Received: 20 October 2022 | Revised: 11 January 2023 | Accepted: 30 January 2023

DOI: 10.1002/hsr2.1108

ORIGINAL RESEARCH

Health Science Reports Open Access WILEY

Machine learning model of imipenem-resistant *Klebsiella pneumoniae* based on MALDI-TOF-MS platform: An observational study

Yu Zeng[1] | Chao Wang[2] | Qing Ye[3] | Gang Liu[4] | Lixia Zhang[4] | Jingjing Wan[1] | Yu Zhu[5]

[1]School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, China

[2]Department of Clinical Laboratory, First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin, China

[3]Department of Hepatology, The Third Central Hospital of Tianjin, Tianjin, China

[4]Department of Clinical Laboratory, Tianjin Haihe Hospital, Tianjin, China

[5]Department of Clinical Laboratory, The Third Central Hospital of Tianjin, Tianjin, China

Correspondence

Yu Zhu, 83 Jintang Road, Hedong District, Tianjin 300350, China.
Email: zhuyutj@126.com

Jingjing Wan, 500 Dongchuan Rd, Minhang District, Shanghai 200241, China.
Email: jjwan@chem.ecnu.edu.cn

Funding information

Basic Research Cooperation Project of Beijing, Tianjin and Hebei Province, China, Grant/Award Number: 20JCZXJC00030

Abstract

Background and Aim: Machine learning is an important branch and supporting technology of artificial intelligence, we established four machine learning model for the drug sensitivity of *Klebsiella pneumoniae* to imipenem based on matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF-MS) and compared their diagnostic effect.

Methods: The data of MALDI-TOF-MS and imipenem sensitivity of 174 cases of *K. pneumoniae* isolated from clinical specimens in the laboratory of microbiology department of Tianjin Haihe Hospital from 2019 January to 2020 December were collected. The mass spectrometry and imipenem sensitivity of 70 cases of imipenem-sensitive and 70 resistant cases were randomly selected to establish the training set model, 17 cases of sensitive and 17 cases of resistant cases were randomly selected to establish the test set model. Mass spectral peak data were subjected to orthogonal partial least squares discriminant analysis (OPLS-DA), the training set data model was established by machine learning least absolute shrinkage and selection operator (LASSO) algorithm, logistic regression (LR) algorithm, support vector machines (SVM) algorithm, neural network (NN) algorithm, the area under the curve (AUC) and confusion matrix of training set and test set model were calculated and selected by Grid search and 3-fold Cross-validation respectively, the accuracy of the prediction model was verified by test set confusion matrix.

Results: The R^2Y and Q^2 of OPLS-DA were 0.546 and 0.0178. The AUC of the best training set and test set models were 0.9726 and 0.9100, 1.0000 and 0.8581, 0.8462 and 0.6263, 1.0000 and 0.7180 evaluated by LASSO, LR, SVM and NN model respectively. The accuracy of the LASSO, LR, SVM and NN model were 87%, 79%, 62%, and 68% in test set, respectively.

Yu Zeng, Chao Wang, and Qing Ye contributed equally to this study.

Conclusion: The LASSO prediction model of *K. pneumoniae* sensitivity to imipenem established in this study has a high accuracy rate and has potential clinical decision support ability.

KEYWORDS

artificial intelligence, imipenem, *Klebsiella pneumoniae*, machine learning

1 | INTRODUCTION

Klebsiella pneumoniae is one of the common clinical pathogens that can cause infection of the respiratory tract, urinary tract, abdominal cavity and other areas, resulting in sepsis, multiorgan dysfunction and even death.[1,2] In recent years, the drug resistance of *K. pneumoniae* has become increasingly serious due to the widespread application of antibacterial drugs and irrational used, which affecting the clinical anti-infection treatment effect and causing harm to the health and life of patients.[3] Specifically, anti-infective therapy should be performed clinically as early as possible according to the susceptibility of *K. pneumoniae* to reduce the case fatality rate in patients with severe or acute infection. At present, it need to conduct susceptibility analysis after the bacterial isolate has been identified by mass spectrometry. At same time, clinical empiric anti-infective therapy has deficiencies such as poor treatment effect and risk of drug resistance.

