**Accurate classification of *Listeria* species by MALDI-TOF mass spectrometry incorporating denoising autoencoder and machine learning**

Author links open overlay panel[YunhongLi](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)[a](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)[ZeyuGan](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)[a](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)[XiZhou](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)[b](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)[ZhiweiChen](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)[a](https://www.sciencedirect.com/science/article/abs/pii/S0167701221002463" \l "!)

**Show more**

Add to Mendeley

Share

Cite

[https://doi.org/10.1016/j.mimet.2021.106378](https://doi.org/10.1016/j.mimet.2021.106378" \o "Persistent link using digital object identifier" \t "_blank)[Get rights and content](https://s100.copyright.com/AppDispatchServlet?publisherName=ELS&contentID=S0167701221002463&orderBeanReset=true" \t "_blank)

**Abstract**

[Listeria](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/listeria) monocytogenes belongs to the category of facultative [anaerobic bacteria](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/anaerobic-bacterium), and is the pathogen of [listeriosis](https://www.sciencedirect.com/topics/immunology-and-microbiology/listeriosis), potentially lethal disease for humans. There are many similarities between L. *monocytogenes* and other non-pathogenic [*Listeria*](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/listeria) species, which causes great difficulties for their correct identification. The level of L. *monocytogenes* contamination in food remains high according to statistics from the Food and Drug Administration. This situation leads to food recall and destruction, which has caused huge economic losses to the food industry. Therefore, the identification of *Listeria* species is very important for clinical treatment and food safety. This work aims to explore an efficient [classification algorithm](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/classification-algorithm) which could easily and reliably distinguish *Listeria* species. We attempted to classify *Listeria* species by incorporating denoising autoencoder (DAE) and machine learning algorithms in matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS). In addition, convolutional neural networks were used to map the high dimensional original mass spectrometry data to low dimensional core features. By analyzing MALDI-TOF MS data *via* incorporating DAE and [support vector machine](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/support-vector-machine) (SVM), the identification accuracy of *Listeria* species was 100%. The proposed [classification algorithm](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/classification-algorithm) is fast (range of seconds), easy to handle, and, more importantly, this method also allows for extending the identification scope of bacteria. The DAE model used in our research is an effective tool for the extraction of MALDI-TOF mass spectrometry features. Despite the fact that the MALDI-TOF MS dataset examined in our research had high dimensionality, the DAE + SVM algorithm was still able to exploit the hidden information embedded in the original MALDI-TOF mass spectra. The experimental results in our work demonstrated that MALDI-TOF mass spectrum combined with DAE + SVM could easily and reliably distinguish *Listeria* species.

**Introduction**

The *Listeria* species are Gram-positive bacteria which can grow in saline and cold environments (Gray and Killinger, 1966). To the best of our knowledge, there are 26 recognized species representing the genus *Listeria* (Carlin et al., 2021). We selected the six most common *Listeria* species as our research objects, including *Listeria monocytogenes*, *Listeria innocua*, *Listeria welshimeri*, *Listeria grayi*, *Listeria seeligeri* and *Listeria ivanovii* (Bakker et al., 2013). It should be mentioned that only these six species are included in the Centers for Disease Control and Prevention (CDC) database (<https://microbenet.cdc.gov/>) and EXS3000 database (Zybio Inc., Chongqing, China, software version 1.0.10). Among the six *Listeria* species, L. *monocytogenes* and *L. ivanovii* are pathogenic to animals, and L. *monocytogenes* is the only species that has been associated with human diseases, such as mononucleosis*,* meningitis and encephalitis (Schlech, 1996; Vázquez-Boland et al., 2001). According to one report, *L. innocua* also possesses potential pathogenicity (Favaro et al., 2014). The case fatality rate for listeriosis in the untreated susceptible group (pregnant women, immunocompromised individuals, elderly persons, and newborns) has reached 36% (Siegman-Igra et al., 2002; Scallan et al., 2011; Behravesh et al., 2011; Hernandez-Milian and Payeras-Cifre, 2014). The predominant reason for human infection is the occasional contamination of raw food products and ready-to-eat meals. *L. monocytogenes* mainly colonizes livestock and poultry, resulting in easy contamination of meat, eggs, milk, seafood products, vegetables and other foods, which will cause food poisoning and outbreaks of listeriosis (Schlech III and Acheson, 2000; Dalton et al., 1997; Norton et al., 2001). Therefore, the detection and identification of *Listeria* species are crucial for clinical and food safety. The conventional classification approaches for the identification of *Listeria* species are predominantly based on phenotypic characteristics such as biochemical characteristics, biochemical reactions, colony morphology (Jadhav et al., 2012) and serology. Although the above-mentioned methods can reliably classify the species subtypes, they require a series of tedious steps (taking at least 1 to 3 days) and are often time- and resource-consuming (Cherkaoui et al., 2010). Yan et al. (Yan et al., 2021) achieved an accuracy of 93.33% for classifying *Listeria* species based only on their extended Raman spectra (400–1800 cm−1). Quero et al. (Quero et al., 2014) developed a real-time PCR-based method for the rapid identification of *Listeria* species. However, many laboratories still do not have the necessary equipment for such analyses.

