 1 Mar 2024 / Biomedical Optics Express 1878

Biomedical Optics EXPRESS

Rapid and accurate identification of bacteria utilizing laser-induced breakdown spectroscopy

J. H. LIANG,[1,2] S. Q. WANG,[3] W. F. ZHANG,[4] Y. GUO,[4] Y. ZHANG,[5] F. CHEN,[1,2] L. ZHANG,[1,2,*] W. B. YIN,[1,2,6] L. T. XIAO,[1,2] AND S. T. JIA[1,2]

[1]*State Key Laboratory of Quantum Optics and Quantum Optics Devices, Institute of Laser Spectroscopy, Shanxi University, Taiyuan, China*
[2]*Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan, China*
[3]*SINOPEC Research Institute of Petroleum Processing Co., Ltd., Beijing, China*
[4]*Shanxi Xinhua Chemical Defense Equipment Research Institute Co., Ltd., Taiyuan, China*
[5]*School of Optoelectronic Engineering, Xi'an Technological University, Xian, China*
[6]*ywb65@sxu.edu.cn*
[*]*k1226@sxu.edu.cn*

Abstract: Timely and accurate identification of harmful bacterial species in the environment is paramount for preventing the spread of diseases and ensuring food safety. In this study, laser-induced breakdown spectroscopy technology was utilized, combined with four machine learning methods - KNN, PCA-KNN, RF, and SVM, to conduct classification and identification research on 7 different types of bacteria, adhering to various substrate materials. The experimental results showed that despite the nearly identical elemental composition of these bacteria, differences in the intensity of elemental spectral lines provide crucial information for identification of bacteria. Under conditions of high-purity aluminum substrate, the identification rates of the four modeling methods reached 74.91%, 84.05%, 85.36%, and 96.07%, respectively. In contrast, under graphite substrate conditions, the corresponding identification rates reached 96.87%, 98.11%, 98.93%, and 100%. Graphite is found to be more suitable as a substrate material for bacterial classification, attributed to the fact that more characteristic spectral lines are excited in bacteria under graphite substrate conditions. Additionally, the emission spectral lines of graphite itself are relatively scarce, resulting in less interference with other elemental spectral lines of bacteria. Meanwhile, SVM exhibited the highest precision rate and recall rate, reaching up to 1, making it the most effective classification method in this experiment. This study provides a valuable approach for the rapid and accurate identification of bacterial species based on LIBS, as well as substrate selection, enhancing efficient microbial identification capabilities in fields related to social security and military applications.

1. Introduction

With the rapid development of microbiology and deepening research on microbial diversity, precise identification and classification of microorganisms have become crucial. There are various types of microorganisms, such as bacteria, fungi, and algae, which play an indispensable role in maintaining ecosystem function and balance. Moreover, microorganisms have broad application potential in fields such as medicine, environmental science, food industry, and agriculture. The diverse nature of microorganism species presents a distinction between beneficial and harmful ones. Some beneficial microorganisms contribute positively to activities such as food fermentation, biotherapy, and environmental restoration. However, harmful microorganisms like pathogenic bacteria can trigger various diseases, including food poisoning, respiratory infections, and skin infections, while posing potential threats to the environment and ecosystems. Therefore, efficient and accurate identification of microorganism species not only promotes the development

#517213 https://doi.org/10.1364/BOE.517213
Journal © 2024 Received 28 Dec 2023; revised 16 Feb 2024; accepted 20 Feb 2024; published 26 Feb 2024

Biomedical Optics EXPRESS

of life sciences but also helps improve disease prevention and enhance the application value of microorganisms.

Traditional methods for microbial identification, including morphological identification, immune diagnosis, and polymerase chain reaction (PCR), which rely on the study of biological characteristics, biochemical properties, and gene sequences. However, the accuracy and sensitivity of morphological identification is relatively low, while the other two methods require corresponding antibodies or molecular chains, as well as comprehensive sample libraries [1–3]. In addition, these methods involve a long incubation period, precise gene sequencing, or complex biochemical tests, making the detection process quite time-consuming. In recent years, some rapid microbial identification methods have emerged, such as antimicrobial susceptibility testing (AST), multiplex PCR, fluorescent indicators, etc., which offer faster detection speeds. However, their high costs and dependence on consumables have limited their widespread application. Therefore, there is an urgent need to explore a new microbial identification method that is fast, accurate, convenient, and economical.

Laser-induced breakdown spectroscopy (LIBS) is a fast, efficient, multi-element, non-contact atomic and molecular analysis technique [4]. This technology employs a pulsed laser with high peak power focused on the sample surface, causing instantaneous melting, evaporation, gasification, and ionization of the sample in the interaction region, generating transient plasma. Qualitative and quantitative analysis of the elements present in the sample is achieved by processing the spectral fluorescence radiation [5–7]. LIBS has opened up new avenues for microbial identification as it enables direct analysis of cellular components. In recent years, several studies have reported the use of LIBS technology for the detection and identification of biological strains [8]. For example, Farooq et al. demonstrated the potential application of LIBS technology in bacterial identification by extracting and analyzing plasma spectra from various bacteria on a glass substrate [9]. Kim et al. preprocessed LIBS spectra of 5 types of bacteria, including E. coli, utilizing a fingerprint spectral line normalization method and achieved effective differentiation in a two-dimensional plot using two elemental spectral lines [10]. Manzoor et al. identified 40 different types of bacteria, including E. coli, in culture dishes using LIBS combined with a neural network algorithm, achieving an accuracy rate of 95% [11]. Sun et al. constructed a classification model for pathogenic bacteria using support vector machines (SVM) and backpropagation neural networks (BPNN), achieving the identification rate of up to 98% for 5 types of bacteria on silicon chips [12]. Rao et al. achieved the identification rate of over 90% for 10 types of bacteria on filter paper using random forest (RF) combined with principal component analysis (PCA) [13]. Mohaidat et al. achieved the identification rate of 100% for non-pathogenic E. coli and non-toxic derivatives of pathogenic bacteria on agar media under three metabolic conditions using a deterministic finite automaton algorithm model [14]. Marcos et al. achieved the identification rate of 95% for Pseudomonas aeruginosa, E. coli, and Salmonella typhimurium on culture dishes using a neural network algorithm [15]. Furthermore, Wang et al. achieved the identification rate of up to 98% for 6 types of bacteria on glass slides by combining PCA dimensionality reduction methods [16]. In summary, there is a variety of classification models and substrate materials currently used for LIBS identification of bacteria, with varying identification rates.