Machine learning has become an important direction for intelligent laboratory medicine. It is important to obtain clinical value in the diagnosis, classification, efficacy evaluation, and prognosis prediction from abundant medicine data with high dimensionality and redundance. In recent years, machine learning which is an important branch and supporting technology of artificial intelligence via building models and predict new samples through data analysis, has become a hot spot in medicine research fields. Traditional machine learning algorithms are mainly east absolute shrinkage and selection operator (LASSO), logistic regression (LR), support vector machines (SVM), neural network (NN), and so on. Therefore, machine learning technology is becoming a powerful tool for high-precision predictive analysis and auxiliary clinical diagnosis, which can replace subjective judgment by establishing prediction results given by complex digital models and improve patient diagnostic results, which has shown great potential and is a development trend of assisting clinical decision-making technology.[4]

Imipenem is a carbapenem antibacterial drug with a good therapeutic effect on *K. pneumoniae*, however, the resistance of *K. pneumoniae* to imipern has increased year by year, resulting the great difficulties to clinical experience.[5,6] This study is proposed to establish a prediction model of *K. pneumoniae* to imipenem drug sensitivity by LASSO, LR, SVM, and NN machine learning algorithms based on matrix-assisted laser desorption/ionization time-of-flight mass (MALDI-TOF MS) spectrometry, and compare the diagnostic efficacy of different algorithms to explore the potential clinically assisted decision support methods.

2 | METHODS

2.1 | Study participants

The nonrepetitive samples of 174 patient with *K. pneumoniae* infection in Tianjin Haihe Hospital from 2019 January to 2020 December were collected and patient's characteristic were shown in Table 1. The present study was approved by the Medical Ethics Committee of Tianjin Haihe Hospital, was performed according to the principles of the Declaration of Helsinki (2019HHKT-016) and all patients provided written informed consent before participation in the study. All waste materials were well sterilized before disposing to the Environment.

2.2 | Bacteria isolation collection and identification

2.2.1 | Bacteria isolation preparation

All sample were inoculated on MacConkey agar plates, and individual colonies were isolated after 24 h of incubation at 37°C.

2.2.2 | MALDI-TOF MS assay

Individual colony was selected and transfer to MALDI-TOF MS target plate with 1 μL of α-cyano-4-hydroxycynnamic acid-matrix solution (HCCA; Bruker Daltonik) and dried at room temperature. The acquisition and analysis of mass spectrum data were carried out by Microflex LT spectrometer (Bruker Daltonik) with a linear positive mode covering the molecular weight range of 2 to 20 kDa, laser frequency of 60 Hz and 240 shots in several points with a pulsed nitrogen laser ($\lambda = 337$ nm). The mass spectrum identification scores ≧2.0 was considered as high confidence identification. Three replicates for every isolate were performed.

2.2.3 | Drug sensitive assay

Imipenem resistance was analyzed by minimum inhibitory concentration (MIC) method and VITEK-2 compact system (BioMerieux), ≦1 μg/mL is defined as sensitive and ≧4 μg/mL is defined as resistance based

TABLE 1 Demography of study participants.

Characteristics	*n* (%)
Age	
<21	1 (0.6)
21–30	4 (2.3)
31–40	11 (6.3)
41–50	15 (8.6)
51–60	28 (16.1)
>60	115 (66.1)
Median age	67
Age range	18–88
Gender	
Male	123 (70.7)
Female	51 (29.3)
Imipenem sensitive	
Sensitive	87 (50)
Resistance	87 (50)
Sample type	
Broncho-alveolar lavage	13 (7.5)
Blood	9 (5.2)
Pleural fluid	2 (1.2)
Pus	7 (4.0)
Secretion	2 (1.1)
Sputum	130 (74.7)
Urine	11 (6.3)

on Clinical and Laboratory Standards Institute (CLSI). Three replicates for every isolate were performed.