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a relatively new technology that is reliable and achieves fast microbial identification based on the whole-cell proteome fingerprint of the microorganisms (Bilecen et al., 2015; Fenselau and Demirev, 2001; Holland et al., 1996; Barbuddhe et al., 2008a, Barbuddhe et al., 2008b), and it has been introduced into many clinical microbiology laboratories in recent years. MALDI-TOF MS is easy-to-use and high-throughput compared to the traditional phenotypic identification and molecular biology techniques, greatly improving the identification accuracy and efficiency for clinical microorganisms (Dubois et al., 2012) and is gradually replacing the conventional identification approaches (Cherkaoui et al., 2010; Sauer and Kliem, 2010; Van Veen et al., 2010; Bizzini et al., 2010; Martiny et al., 2012). MALDI-TOF MS has been successfully used to classify L. *monocytogenes* in previous reports, such as identification of L. *monocytogenes* from selective enrichment broth (Jadhav et al., 2014), the rapid identification of *Listeria* species (Marklein et al., 2009; Rychert et al., 2013; Hsueh et al., 2014; Jadhav et al., 2015; Ojima-Kato et al., 2016), the discrimination of lineages (Ojima-Kato et al., 2016) and the subtyping of L. *monocytogenes* (Marklein et al., 2009; Jadhav et al., 2015; Ojima-Kato et al., 2016). Microbiological laboratories usually apply database matching algorithms to classification of unknown mass spectra. Normally, a standard spectrum is obtained from multiple samples of a single bacterial strain and a sample spectrum to be matched should be compared with the standard spectrum in the database by calculating the similarity among multiple parameters, such as peak position and the area under the peak. After that, a matching score list will be obtained. The species with the highest score in the list guides the final identification result. The database matching algorithms can identify *Listeria* genus 100% correctly. However, the database matching algorithm faces difficulties in identifying similar microbial species with similar MALDI-TOF mass spectral patterns such as L. *monocytogenes* and L. *innocua*, and thus accurate species-level identification results cannot be obtained (Farfour et al., 2012). Besides, the accuracy of mass spectra identification depends on the quality of the utilized database. Incomplete information in the strain database will limit the accuracy of identification. Enriching the strain database can improve the accuracy of identification results.

The analysis of MALDI-TOF MS based on the database matching algorithm relies on less attributes of spectral, such as the area under the peak and the peak height that are associated with microbial species (Weis et al., 2020). Hence, there is an abundance of information included in MALDI-TOF MS that remains unused. Machine learning (ML) is a group of methods for finding patterns from specific datasets. Various ML algorithms, including k-nearest neighbors (KNN), naive Bayes (NB), random forest (RF) and support vector machine (SVM), are stable and robust (Wang et al., 2018a). The ML model could utilize a large number of calculations to discover unintuitive or even counter-intuitive statistical information from the training set, and use the learned pattern to classify the unknown test set (Goodwin et al., 2014). Many prior researches have been proven to succeed in utilizing ML in medical decision-making fields (Wang et al., 2018b, Wang et al., 2018c; Wang et al., 2016; Tseng et al., 2019).

For high dimensional MALDI-TOF MS data, traditional ML algorithms may fail due to the curse of dimensionality. Hence, dimensionality reduction is an essential step for the application of ML algorithms, especially when using a dataset with a relatively small sample size to train the ML model. In addition, due to high dimensionality in MS data (Goodwin et al., 2014), the detection model is prone to overfitting.