This article focuses on research involving the LIBS identification of bacteria based on various machine learning models. The main objectives are to explore methods for extracting bacterial spectral information, selecting substrate materials, and evaluating the rate of identification, precision, and recall of each model. The aim is to provide references for the efficient and accurate identification of bacterial species.

Biomedical Optics EXPRESS

2. Experiment

2.1. Bacteria preparation

The experiment selected seven common pathogenic bacteria as identification targets, including Escherichia coli (ATCC25922 E. coli), Enterococcus faecalis (ATCC29212 E. faecalis), Bacillus megatherium (CGMCC1.217 B. megatherium), Bacillus subtilis (ATCC 6633 B. subtilis), Bacillus thuringiensis (ATCC10792 B. thuringiensis), Pseudomonas aeruginosa (ATCC9027 P. aeruginosa), and Pseudomonas fluorescens (ATCC13524 P. fluorescens), numbered 1 to 7. In addition, high-purity aluminum (99.99%) and graphite were chosen as the substrate materials, each with dimensions of 50 mm × 50 mm.

The preparation process for the bacteria is shown in Fig. 1. First, the inoculation rings and substrates are sterilized at 121 °C for 15 minutes. Then, well-isolated colonies from slant agar are subcultured in liquid medium for enrichment, with an enrichment time of 12 hours. Subsequently, the harvested bacteria are dissolved in deionized water to prepare a 3 ml bacterial solution, which is then centrifuged at 10000 r/min. The supernatant is removed, and 1 ml of the precipitate is evenly spread on the surface of the substrate using a pipette. Finally, it is left to dry naturally, forming a 30 mm × 30 mm bacterial film on the substrate surface. It is worth noting that all the above operations are carried out in a sterile ultra-clean bench to ensure the aseptic conditions of the experiment.

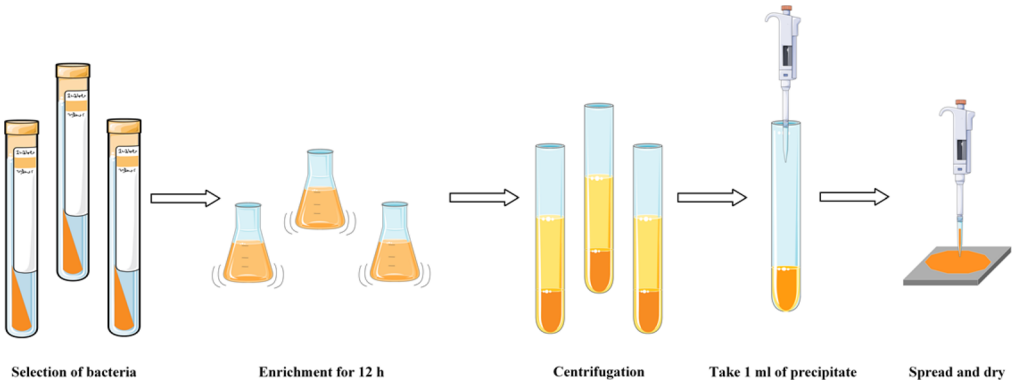

Fig. 1. Flowchart of bacterial preparation.

2.2. Experimental setup

The constructed experimental device for LIBS identification of bacteria is shown in Fig. 2. The high-power pulsed laser is split into two beams after passing through a λ/2 waveplate, beam expander and polarizing beam splitter (PBS). One beam enters the power meter to monitor the laser power in real time, while the other beam is reflected by a 45° mirror and focused on the sample surface after passing through a plano-concave lens (f = 75 mm). At this point, the fiber collects the plasma fluorescence and directs it into the spectrometer (AvaSpec-ULS2048L). The Nd: YAG laser used in the experiment is a Quantel Q-smart 450, with an output pulse width of 5 ns, energy of 150 mJ/pulse, repetition rate of 1 Hz, and a central wavelength of 532 nm. The spectrometer has three channels, covering wavelengths from 192 to 880 nm, with a resolution of 0.14 to 0.18 nm. The integration time and delay time are set to 1 ms and 300 µs, respectively. The test sample is placed on an electrically driven two-dimensional displacement stage. Each point on the sample surface is irradiated by 5 laser pulses, and the average spectrum is recorded as one group of spectra. A total of 200 groups of spectra are collected for each sample.

Biomedical Optics EXPRESS

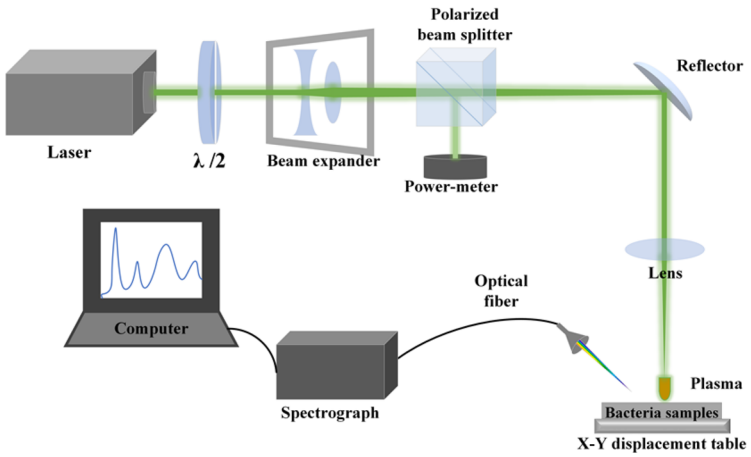

Fig. 2. LIBS experimental setup for identification of bacteria.

3. Identification of bacteria results and discussion

3.1. Spectral pre-processing

Bacteria are composed of both organic and inorganic components. The former including proteins, nucleic acids, polysaccharides, and lipids, while the latter includes water, inorganic salts, and trace elements. Different bacterial species exhibit varying organic and inorganic content and proportions. Furthermore, essential cations like Na^+ and K^+, closely associated with bacterial cell viability, play significant roles in controlling the intra- and extracellular homeostasis. The requirements for osmotic pressure also vary within a certain range among different bacterial species. Consequently, through LIBS analysis of the elemental composition and components of bacteria, classification and identification of bacterial species can be achieved.