2.2.4 | Data quality control

K. pneumoniae ATCC700603 was as extended spectrum beta-lactamases (ESBLs) positivity control bacterial isolate, *Klebsiella Pneumoniae* (ATCC BAA-1706), *Escherichia coli* ATCC25922 and ATCC8739 was as ESBLs negative control bacterial isolates, which obtained from American Type Culture Collection (ATCC). MALDI-TOF-MS and drug sensitive analysis was operated by experienced operator, each bacterial isolate is repeatedly tested three times. MALDI-TOF-MS was calibrated by *E. coli* ATCC8739 every month. The positive control *K. pneumoniae* ATCC700603 and ESBLs negative control *E. coli* ATCC25922 are used for quality control monitoring for drug sensitive assay very week. Daily monitoring of laboratory temperature and humidity environments is carried out, reagents such as mass spectrometry matrix and drug sensitive reagent are examined to avoid interference caused by personnel, environment and reagent factors.

2.3 | Machine learning

2.3.1 | Data filtering

According to the drug sensitivity distribution of imipenem bacterial isolates, 70 cases of pneumonia Kraber resistant bacterial isolates and 70 cases of sensitive bacterial isolates were randomly selected from 174 samples as training sets, and 17 cases of resistant bacterial isolates and 17 cases of sensitive bacterial isolates were randomly selected as test sets, and the mass spectrometry data peaks of the above bacterial isolates were preprocessed according to the following steps.

2.3.2 | Data preprocessing

The mass-to-charge ratio (m/z) data of mass spectrometry is standardized and normalized, the starting m/z value is set to 1962DA, the ending m/z value is 19998DA, the resampling number is 15000, the Gaussian filtering algorithm is used for data smoothing, the σ value is set to 1, the White Top Hat algorithm is used for baseline correction, the baseline correction window value is set to 5, and the mass spectral peak m/z is processed by equal-distance binning. The signal processing function in the Python 3.8.8 software (www.python.org) scipy package was used for mass spectrometry peak extraction, and finally 3457 mass spectrometry peaks were extracted from the mass spectrometry analysis results of each bacterial isolate, and the above mass spectrometry peaks were used as parameters to participate in the model construction.

2.3.3 | Clustering

All mass spectrometry peak intensity and m/z data of bacterial isolate were selected and orthogonal partial least squares discriminant analysis (OPLS-DA) was performed using MetaboAnalyst 5.0 software (https://www.metaboanalyst.ca).

2.3.4 | Data modeling and accuracy verification

The mass spectrometry peak data of the training set was selected and the LASSO, LR, SVM and NN algorithm models were obtained by the Python 3.8.8 software Scikit-learn machine learning package according to the following formula (Figure 1), and the cut off for calculating the probability of sensitivity or resistance was >0.6 to defined as sensitive or resistant. The above-mentioned optimal model was screened and area under curve (AUC) was by analyzed by Grid search and 3-fold Cross-validation. The confusion matrix of training set and

LASSO

$$Q(\beta) = ||y - \boldsymbol{X}\beta||^2 + \lambda||\beta||_1$$

$$\iff \arg\min ||y - \boldsymbol{X}\beta||^2 \quad s.t. \sum |\beta_j| \le s$$

SVM

$$\max_{(\vec{w},b)} \min_{i} \frac{1}{\|\vec{w}\|}|\vec{w}\cdot\vec{x}_i + b|$$

$$\text{s.t.} \quad y_i(\vec{w}\cdot\vec{x}_i + b) > 0, \quad i = 1, 2, \ldots, m$$

LR

$$y = \sigma(f(\boldsymbol{x})) = \sigma(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^T\boldsymbol{x}}}$$

NN

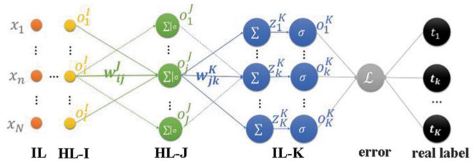

FIGURE 1 The formula of machine learning. HL, hidden layer; IL, input layer; LASSO, least absolute shrinkage and selection operator; LR, logistic regression; NN, neural network; SVM, support vector machines.

test set was established using the Python 3.8.8 software Scikit-learn machine learning package. The drug sensitivity, drug resistance and total accuracy was calculated, respectively.

