Università degli Studi
di Milano-Bicocca

Advanced Machine Learning

Final Project

# Deep Learning for Bacteria Identification using Raman Spectroscopy

*Authors:*
Davide Marchetti - 815990 - d.marchetti8@campus.unimib.it
Francesco Stranieri - 816551 - f.stranieri1@campus.unimib.it

February 2021

**Abstract**

Considering the spectral data, classical machine learning techniques have been widely applied, while relatively little work was done in adapting deep learning models [1]. In recent years, CNNs have been applied with great success to a wide variety of computer vision tasks, such as ImageNet classification [2]. Here we apply deep learning approaches to accurately identify 30 common bacterial pathogens, reaching an average isolate-level accuracy exceeding 78%, and an antibiotic treatment identification accuracy of 95%.

# 1 Introduction

Bacterial infections are a leading cause of death in almost every nation, causing more than 6.7 million deaths every year [3][4]. In the United States alone, these infections require $33 billion for annual healthcare spending, equivalent to 8.7% of the total healthcare expenditure [5], and hence are particularly costly.

To detect and identify the bacteria and its antibiotic resistance, current diagnostic methods include sample culture. However, this is a slow process, that can take days even in state-of-the-art laboratories [6][7]. New methods for rapid culture-free diagnosis of bacterial infections are therefore required. *Raman spectroscopy* has the potential to identify the species of bacteria and, consequently, its antibiotic susceptibility. Molecular compositions describe different bacterial phenotypes, leading to subtle variations in their corresponding Raman spectra.

In this project, we initially conceive a *Convolutional Neural Network* (CNN) to classify bacterial spectra and the related antibiotic treatment, measuring the isolate-level accuracies and comparing it with a state-of-the-art CNN. Secondly, we compare two popular machine learning algorithms against our CNN model, in order to evaluate if a *deep approach* could lead to better performance.

# 2 Dataset

We consider a training dataset of 60 000 spectra, along with fine-tuning and test datasets, the latter gathered from *independently* cultured samples, each consisting of 3 000 spectra. All of these datasets[1] represent the same 30 bacterial isolate classes, utilized to produce the results in the paper taken as a reference, namely *"Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning"* [8]. These 30 isolate classes cover more than 94% of all bacterial infections treated

[1] https://github.com/csho33/bacteria-ID

at Stanford Hospital in the years 2016–2017, and are therefore indicative of most infections in intensive care units worldwide [9].

Raman spectra of bacterial species can be difficult to distinguish clearly, as observable in Figure 1.