The identification of bacteria is accomplished by utilizing the spectral characteristics of their plasma, which primarily depend on the differences in morphology, shape, composition, etc. To highlight these spectral features, it is necessary to preprocess the plasma spectra of bacteria.

Figure 3 demonstrates the comparison of LIBS spectra between the untreated clean substrate and the substrate contaminated with bacteria. From the figure, it is clearly observed that the spectral lines near the ultraviolet range are more easily excited in the aluminum substrate, while the graphite substrate is more easily excited in the visible light range. Taking E. coli as an example, Fig. 3(c) shows a stronger excitation of the substrate spectral lines such as Al II 199.05 nm, Al II 281.62 nm, and Al II 358.66 nm compared to Fig. 3(a) and the spectral lines Mg II 280.27 nm, K I 766.49 nm, and K I 769.90 nm, among others. On the other hand, Fig. 3(d) of the graphite substrate, in comparison to Fig. 3(b), although shows a decreased intensity of the substrate line C I 247.86 nm, exhibits stronger excitation of spectral lines such as Mg II 280.27 nm, Ca I 422.67 nm, Na I 589.00 nm, Ca I 671.78 nm, K I 766.49 nm, and K I 769.90 nm, among others. Comparing Fig. 3(c) and 3(d), it can be concluded that the graphite substrate, although has an influence on the C I 247.86 nm line in bacteria, it is capable of exciting more and stronger bacterial characteristic lines. While the aluminum substrate can excite some bacterial lines, it also generates more interference from aluminum lines. The reason for this difference may be attributed to the fact that the reflectivity of the aluminum substrate ($R_{Al} = 0.91447$) is much higher than that of the graphite substrate ($R_C = 0.31626$). Therefore, the laser energy used for bacterial ablation is smaller, resulting in weaker bacterial spectra.

For each group of 3×4094 LIBS spectra, Savitzky-Golay smoothing filtering is first applied. The core of this method lies in weight-filtering the data within a window, where the weighted

Biomedical Optics EXPRESS

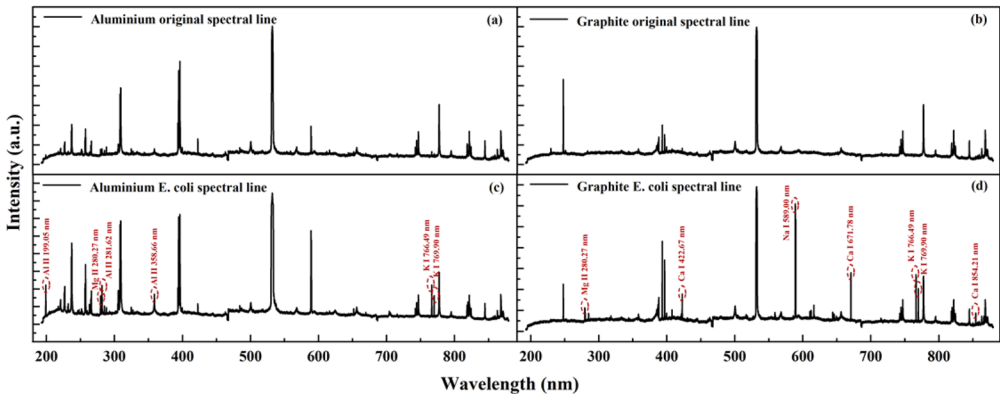

Fig. 3. Comparison of LIBS spectra of basal and bacterial samples.

weights are obtained through the least squares fitting of a given high-order polynomial. This preprocessing is primarily used to address random errors present in LIBS measurements. It not only effectively preserves the information on signal changes, but also improves the signal-to-noise ratio of the spectra. Subsequently, the smoothed spectra are subjected to automatic peak finding and peak line fitting based on predefined heights.

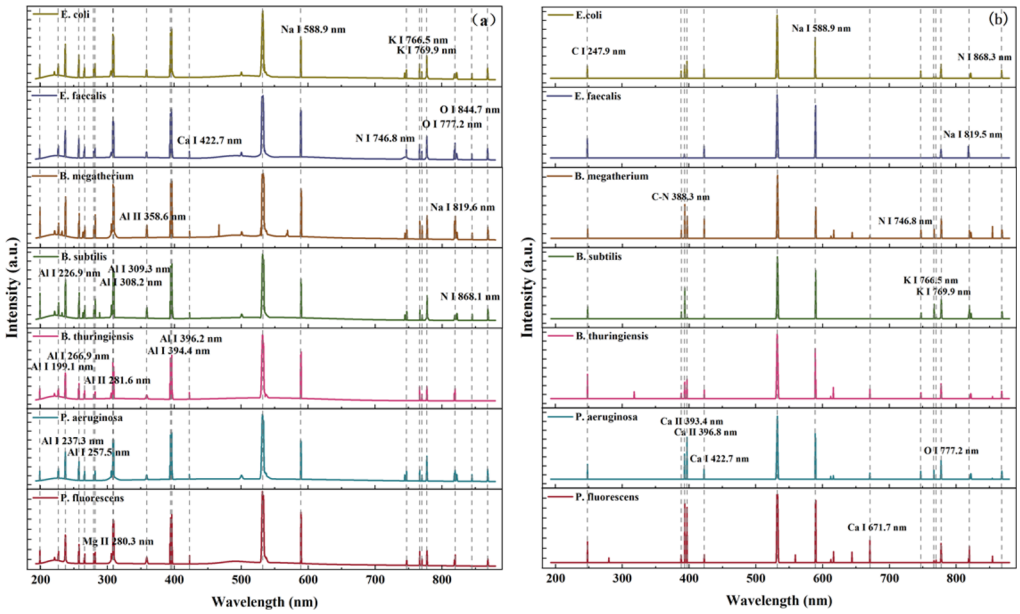

Fig. 4. Pretreatment spectra of seven types of bacteria with (a) aluminum and (b) graphite substrates.