3 | RESULTS

3.1 | Clustering analysis of imipernem-resistant and sensitive bacterial isolates of *K. pneumoniae*

The MALDI-TOF-MS data of imipenem resistant bacterial isolates (R group) and imipenem sensitive bacterial isolates (S group) was analyzed by OPLS-DA analysis (Figure 2), the results show that R²Y and Q² of OPLS-DA analysis are 0.546 and 0.0178, respectively, indicating that the poorer the fitting accuracy for imipenem sensitive model of *K. pneumoniae*. Therefore, we need to established a new mass spectrometry analysis model to distinguish imipenem sensitivity and resistance.

3.2 | Diagnostic efficiency of imipenem drug sensitive training set and test set

The LASSO, LR, SVM, and NN algorithms were all selected by grid search algorithm and 3-fold cross-validation to select the optimal model. The lambda value of the selected LASSO model is 0.195, the SVM kernel function is Gaussian core, NN has 2 hidden layers, the first layer has 64 neurons, the second layer has 32 neurons, using ReLU (rectified linear unit) activation function, Adam (adaptive momentum) algorithm optimization parameters, cross entropy as the loss function. The imipenem drug sensitive training set and test set AUC of the LASSO, LR, SVM, and NN algorithms are 0.9726 and 0.9100, 1.0000 and 0.8581, 0.8462 and 0.6263, 1.0000 and 0.7180, respectively (Figure 3).

3.3 | Algorithm verification

The confusion matrix of the training set and the test set is established, and the training set of the imipenem drug sensitive prediction model (LASSO, LR, SVM, and NN algorithms) was verified by the test set, and the prediction accuracy and drug sensitivity/resistance prediction accuracy are shown in Figures 4 and 5.

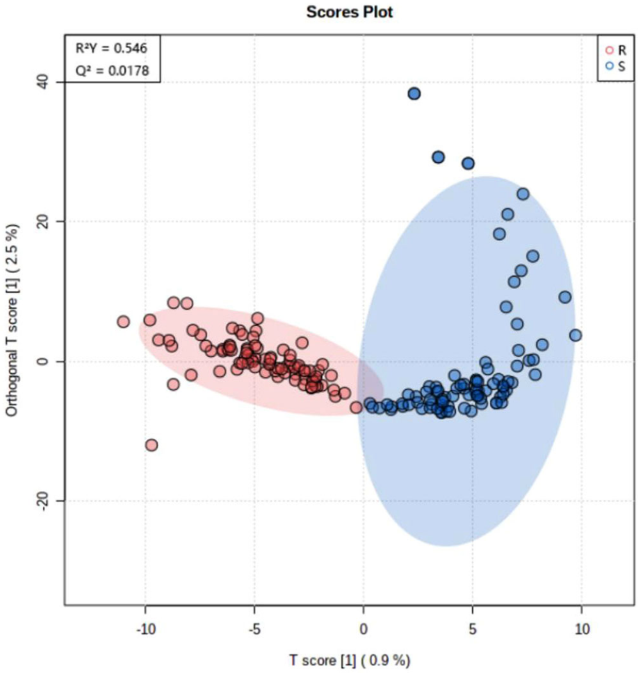

FIGURE 2 The OPLS-DA analysis of imipenem resistant and sensitive of *Kleborgs pneumoniae*. R, imipenem resistance; S, imipenem sensitive.

4 | DISCUSSION

K. pneumoniae is a conditioned pathogen of Enterobacter in the intestines and respiratory tracts of normal people, which can lead to severe infections such as intracranial infection, sepsis, septic pulmonary embolism and so on, even threat the life of patients. It requires more than 48 h to cultured bacterial isolate and mass spectrometry identification before obtain microbial drug sensitivity results, resulting in worsening patient's condition.

Previous research has focused on the analysis of microbial drug sensitive through methods such as cluster analysis. However, this