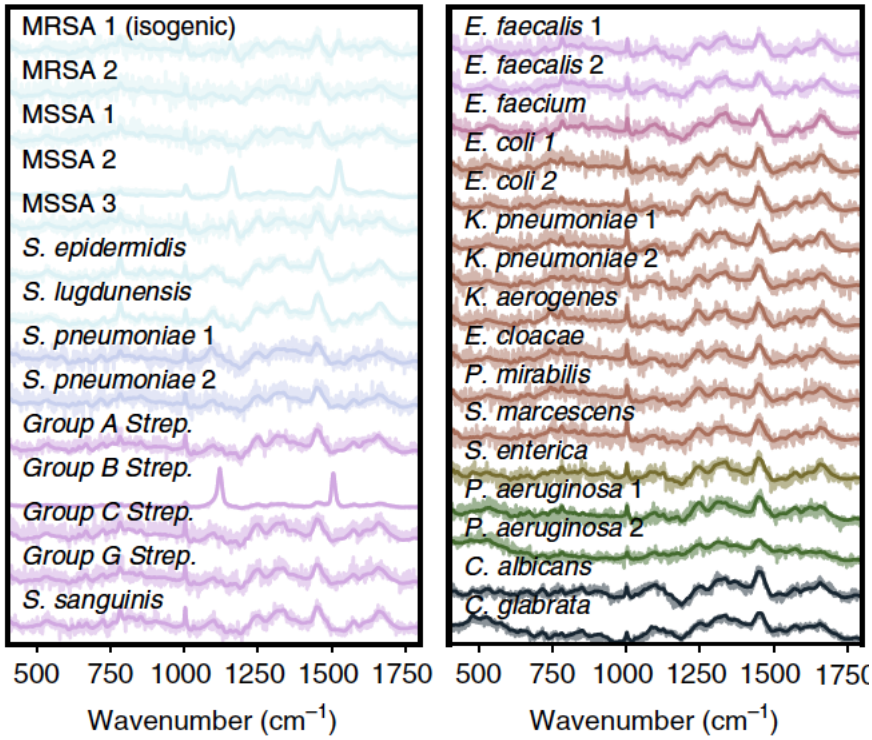


**Figure 1:** Raman spectra of considered bacterial species. The training dataset contains 2 000 spectra for each of the 30 reference isolates, while the fine-tuning and test datasets each consist of 100 spectra for each of the 30 reference isolates.

It is important to note that the input of the utilized algorithms does not consist of two-dimensional images, but rather *one-dimensional* spectra. It is also relevant to consider that intra-sample variance is high [8]. For 19 out of 30 isolates, spectra from at least one other isolate are more similar, on average, than spectra from the same isolate. Consequently, a large number of spectra per sample may help to better reflect the full distribution of data, in order to provide better predictive performance.

# 3 The Methodological Approach

The first question concerning this project is whether the CNN architecture used in the paper taken as a reference is *overkill*, that is if similar performance can be achieved with a simpler model.

The reference architecture is adapted from Resnet [10] and consists of an initial convolution layer with 64 filters, followed by 6 residual layers. Each residual layer contains 4 convolutional layers, each of which has 100 filters. Finally, there is a fully connected layer. The total depth of the network is then 26 layers.

Our CNN architecture, inspired from DeepCID [11] and shown in Figure 2, consists of two pairs of convolutional and max-pooling layers followed by a fully connected layer. The total depth of our network is decidedly smaller, being only 6 layers. Both the CNNs outputs a probability distribution across the 30 reference isolates, and the maximum value is taken as the predicted class.

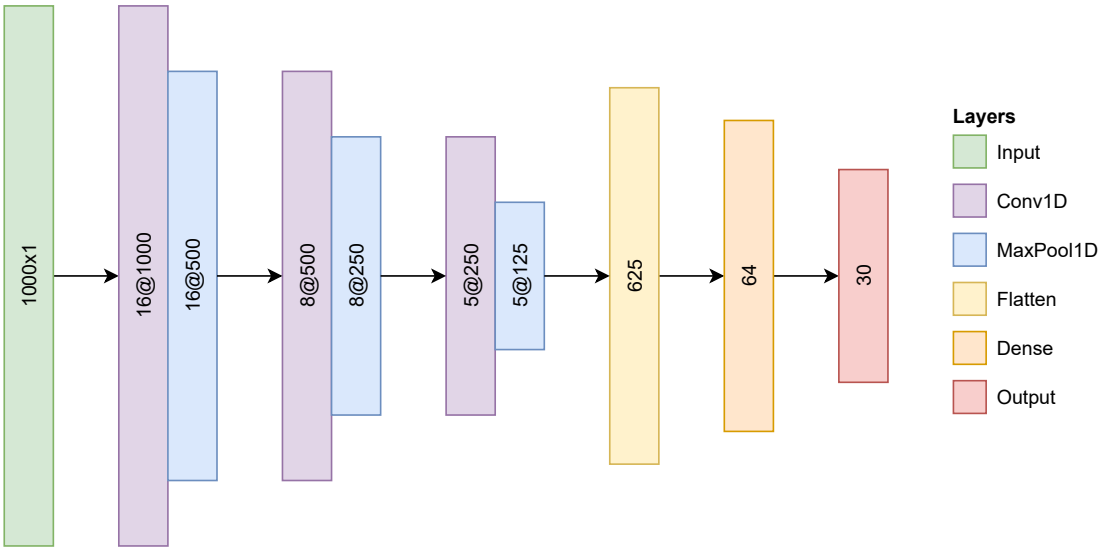


**Figure 2:** Presented CNN architecture, with a total depth of 6 layers.

The hyperparameters are selected through 3-fold *cross-validation* on the training dataset via *grid search*, according to the experimental plan described in Table 1. Across all experiments, the Adam optimizer [12] with a learning rate of 0.001 is maintained.