In recent years, deep learning has generated breakthrough results across multiple fields including image recognition, speech recognition, drug discovery, genomics and natural language processing (LeCun et al., 2015). The use of multi-layer deep neural networks can facilitate the identification of highly complex and otherwise indiscoverable patterns in large and often noisy datasets. The features extracted from large-scale datasets by the pre-trained convolutional neural networks (CNN) models have been shown to achieve excellent performance in image classification tasks. Varshni et al. (Varshni et al., 2019) presented a pneumonia detection method, which used CNN to extract features. Feng *et al* (Feng et al., 2021) integrated artificial neural networks to extract multi-wavelength transmission spectrum features for detecting waterborne bacteria. Their model presented validation accuracy above 97.3% and test accuracy of 100%. In this paper, we utilized a denoising autoencoder (DAE) model and CNN to extract latent features, which might be important for classification to fully exploit the information embedded in MALDI-TOF MS, and to prevent overfitting. DAE is an artificial neural network that works in an unsupervised manner. DAE could efficiently reduce the redundancy of the input data and encode it. DAE is often utilized for dimensionality reduction and compression of data, more specifically, training this network model can remove noise and learn advanced and abstract features from the original input data (Hofer-Schmitz et al., 2018). In addition to being used as a feature extraction method, DAE performs a noise addition process on the original input data in advance, then in the training process, and finally tries to reconstruct the input dataset as close as possible to the raw input dataset (Xing et al., 2016). The noise addition operation is effective during training and automatically fails during prediction. DAE not only prevented our model from overfitting but also reduced the dimensionality of the MS data that were used to train the classification models. Various classifiers, such as KNN, NB, RF and SVM, were used for the classification task after performing feature extraction.

Five-fold cross-validation was used to test and evaluate models. Training sets were divided into five equal partitions. One partition was selected as the validation set, and the remaining partitions were selected as the training set. In this way, five times independent model training and validation were carried out. Finally, the average results of the five times of validation results were used to calculate the model's validation error. After that, the trained model could be used to make predictions on the independent testing dataset. Seven metrics were used in the present paper, including accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1 score, weighted-average precision, weighted-average recall and weighted-average F1 score to evaluate the performance of the classification models. The prediction results of classification problems could be divided into the following four categories: true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

Accuracy is a commonly used metric to evaluate the classification results. Accuracy represents the proportion of the correctly classified samples over the total number of samples and is defined as follow:Accuracy=TP+TNTP+TN+FP+FN

However, if the accuracy of the model is high for the majority class and low for the minority class, and the sample size of the former is much higher than the one of the latter, the accuracy of direct calculation of the whole can be easily misleading. Precision and recall are also two important evaluation metrics used in classification problems. It is often not objective to use only one of these two metrics to evaluate the classification performance of the model. Traditionally, the F1 score is usually used to balance the tradeoff between these two metrics. However, in the multi-classification problem, to comprehensively evaluate the classification effect of the model for each category, the macro average is used as the evaluation metric. In particular, the accuracy, recall and F1 score of each category are calculated separately, and then the average is taken. Furthermore, the number of samples in each category in our dataset was imbalanced. The weighted average method was used to partially overcome the misunderstandings caused by the imbalanced nature of the dataset. This was accomplished by calculating the accuracy, recall and F1 score of each category separately and then taking the weighted average according to the number of samples in each category. The equations used to quantify these performance metrics are presented below (N represents the number of categories in a multi-category, Ci represents the weight of each category):Recall=TPTP+FNPrecision=TPTP+FPF1Score=2×Precision×RecallPrecision+RecallMacro Precision=∑i=1NPrecisioniNMacro Recall=∑i=1NRecalliNMacroF1Score=∑i=1NF1ScoreiNWeighted Precision=∑i=1NPrecisioni×Ci∑i=1NCiWeighted Recall=∑i=1NRecalli×Ci∑i=1NCiWeightedF1Score=∑i=1NF1Scorei×Ci∑i=1NCi

**Section snippets**

**Materials and methods**

Fig. 1 shows the steps of the proposed classification algorithm. In the following section, we briefly describe each of these steps.

**Identification by database matching**

348 testing data acquired by MALDI-TOF MS were analyzed using the EXS3000 and CDC database. After that, the bacterial identification results and their matching score list were obtained. Table 2 shows the real category and the EXS3000's classification results and Table 3 shows the real category and the CDC's classification results of the six categories with the confusion matrix, providing details about the incorrect classifications. Entries in the shaded area represent correct classifications

**Conclusion**

In recent years, mass spectrum analysis has become increasingly important in the field of bioinformatics, especially in the analysis of bacterial data (Burckhardt and Zimmermann, 2018). In this paper, CNN and DAE were used to extract features from the original MALDI-TOF mass spectrum for *Listeria* classification at the species level. Meanwhile, four machine learning models including KNN, NB, RF, SVM, and a CNN model were utilized to classify *Listeria* species based on learned mass spectrum

**Authors' contributions**

All authors have participated in the conception and design, or analysis and interpretation of the data, drafting the article or revising it critically for important intellectual content, and approval of the final version.

**Funding information**

This work was supported by the China Scholarship Council (No. 201806315005) and the scholarship for overseas studies in Fujian Province.

**Disclosure statement**

No competing financial interests exist.