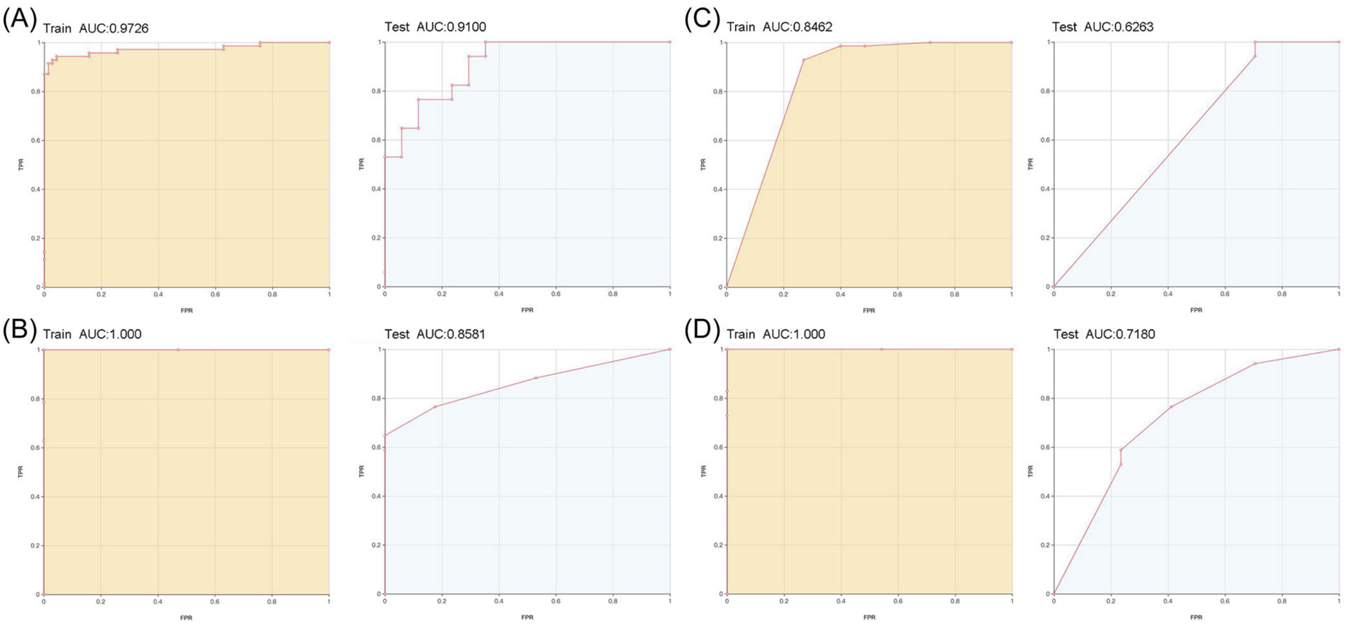

FIGURE 3 The ROC analysis of train and test set for imipenem sensitive of *Kleborgs pneumoniae*. (A) Least absolute shrinkage and selection operator algorithm, (B) logistic regression algorithm, (C) support vector machines algorithm, (D) neural network algorithm.