Following the aforementioned preprocessing, the LIBS spectra of various bacteria with graphite and aluminum substrates are shown in Figs. 4(a) and 4(b), respectively. The selected characteristic spectral lines are labeled in the figures and specified in Table 1. For instance, in the case of the K I line at 766 nm, the relative standard deviation (RSD) of characteristic spectral lines under different substrate conditions consistently remains below 10%, which provides a good stability. It can be observed that, compared to the original spectra, the preprocessing significantly reduces

Biomedical Optics EXPRESS

background noise. Bacterial samples mainly contain elements such as C, H, O, N, Na, K, Ca, etc. While the spectral lines of different bacterial species are quite similar, but the intensities differ significantly. For example, in Fig. 4(a), the signal intensities of the three spectral lines Mg II 280.3 nm, K I 766.5 nm, and K I 769.9 nm are noticeably different. In Fig. 4(b), when comparing the spectra of B. thuringiensis and B. subtilis, it can be observed that the former has the lines Ca I 422.7 nm and Ca I 671.7 nm, while the latter has the line K I 766.5 nm. For E. coli and E. faecalis, the former has the lines Ca II 396.8 nm and N I 746.8 nm, while the latter does not. Similarly, comparing the spectra of P. fluorescens and other bacterial species also reveals their differences. Although it is easy to distinguish between bacterial species based on the differences in the characteristic spectral lines, the fluctuation of line intensities due to random factors such as laser energy and environmental influences can easily lead to misclassification. Compared to visual observation, machine learning methods not only more effectively utilize visually significant spectral line features but also excel in identifying hidden relevant features through extensive data-driven learning, achieving faster and more accurate classification. Next, we will try four algorithms—PCA, PCA-KNN, RF, and SVM—to compare and identify the spectra of bacteria with aluminum and graphite substrates.

Table 1. Elemental emission lines observed in LIBS spectra bacteria

No.	Element	Wavelength (nm)
1	C I	247.856
2	N I	818.487
3	O I	777.194
4	Mg II	279.553, 280.271
5	C-N	388.300
6	Na I	588.995, 589.592, 819.482
7	K I	532.328, 533.969, 766.490, 769.896
8	Ca I	422.673, 612.222, 616.217, 643.907, 671.770
	Ca II	317.933, 393.366, 396.847, 854.209

3.2. Identification of bacteria based on the K-nearest neighbor algorithm

Principal Component Analysis, is a widely used statistical method suitable for dimensionality reduction and feature extraction in large multivariate datasets [17–19]. In practical applications, there may be correlations among multiple variables, leading to data redundancy and noise. PCA achieves data compression and feature extraction by linearly transforming the original data into a new feature space. It retains dimensions of features with the highest variance while ignoring those with almost zero variance, thereby improving the efficiency of data utilization.

In this study, PCA is used for dimensionality reduction and feature extraction of bacterial spectral data. Using feature spectra as input for classification reduces noise interference and simplifies data processing, thereby shortening the analysis time. Each feature spectrum is represented in its principal components, and the proportion of retained original information for each spectrum is indicated by the variance of each principal component. Figure 5 displays the scores of the first three principal components obtained from PCA, with a cumulative contribution rate exceeding 90%, indicating that these three principal components can effectively identify the four types of bacteria shown in the figure. However, as the number of bacterial species increases, the score distributions of different species will substantially overlap. Therefore, the classification performance of PCA alone is unsatisfactory, necessitating further integration with other machine learning methods.

Biomedical Optics EXPRESS

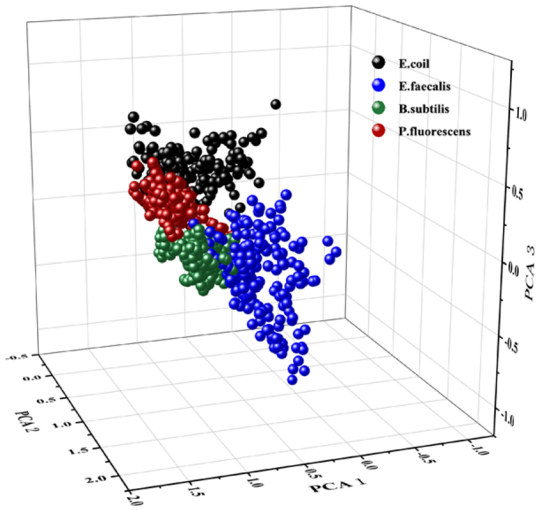

Fig. 5. Classification results of four types of bacteria on graphite substrates using the first three principal components in PCA.

The K-Nearest Neighbor (KNN) algorithm is one of the most commonly used multi-classification algorithms in data mining [20,21]. This algorithm compares the input new prediction set data with the K-nearest values in the known category training set, and the category of the prediction set depends on the majority belonging of the K values. As the KNN classification algorithm does not rely on determining the discriminatory class domain, it performs well in dealing with species issues with a high degree of spectral feature overlap. In the KNN algorithm, the calculation of distance is crucial, and in the experiment, Euclidean distance (the true distance between two points in multidimensional space) is used to calculate the distance between the testing and training samples. The selection of the K value is also crucial, as a smaller K value is susceptible to the influence of outliers, leading to overfitting, while a larger K value makes the decision boundary become blurred.

In the KNN model, 20 and 17 characteristic spectral lines obtained through preprocessing and selection from aluminum and graphite substrates, respectively, are used as input variables for the model. However due to the high time complexity of sample distance calculation in the KNN algorithm, especially when handling large-scale datasets and high-dimensional data, there are difficulties and challenges. Therefore, the experiment attempts to use PCA dimension reduction to optimize the algorithm. Firstly, the data is divided into independent and dependent variables, and then the principal components obtained from PCA dimension reduction are taken as the input of the KNN model. In this classification model, the cumulative contribution rate of the scores of the first 8 principal components corresponding to the aluminum substrate and the first 9 principal components corresponding to the graphite substrate reaches 99%. Therefore, these principal components represent all the information of the original spectrum features. 80% of the total samples are selected as the calibration set, and the remaining samples are used as the testing set. Figure 6 shows the traversal curve of the K value, and the classification model achieves optimal results when the K value is set to 3 and 4, respectively. It is worth noting that, compared to the original spectral features, the classification model after PCA dimension reduction exhibits relatively small fluctuations in accuracy as the K value increases.