**References (53)**

* H.D. Bakker*et al.*

**Genome sequencing identifies listeria fleischmannii subsp. coloradonensis subsp. nov. isolated from a ranch**

Int. J. Syst. Evol. Microbiol.

(2013)

* S.B. Barbuddhe*et al.*

**Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry**

J. Appl. Environ. Microbiol.

(2008)

* S.B. Barbuddhe*et al.*

**Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry**

Appl. Environ. Microbiol.

(2008)

* B.C. Behravesh*et al.*

**Deaths associated with bacterial pathogens transmitted commonly through food: foodborne diseases active surveillance network (FoodNet), 1996–2005**

J. Infect. Dis.

(2011)

* K. Bilecen*et al.*

**Performances and Reliability of Bruker Microflex LT and VITEK MS MALDI-TOF Mass Spectrometry Systems for the Identification of Clinical Microorganisms**

Biomed. Res. Int.

(2015)

* A. Bizzini*et al.*

**Performance of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of bacterial strains routinely isolated in a clinical microbiology laboratory**

J. Clin. Microbiol.

(2010)

* I. Burckhardt*et al.*

**Susceptibility testing of bacteria using MALDI-TOF mass spectrometry**

Front. Microbiol.

(2018)

* C.R. Carlin*et al.*

**Listeria cossartiae sp. nov., Listeria immobilis sp. nov., Listeria portnoyi sp. nov. and Listeria rustica sp. nov., isolated from agricultural water and natural environments**

Int. J. Syst. Evol. Microbiol.

(2021)

* A. Cherkaoui*et al.*

**Comparison of two matrix-assisted laser desorption ionization-time of flight mass spectrometry methods with conventional phenotypic identification for routine identification of bacteria to the species level**

J. Clin. Microbiol.

(2010)

* C.B. Dalton*et al.*

**An outbreak of gastroenteritis and fever due to Listeria monocytogenes in milk**

N. Engl. J. Med.

(1997)

* D. Dubois*et al.*

**Performances of the Vitek MS matrix-assisted laser desorption ionization–time of flight mass spectrometry system for rapid identification of bacteria in routine clinical microbiology**

J. Clin. Microbiol.

(2012)

* E. Farfour*et al.*

**Evaluation of the Andromas matrix-assisted laser desorption ionization–time of flight mass spectrometry system for identification of aerobically growing gram-positive bacilli**

J. Clin. Microbiol.

(2012)

* M. Favaro*et al.*

**First case of Listeria innocua meningitis in a patient on steroids and eternecept**

JMM Case Rep.

(2014)

* C. Feng*et al.*

**Artificial Neural Networks Combined Multi-Wavelength Transmission Spectrum Feature Extraction for Sensitive Identification of Waterborne bacteria**

Spectrochim. Acta. A. Mol. Biomol. Spectrosc.

(2021)

* C. Fenselau*et al.*

**Characterization of intact microorganisms by MALDI mass spectrometry**

Mass Spectrom. Rev.

(2001)

* C.R. Goodwin*et al.*

**Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data**

Anal. Chem.

(2014)

* M.L. Gray*et al.*

**Listeria monocytogenes and listeric in infections**

Bacteriol. Rev.

(1966)

* A. Hernandez-Milian*et al.*

**What Is New in Listeriosis?**

Biomed. Res. Int.

(2014)

* K. Hofer-Schmitz*et al.*

**One-class autoencoder approach to classify Raman spectra outliers**

ESANN.

(2018)

* R.D. Holland*et al.*

**Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry**

Rapid Commun. Mass Spectrom.

(1996)

* P.R. Hsueh*et al.*

**Bruker Biotyper matrix-assisted laser desorption ionization–time of flight mass spectrometry system for identification of Nocardia, Rhodococcus, Kocuria, Gordonia, Tsukamurella, and Listeria species**

J. Clin. Microbiol.

(2014)

* S. Jadhav*et al.*

**Methods used for the detection and subtyping of Listeria monocytogenes**

J. Microbiol. Methods

(2012)

* S. Jadhav*et al.*

**Detection of Listeria monocytogenes from selective enrichment broth using MALDI–TOF mass spectrometry**

J. Proteome

(2014)

* S. Jadhav*et al.*

**Rapid identification and source-tracking of Listeria monocytogenes using MALDI-TOF mass spectrometry**

Int. J. Food Microbiol.

(2015)

* D. Kingma*et al.*

**Adam: A Method for Stochastic Optimization**

(2014)

* J. Kolibal*et al.*

**MALDI-TOF Baseline Drift Removal Using Stochastic Bernstein Approximation**

EURASIP J. Adv. Sig. Pr.

(2006)