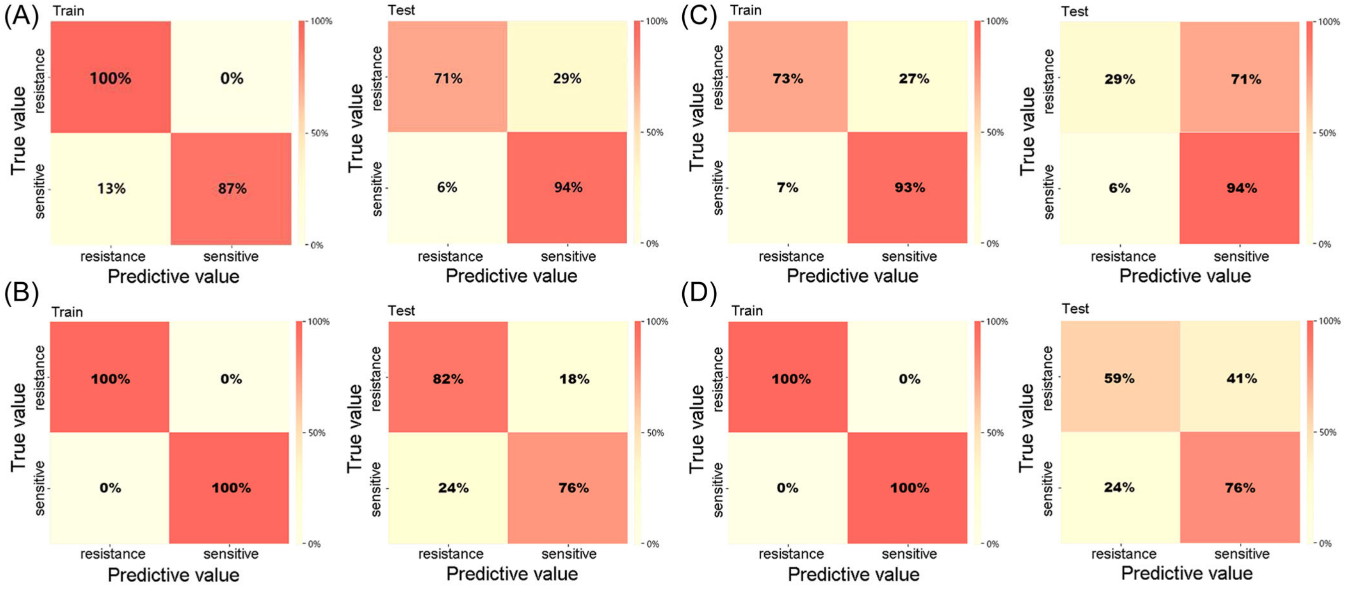

FIGURE 4 The confusion matrix of imipenem sensitive of *Kleborgs pneumoniae* in train set and test set. (A) Least absolute shrinkage and selection operator algorithm, (B) logistic regression algorithm, (C) support vector machines algorithm, (D) neural network algorithm.

study also found that the OPLS-DA analysis was not suitable for prediction to drug sensitive analysis of *K. pneumoniae* imipenem, which may be related to the complexity of mass spectrometry information.

With the increasing volume of medical data, high-dimensional data is difficult to be processed by traditional statistical models, and machine learning can establish efficient and accurate mathematical models based on big data, changing the traditional clinical path, and provide patients with the best possible treatment.

Machine learning technology has broad clinical application prospects, such as CT and MRI image intelligent recognition, drug metabolism based on gut microbiota and so on.[7,8] In terms of microbial identification, Mortier and colleagues and Feucherolles and colleagues try to bacterial isolate identification and Campylobacter diversity analysis based on MALDI-TOF combines machine learning. Tran and colleagues and Deulofeu and colleagues studied the machine learning model for intelligent diagnosis of the novel

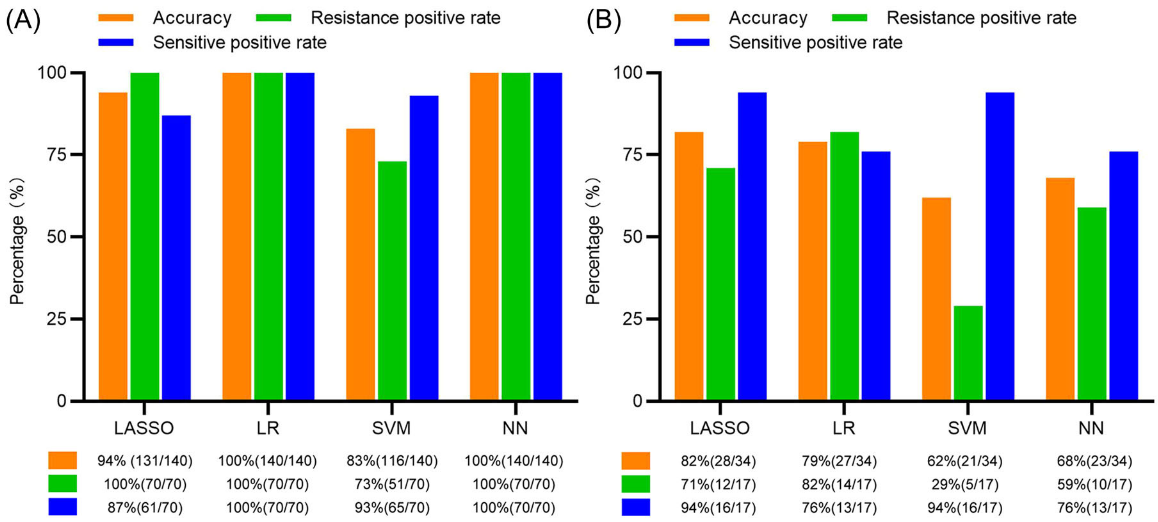

FIGURE 5 The accuracy and positive rate of different algorithm in imipenem sensitive and resistance of *Kleborgs pneumoniae*. (A) Train set; (B) test set. LASSO, least absolute shrinkage and selection operator; LR, logistic regression; NN, neural network; SVM, support vector machines.