Accuracy rate is one of the most direct and commonly used evaluation metrics in classification models, measuring the predictive accuracy of the model on the overall samples. Accuracy rate is defined as the ratio of the number of samples classified correctly (Actual label = Predicted

Biomedical Optics EXPRESS

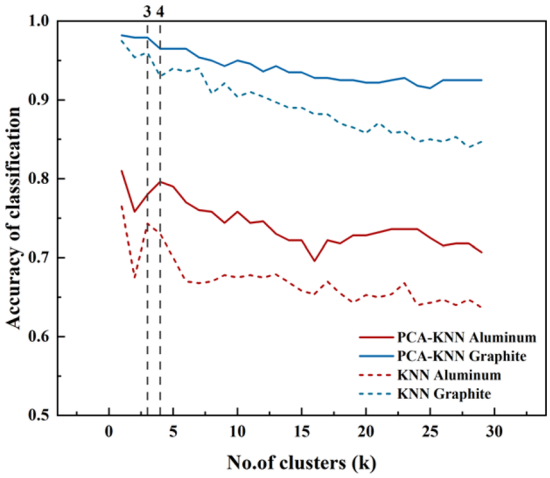

Fig. 6. K value curve for KNN traversal of bacteria.

label) to the total number of samples. When evaluating the KNN and PCA-KNN models, the corresponding accuracy rates for aluminum substrates were found to be 74.91% and 84.05% respectively, while for graphite substrates, they were 96.87% and 98.11% respectively. This indicates that for both models, the identification accuracy is superior when using graphite as the substrate compared to aluminum. The improvement in accuracy may be attributed to the more comprehensive expression of bacterial elements in the spectra of graphite substrates, making graphite substrates more advantageous in LIBS identification of bacteria.

3.3. Identification of bacteria based on the random forest algorithm

Random Forest algorithm is a machine learning model that has strong adaptability and can handle high-dimensional data [22–25]. This algorithm constructs a collection of decision trees (forest) by randomly selecting feature subsets and data subsets, ensuring that each decision tree is independent of the others. The data is then classified and identified based on the judgments made by each decision tree in the forest. After building the forest with the training set, when new samples from the prediction set are inputted, each decision tree in the forest would make its own judgment. Based on the classification results of all decision trees, the most frequently selected category is chosen as the final prediction result.

After preprocessing the bacterial spectra with aluminum and graphite as substrates, 20 and 17 characteristic spectra lines respectively were selected as input variables for the model. The RF learning method was used to identify 7 types of bacteria, where if there is a significant difference or no characteristic spectra line in comparison to the random two types of bacterial samples, the decision tree can directly distinguish them at that node. A total of 1120 randomly selected samples were used as the training set, while the remaining 280 samples were used as the testing set. The classification prediction results are shown in Fig. 7, where different colors and shapes represent different types of bacteria, and solid dots represent misjudgment samples. From the figure, it can be observed that for the samples based on the aluminum substrate, there were 41 misjudgment samples, while for the samples based on the graphite substrate, there were only 3 misjudgment samples, indicating that the RF model has a significantly higher accuracy in identifying bacteria based on the graphite substrate. Multiple random repeated tests were conducted on the samples based on graphite substrate, and the accuracy rate was consistently above 98%. Figure 7(b) shows the testing set randomly selected in one of the tests, where the rate of identification reached 98.93%. Figure 8 displays the variation curve of the accuracy

Biomedical Optics EXPRESS

rate under different decision trees, showing a generally positive correlation between the overall accuracy rate and the number of decision trees. Under the condition of the aluminum substrate, the accuracy rate of bacteria classification generally stabilized after the 4th decision tree.

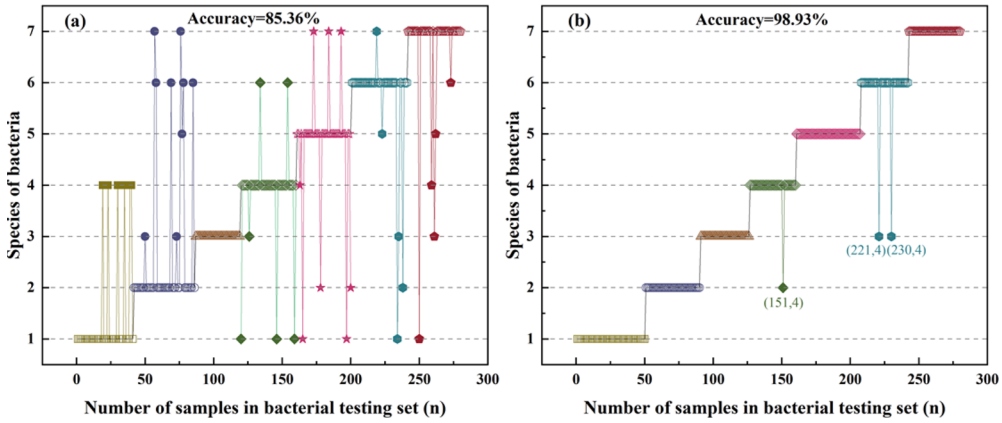

Fig. 7. Identification results of RF classifier for 7 types of bacteria testing set.

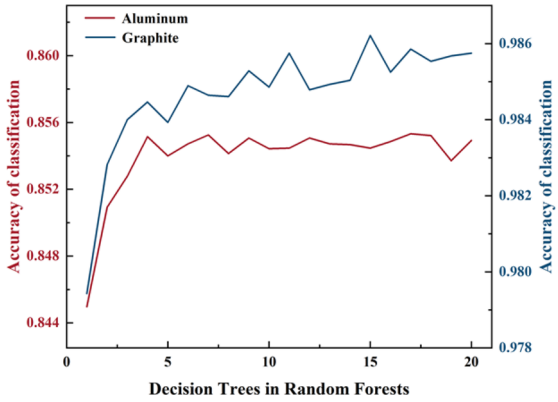

Fig. 8. Decision trees in RF classifier.