For each configuration of hyperparameters, we use the average test accuracy measured on the 3 splits of the cross-validation to perform *model selection*, with the aim of obtaining the most accurate model.

| Hyperparameter | Values |
|---|---|
| `batch_size` | 16, 32, *64* |
| `conv_layer` | 2, *3* |
| `filters` | *16*, 32 |
| `kernel_size` | *3*, 5 |
| `units` | *256*, 512, 1024 |
| `dropout_rate` | .3, *.5* |

**Table 1:** Experimental plan for the choice of the CNN's hyperparameters. In total, we performed 144 experiments. The used configuration is highlighted in italic.

To achieve results comparable with the reference paper, a *fine-tuning process* is needed. As a consequence, the fully connected layers are replaced with new ones, while the convolutional layers are frozen. The fine-tuning dataset is then split into 80% train and 20% validation, and the model is trained once more. Moreover, the corresponding accuracies are visualized to detect potential *overfitting*. Finally, we evaluate the performance on the test data and compare it with the reference test accuracy.

The second question is if *conventional* machine learning algorithms can be used on a smaller 15-class classification task, and if the performances are comparable to those achieved by our CNN model. It is important to point out that, for this task, only the fine-tuning dataset has been considered.

The chosen machine learning algorithms are *Support Vector Machine* (SVM) and *Multi-Layer Perceptron* (MLP), which are two of the more common classification techniques in the field of spectral data [1].

To define the hyperparameters for both models, we use grid search on fine-tuning dataset through 5-fold cross-validation, according to the experimental plans described in Tables 2 and 3. To perform model selection, we still consider the best average test accuracy across the 5 splits of the cross-validation.

With regard to SVM, we train the model on the entire fine-tuning dataset. In order to decreases computation costs, we apply *Principal Component Analysis* (PCA) reducing the input dimension from 1000 to 20.

For MLP training, we instead split the fine-tuning data into 80% train and 20% validation and then visualize the corresponding accuracies, in order to detect possible overfitting.

Lastly, for the CNN model we exploit the fine-tuning process previously described once again, but this time on the 15-class.

| Hyperparameter | Values |
|---|---|
| `kernel` | rbf, *linear* |
| `gamma` | *0*, .001, .0001 |
| `C` | *1*, 10, 100, 1000 |

**Table 2:** Experimental plan for the choice of the SVM's hyperparameters. In total, we performed 32 experiments. The used configuration is highlighted in italic.

| Hyperparameter | Values |
|---|---|
| `batch_size` | 16, 32, *64* |
| `units` | *256*, 512 |
| `hidden_layers` | *1*, 2, 3 |
| `dropout_rate` | .3, *.5* |

**Table 3:** Experimental plan for the choice of the MLP's hyperparameters. In total, we performed 36 experiments. The used configuration is highlighted in italic.

# 4 Results and Evaluation

On the *30-class task*, the average isolate-level accuracy of our CNN model on the test data reaches 78.3%. This result is not far from that reached in the reference paper, which is 82.2±0.3%. The confusion matrix in Figure 3 display an overview of the performance by class. Entries$(i, j)$ represents the percentage out of 100 test spectra that are predicted as class $j$, given a ground truth of class $i$. Consequently, entries along the diagonal represent the accuracies for each class.

On the *15-class task*, instead, the average isolate-level accuracy of our CNN model on the test dataset is 85.1%, confirming good performance. In comparison, the implementations of SVM and MLP achieve accuracies of 82.7% and 85.0% respectively. The confusion matrices concerning these three models are shown in Figures 4 to 6.