coronavirus pneumonia (COVID-19) based on mass spectrometry data.[9–12]

At present, there are also some studies on the application of machine learning techniques to the drug sensitive analysis of *K. pneumoniae*, such as Lu and colleagues and Liu and colleagues, which based on Raman spectroscopy and whole genome sequencing technology,[13,14] however Raman spectroscopy and whole genome sequencing technology have not yet been used as routine techniques for microbial identification.

There also was a study about rapid detection of carbapenem-resistant *K. pneumoniae* using machine learning and MALDI-TOF MS Platform by Wang et al.[15] RF, SVM and SVM-K (radial basis function kernel) model to distinguish carbapenem sensitive and resistant *K. pneumoniae*, indicating a best AUC 0.9356 and accuracy 0.91 by SVM-K model with peak dimensionality reduction in train set.

In this study, all of 3457 mass spectrometry peaks information were extracted as parameters to participate in model construction for more complex machine learning algorithm analysis. The LASSO, LR, SVM and NN algorithms were applied to imipenem sensitivity for *K. pneumoniae* based MALDI-TOF-MS data by Python software, and optimal model was selected by Grid search algorithm and 3-fold cross-verification. It was found that the training set ACU of LASSO, LR, and NN algorithms could nearly reach 1.0000, but the test set of LASSO algorithm had a highest AUC.

We also found that the accuracy of LASSO, LR, SVM, and NN algorithm were 94%, 100%, 83%, and 100% in training sets, respectively, meanwhile, the accuracy of LASSO, LR, SVM, and NN algorithm were 82%, 79%, 62%, and 68% in test set, respectively. Compared with Raman spectroscopy-based machine learning models based on the study of Lu et al.[13] The LASSO algorithm in this study has a well accuracy, but the SVM algorithm has a lower accuracy in test set. Compared with Wang et al.[15] there was a better AUC and accuracy of LASSO algorithm in train set. Through the comparison of the diagnostic efficiency of the above algorithms, it is found LASSO may be an more ideal predictive model algorithm with better AUC and accuracy.

5 | CONCLUSION

In this study, we established an inexpensive and well-accuracy machine learning model based on the existing conventional application of MALDI-TOF-MS technology to predict the drug sensitivity of *K. pneumoniae* to imipenem, to assist clinical antimicrobial application decision-making.

This study is an exploration on the prediction of drug sensitivity of imipenem for *K. pneumoniae* using machine learning methods, demonstrating the potential of artificial intelligence in the field of microbial resistance identification. However, multicenter big data studies need to be established to further improve overall accuracy to drive clinical application. At same time, it is still an early stage of exploration for the application of artificial intelligence diagnostic methods such as machine learning to the clinical diagnosis, but there will be a broad application prospects for integration of clinical expertise and big data analysis. Moreover, there would be many preparations before the clinical application based machine learning model, for example, it was need to established specialized machine learning model for every common bacterial due to the different characteristics and prediction effect for different bacterial, meanwhile, the accuracy need be improved and so on.

AUTHOR CONTRIBUTIONS

Yu Zeng: Formal analysis; investigation; methodology; writing—review & editing. **Chao Wang**: Methodology; writing—original draft. **Qing Ye**: Investigation; methodology; writing—original draft. **Gang Liu**: Data curation; writing—original draft. **Lixia Zhang**: Funding acquisition; project administration. **Jingjing Wan**: Investigation; methodology; project administration. **Yu Zhu**: Conceptualization; investigation; project administration.

ACKNOWLEDGMENTS

The present study was funded by Basic Research Cooperation Project of Beijing, Tianjin, and Hebei Province, China (20JCZXJC00030). The financial relationships had no involvement in the study design, collection, analysis, and interpretation of data or writing or submitting of the report for publication. We don't have any other financial relationships to disclose.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

No animals were used in this research. All human research procedures followed were in accordance with the ethical standards of the committee responsible for human experimentation (institutional and national), and with the Helsinki Declaration of 1975, as revised in 2013. This study was approved by the Medical Ethics Committee of Tianjin Haihe Hospital.