3.4. *Identification of bacteria based on the support vector machine algorithm*

Support Vector Machine (SVM) algorithm is a theoretical model based on statistical principles that minimizes structural risk [7,26–28]. The more information (feature spectra) this model uses during training, the better its predictive ability becomes. SVM is capable of classifying non-linearly separable data without increasing complexity. The value of introducing a kernel function lies in its ability to map the features of linearly inseparable examples to higher-dimensional space, depicting high-dimensional classification results in a lower-dimensional space through inner products. Common kernel functions include linear kernel function ($K(x, z) = x \cdot z$), polynomial kernel function ($K(x, z) = (\gamma\, x \cdot z + r)^d$), radial basis kernel function ($K(x, z) = exp(-\gamma\, \|x - z\|^2)$), sigmoid kernel function ($K(x, z) = tanh(\gamma\, x \cdot z + r)$), among others. The radial basis kernel function is commonly used for non-linear classification problems.

For this experiment, each type of substrate was collected with 1400 valid spectra, which were preprocessed and used as input variables for the model. The aluminum substrate had 20 valid

Biomedical Optics EXPRESS

feature spectra, while the graphite substrate had 17. The SVM machine learning method was used for identification of bacteria, where 1120 randomly selected samples were used as the training set and the remaining 280 samples were used as the testing set. The predicted results for the 7 bacterial species are shown in Fig. 9, with different colors and shapes representing different types of bacteria. Solid dots indicate the training set, while hollow dots indicate the testing set. From the figure, it can be observed that there were 13 misclassified bacterial samples in the testing set corresponding to the aluminum substrate. However, after performing repeated tests with multiple random selections of the training set and testing set, the rate of identification for the graphite substrate consistently reached 100%. This indicates that the SVM model built based on the graphite substrate has excellent generalization ability and robustness. Compared to RF and KNN, SVM significantly improves the accuracy of classifying bacterial species with the aluminum substrate, although it is still inferior to the classification performance achieved with the graphite substrate.

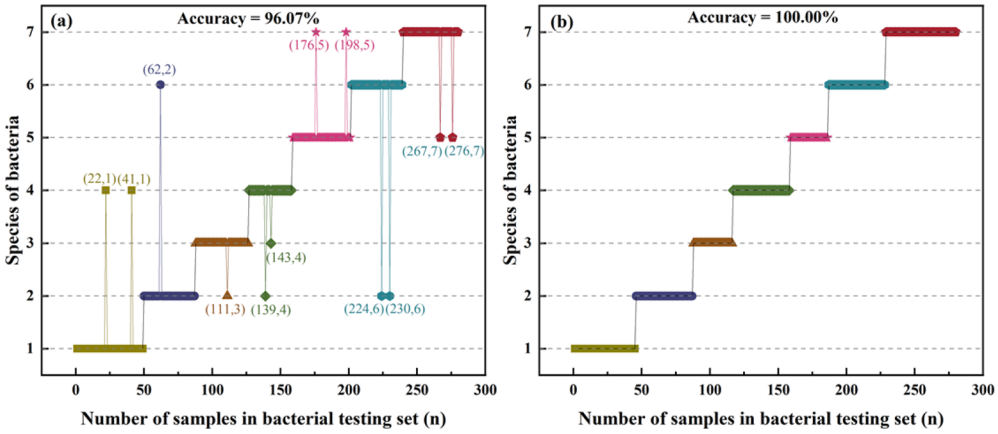

Fig. 9. Identification results of SVM classifiers for 7 types of bacteria testing set.

3.5. *Modelling evaluation*

Classification models, also known as discrete variable prediction models, are typically evaluated using metrics such as Accuracy, Precision, Recall, and F1-score. Accuracy reflects the overall correctness of predictions, including both positive and negative samples, and is used to measure the overall predictive accuracy of classification models. Its goal is to maximize overall predictive accuracy, suitable for scenarios where all classes of samples have equal importance. However, when there is an imbalance in the proportion of different classes in the dataset, accuracy alone may not effectively evaluate the results, thus necessitating independent evaluation for each class.

Precision represents the proportion of samples predicted as a particular class that truly belong to that class, reflecting the accuracy in predicting positive class samples. The goal is to maximize the accuracy of samples predicted as positive class, making it particularly suitable for scenarios where precision in predicting positive classes is of concern. Both accuracy and precision have different definitions, calculation methods, and application goals in the evaluation of classification models.

Recall represents the proportion of samples in a particular class that are correctly predicted by the classifier, out of the total samples in that class. F1-score is the harmonic mean of precision and recall, comprehensively considering the accuracy and recall of the classifier. These metrics are widely used to evaluate the performance of classification models. Precision reflects the model's ability to identify negative samples, with a higher value indicating a stronger ability to

Biomedical Optics EXPRESS

distinguish negative samples. Recall, on the other hand, reflects the model's ability to recognize positive samples, with a higher value indicating a stronger ability to identify positive samples. F1-score considers both of these metrics, with a higher value indicating a more robust model. By denoting samples predicted as positive as TP (True Positive) and those predicted as negative as TN (True Negative), and classifying false negatives (FN) as actual positives mistakenly classified as negatives and false positives (FP) as actual negatives mistakenly classified as positives, these metrics can be calculated using the following formulas:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Table 2 summarizes the bacterial identification rates of four models under two different substrate conditions. Figures 10–12 illustrate the evaluation metrics for different models, with Support denoting the number of samples in the testing set. It should be noted that the RF and SVM, as classification models, require input labels for training, while KNN and PCA-KNN can directly classify input data based on distances without the need for labels.

From the perspective of classification accuracy, it can be observed that the graphite substrate exhibits better overall identification performance compared to the aluminum substrate. This may be due to the fewer spectral lines of graphite, which results in less interference from other elements and better stability and durability, effectively reducing signal distortion or degradation. Although KNN and PCA-KNN classification models are faster as they do not require a training set, their performance is relatively poorer. Dimensionality reduction improves the accuracy of the KNN model, while the RF model achieves higher accuracy than the PCA-KNN model.

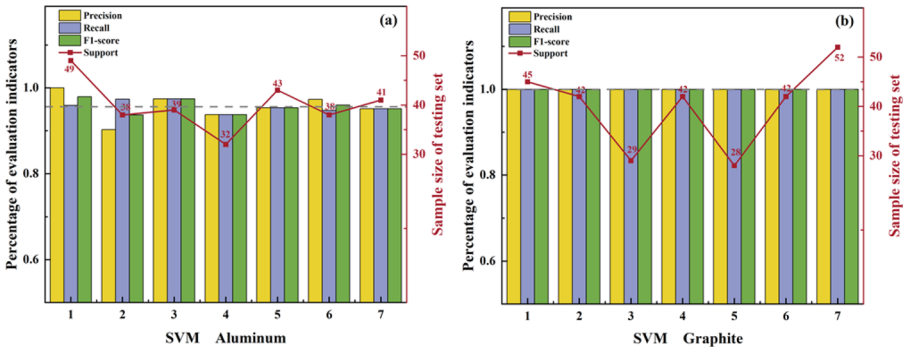

Fig. 10. SVM model evaluation metrics and sample size of testing set.

By observing the evaluation metrics of the four models, for the metal aluminum substrate, the classification performance of class 2 in the SVM model, class 4 in the RF model, class 6 in the KNN model, and classes 5 and 6 in the PCA-KNN model is relatively poor. For the non-metal graphite substrate, the SVM model achieves a rate of identification are 100%, while the RF, KNN, and PCA-KNN models exhibit relatively poor classification results for classes 3 and 6. Overall, the classification evaluation metrics for class 6 are relatively low. It can be seen that different models exhibit varying classification performance for different types of bacteria.

Biomedical Optics EXPRESS

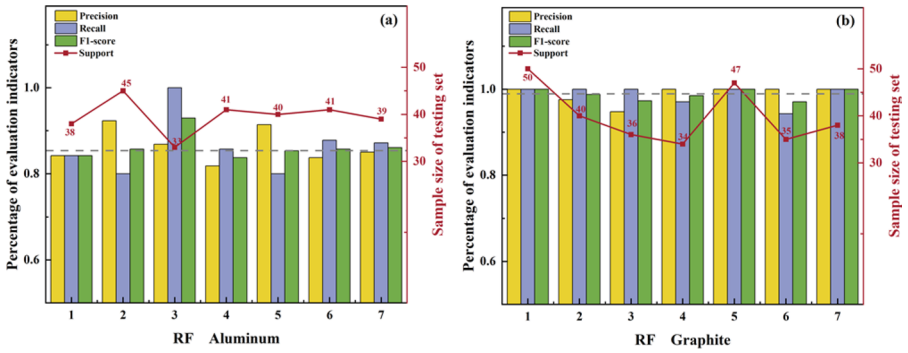

Fig. 11. RF model evaluation index and sample size of testing set.

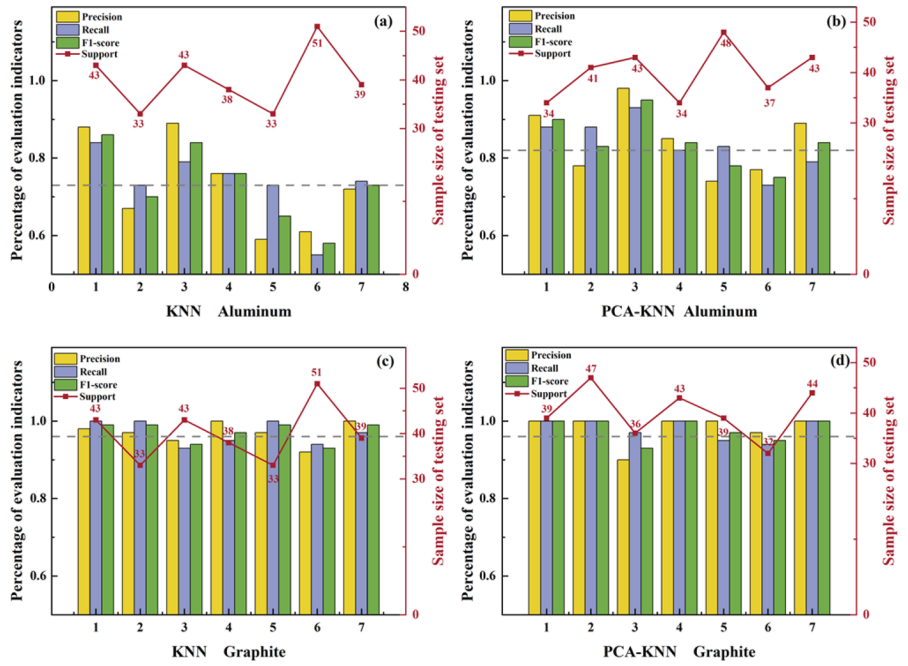

Fig. 12. Evaluation metrics of KNN, PCA-KNN models and sample size of testing set.

Table 2. Number of input variables and identification rates for different models

	Number of variables (n)		Rate of identification (%)	
	Aluminum	Graphite	Aluminum	Graphite
KNN	20	17	73.25%	95.88%
PCA-KNN	8	9	82.53%	96.19%
RF	20	17	85.36%	98.93%
SVM	20	17	96.07%	100.00%

Biomedical Optics EXPRESS

Under the graphite substrate, the SVM identification model exhibits the highest precision, with a Precision, Recall, and F1-score of 1 for all 7 types of bacteria, and its parameter selection is relatively simpler.

4. Conclusion

This study conducted research on the identification of 7 types of bacteria based on LIBS, with a focus on exploring the optimization of machine learning methods and substrates. Experimental results showed that under graphite substrate conditions, KNN, PCA-KNN, RF, and SVM achieved identification rates of 96.87%, 98.11%, 98.93%, and 100% respectively, while under aluminum substrate conditions, the rates are 74.91%, 84.05%, 85.36%, and 96.07% respectively. Therefore, the combination of graphite substrate and SVM machine learning method exhibited the best identification performance. This is attributed to the graphite substrate can excitate more bacterial characteristic spectra and the interference of graphite's C line with other bacterial element spectra is relatively minimal. The LIBS combined with machine learning method for bacterial detection has the advantage of being fast and requiring no complex sample pretreatment, providing a new approach for rapid identification of bacterial species. This method holds broad application prospects in fields such as public security, military warfare, and environmental biology. In the next step, we will continue to expand the types of bacterial samples and substrate materials for further research, and attempt to develop a prototype for precise microbial identification based on LIBS.