TRANSPARENCY STATEMENT

The lead author Jingjing Wan, Yu Zhu affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

ORCID

Yu Zhu http://orcid.org/0000-0003-0499-0235

REFERENCES

1. Barnsteiner S, Baty F, Albrich WC, et al. Antimicrobial resistance and antibiotic consumption in intensive care units, Switzerland, 2009 to 2018. *Euro Surveill*. 2021;46:2001537.
2. Wang M, Earley M, Chen L, et al. Clinical outcomes and bacterial characteristics of carbapenem-resistant *Klebsiella pneumoniae* complex among patients from different global regions (CRACKLE-2): a prospective, multicentre, cohort study. *Lancet Infect Dis*. 2022;22:401-412.
3. Vidal-Cortés P, Martin-Loeches I, Rodríguez A, et al. Current positioning against severe infections due to *Klebsiella pneumoniae* in hospitalized adults. *Antibiotics*. 2022;11:1160.
4. Chen D, Liu J, Zang L, et al. Integrated machine learning and bioinformatic analyses constructed a novel stemness-related classifier to predict prognosis and immunotherapy responses for hepatocellular carcinoma patients. *Int J Biol Sci*. 2022;18:360-373.
5. Duan N, Sun L, Huang C, Li H, Cheng B. Microbial distribution and antibiotic susceptibility of bloodstream infections in different intensive care units. *Front Microbiol*. 2021;12:792282.
6. Mendelsohn E, Ross N, White AM, et al. Extracting novel antimicrobial emergence events from scientific literature and medical reports. *F1000Research*. 2020;9:1320.
7. Khamzin S, Dokuchaev A, Bazhutina A, et al. Machine learning prediction of cardiac resynchronisation therapy response from combination of clinical and model-driven data. *Front Physiol*. 2021;12:753282.
8. McCoubrey LE, Thomaidou S, Elbadawi M, Gaisford S, Orlu M, Basit AW. Machine learning predicts drug metabolism and bioaccumulation by intestinal microbiota. *Pharmaceutics*. 2021;13:2001.
9. Mortier T, Wieme AD, Vandamme P, Waegeman W. Bacterial species identification using MALDI-TOF mass spectrometry and machine learning techniques: a large-scale benchmarking study. *Comput Struct Biotechnol J*. 2021;19:6157-6168.
10. Feucherolles M, Nennig M, Becker SL, et al. Investigation of MALDI-TOF mass spectrometry for assessing the molecular diversity of campylobacter jejuni and comparison with MLST and cgMLST: a Luxembourg One-Health Study. *Diagnostics*. 2021;11:1949.
11. Tran NK, Howard T, Walsh R, et al. Novel application of automated machine learning with MALDI-TOF-MS for rapid high-throughput screening of COVID-19: a proof of concept. *Sci Rep*. 2021;11:8219.
12. Deulofeu M, García-Cuesta E, Peña-Méndez EM, et al. Detection of SARS-CoV-2 infection in human nasopharyngeal samples by combining MALDI-TOF MS and artificial intelligence. *Front Med*. 2021;8:661358.
13. Lu J, Chen J, Liu C, et al. Identification of antibiotic resistance and virulence-encoding factors in *Klebsiella pneumoniae* by Raman spectroscopy and deep learning. *Microb Biotechnol*. 2022;15:1270-1280.
14. Liu W, Ying N, Mo Q, et al. Machine learning for identifying resistance features of *Klebsiella pneumoniae* using whole-genome sequence single nucleotide polymorphisms. *J Med Microbiol*. 2021;11.
15. Wang J, Xia C, Wu Y, Tian X, Zhang K, Wang Z. Rapid detection of carbapenem-resistant *Klebsiella pneumoniae* using machine learning and MALDI-TOF MS platform. *Infect Drug Resist*. 2022;15:3703-3710.

How to cite this article: Zeng Y, Wang C, Ye Q, et al. Machine learning model of imipenem-resistant *Klebsiella pneumoniae* based on MALDI-TOF-MS platform: an observational study. *Health Sci Rep*. 2023;6:e1108. doi:10.1002/hsr2.1108