Funding. National Natural Science Foundation of China (No. 61975103, No. 61875108, No. 61775125, No. 11434007); National Key Research and Development Program of China (No. 2017YFA0304203); Changjiang Scholars and Innovative Research Team in University of Ministry of Education of China (No. IRT_17R70); Shanxi Major Science and Technology Projects (201804D131036); 111 Project (D18001).

Acknowledgments. The authors acknowledge support in experimental device by State Key Laboratory of Quantum Optics and Optical Quantum Devices at Shanxi University.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. P. Hiremath, P. Bannigidad, and S. S. Yelgond, "An improved automated method for identification of bacterial cell morphological characteristics," IJATCSE **2**(1), 11–16 (2013).
2. A. M. Alvarez, "Integrated approaches for detection of plant pathogenic bacteria and diagnosis of bacterial diseases," Annu. Rev. Phytopathol. **42**(1), 339–366 (2004).
3. C. Brady, D. Arnold, J. McDonald, *et al.*, "Taxonomy and identification of bacteria associated with acute oak decline," World J. Microbiol. Biotechnol. **33**(7), 143 (2017).
4. L. Brunnbauer, Z. Gajarska, H. Lohninger, *et al.*, "A critical review of recent trends in sample classification using laser-induced breakdown spectroscopy (LIBS)," TrAC, Trends Anal. Chem. **159**, 116859 (2023).
5. D. A. Cremers and L. J. Radziemski, *Handbook of Laser-induced Breakdown Spectroscopy*, (John Wiley & Sons Ltd, 2006).
6. R. S. Harmon, J. Remus, N. J. Mcmillan, *et al.*, "LIBS analysis of geomaterials: Geochemical fingerprinting for the rapid analysis and discrimination of minerals," Appl. Geochem. **24**(6), 1125–1141 (2009).
7. E. Tognoni, G. Cristoforetti, S. Legnaioli, *et al.*, "Calibration-free laser-induced breakdown spectroscopy: State of the art," Spectrochim. Acta, Part B **65**(1), 1–14 (2010).
8. Steven J. Rehse, "A review of the use of laser-induced breakdown spectroscopy for bacterial classification, quantification, and identification," Spectrochim. Acta, Part B **154**, 50–69 (2019).
9. W. A. Farooq, M. Atif, W. Tawfik, *et al.*, "Study of bacterial samples using laser induced breakdown spectroscopy," Plasma Sci. Technol. **16**(12), 1009–0630 (2014).
10. T. Kim, Z. G. Specht, P. S. Vary, *et al.*, "Spectral fingerprints of bacterial strains by laser-induced breakdown spectroscopy," J. Phys. Chem. B **108**(17), 5477–5482 (2004).
11. S. Manzoor, S. Moncayo, F. Navarro-Villoslada, *et al.*, "Rapid identification and discrimination of bacterial strains by laser induced breakdown spectroscopy and neural networks," Talanta **121**, 65–70 (2014).

Biomedical Optics EXPRESS

12. H. R. Sun, C. R. Yang, Y. Y. Chen, *et al.*, "Construction of classification models for pathogenic bacteria based on LIBS combined with different machine learning algorithms," Appl. Opt. **61**(21), 20–21 (2022).
13. GF Rao, L Huang, and MH Liu, "Discrimination of microbe species by laser induced breakdown spectroscopy," Chinese J. Anal. Chem. **46**(7), 1122–1128 (2018).
14. Q. Mohaidat, S. Palchaudhuri, and S.J. Rehse, "The effect of bacterial environmental and metabolic stresses on a laser-induced breakdown spectroscopy (LIBS) based identification of Escherichia coli and Streptococcus viridans," Appl. Spectrosc. **65**(4), 386–392 (2011).
15. D. Marcos Martinez, J. A. Ayala, R. C. Izquierdo, *et al.*, "Identification and discrimination of bacterial strains by laser induced breakdown spectroscopy and neural networks," Talanta **84**(3), 730–737 (2014).
16. Q. Q. Wang, G. Teng, X. L. Qiao, *et al.*, "Importance evaluation of spectral lines in Laser-induced breakdown spectroscopy for classification of pathogenic bacteria," Biomed. Opt. Express **9**(11), 5837–5850 (2018).
17. R. Bro and K. Smilde, "Principal component analysis," Anal. Methods **6**(9), 2812–2831 (2014).
18. H. Abdi and L. J. Williams, "Principal component analysis," WIREs Comp. Stats. **2**(4), 433–459 (2010).
19. B. U. Yude, J. Pan, B. Jiang, *et al.*, "Spectral feature extraction based on the DCPCA method," Publ. Astron. Soc. Aust. **30**, e24 (2013).
20. Z. Y. Deng, X. S. Zhu, D. B. Cheng, *et al.*, "Efficient kNN classification algorithm for big data," Neurocomputing **195**, 143–148 (2016).
21. S. C. Zhang, X. L. Li, M. Zong, *et al.*, "Efficient KNN classification with different numbers of nearest neighbors," IEEE Trans. Neural Netw. Learn Syst. **29**(5), 1774–1785 (2018).
22. B. Gregorutti, B. Michel, and P. S. Pierre, "Correlation and variable importance in random forests," Stat. Comput. **27**(3), 659–678 (2017).
23. L. Sheng, T. Zhang, G. Niu, *et al.*, "Classification of iron ores by laser-induced breakdown spectroscopy (LIBS) combined with random forest (RF)," J. Anal. At. Spectrom. **30**(2), 453–458 (2015).
24. R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," BMC Bioinformatics **7**(1), 3 (2006).
25. A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," Pattern Recognit. **44**(2), 330–349 (2011).
26. H. F. Wang, B. C. Zheng, S. W. Yoon, *et al.*, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," Eur. J. Oper. Res. **267**(2), 687–699 (2018).
27. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn. **20**(3), 273–297 (1995).
28. R. Dietrich, M. Opper, and H. Sompolinsky, "Statistical mechanics of support vector networks," Phys. Rev. Lett. **82**(14), 2975–2978 (1999).