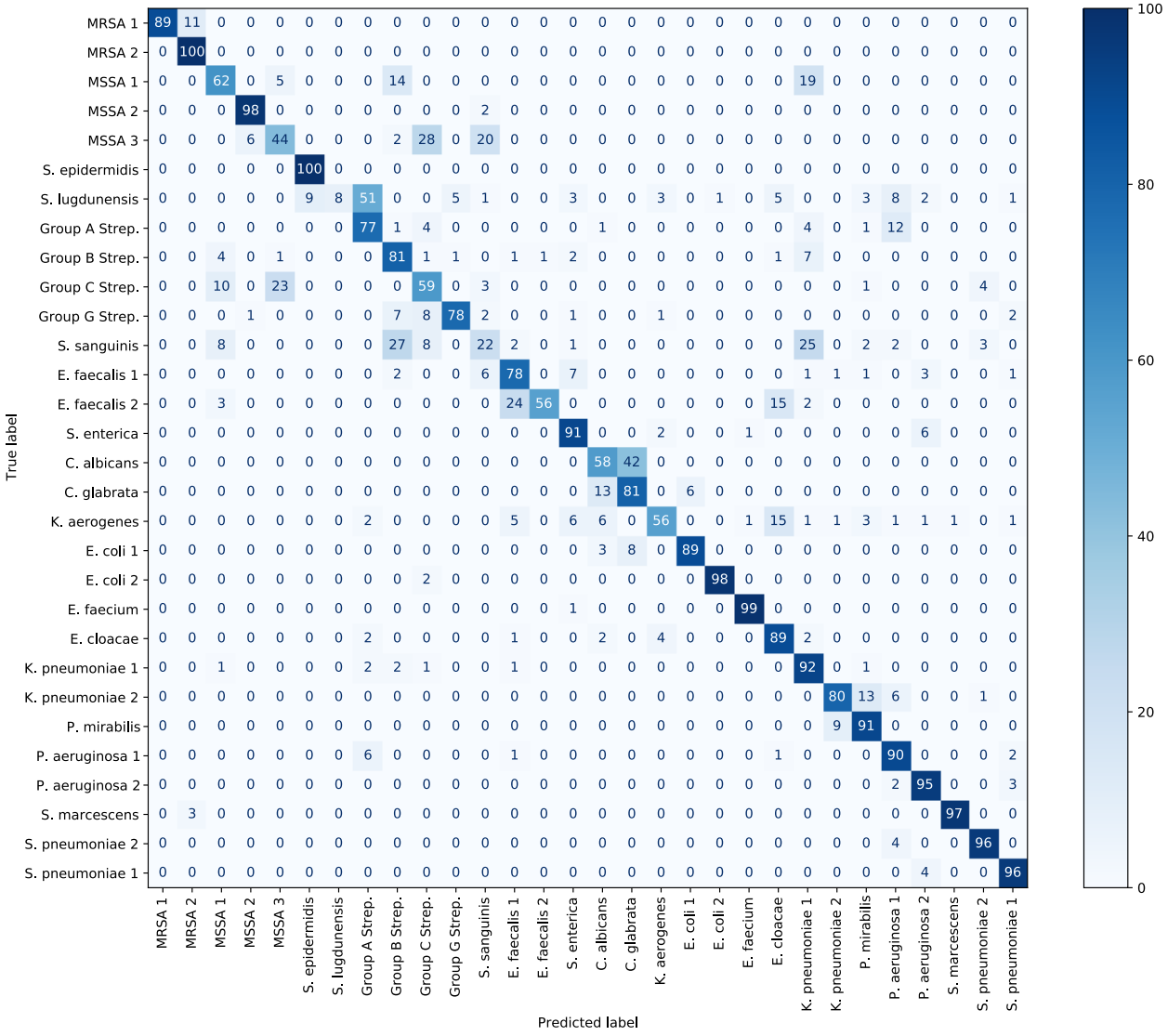


**Figure 3:** Confusion matrix for the 30 bacterial classes, with respect to CNN model.

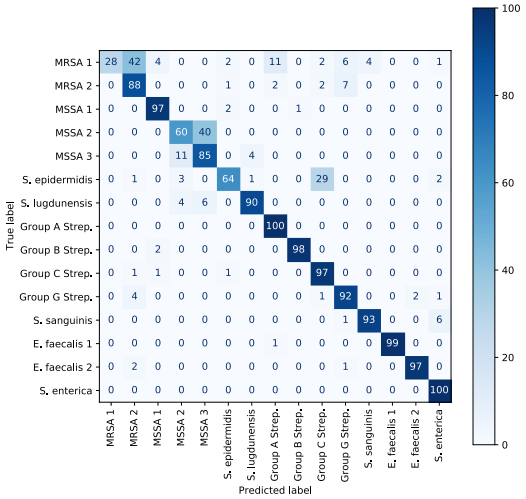


**Figure 4:** Confusion matrix for the 15 bacterial classes, with respect to CNN model.

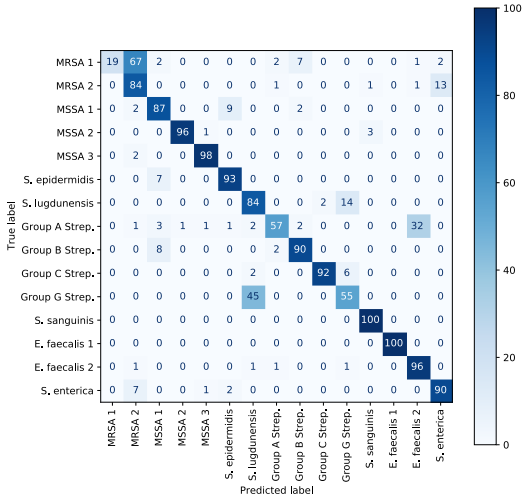


**Figure 5:** Confusion matrix for the 15 bacterial classes, with respect to SVM model.

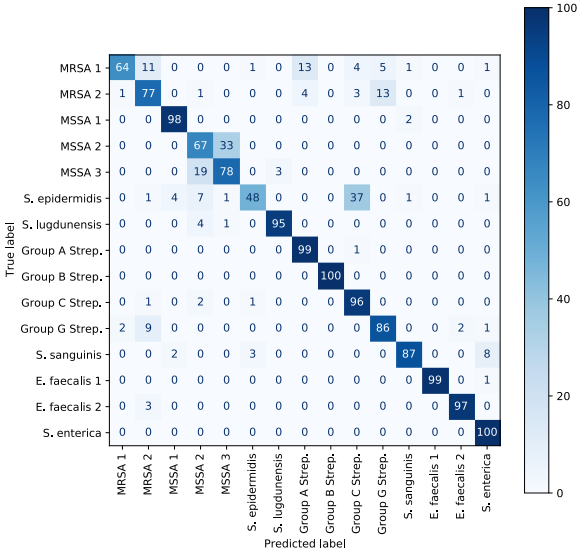


**Figure 6:** Confusion matrix for the 15 bacterial classes, with respect to MLP model.

# 5 Discussion

Classification accuracy at isolate level is the standard metric for bacterial identification, but in practice the priority is choosing the correct antibiotic to treat a patient. Broad spectrum antibiotics are often prescribed while waiting for culture results [13] and, as a consequence, over 30% of patients are treated unnecessarily [14]. The 30 bacterial species considered in this project can be arranged into groupings, based on the recommended empiric treatment. Classification accuracies can thus be condensed into a new confusion matrix grouped by *empiric antibiotic treatment*.

The average accuracy of our CNN model on the empiric treatment task becomes 94.9%, as shown in the confusion matrix in Figure 7. This value is very close to the one of the reference model, that is 97.0±0.3%. The achieved result is impressive especially if we consider that our model has about 42 000 parameters, compared to the reference architecture which has more than 1 340 000 parameters.

Considering the 15-class, the average accuracy of our CNN model on the empiric treatment task is 98%. The SVM and MLP models instead achieve 97.5% and 97.7% respectively. The corresponding confusion matrices are shown in Figures 8 to 10. It is important to note that these values are very close to each other, despite the bacterial identification accuracies were lower. This can be explained by the fact that *misclassifications* are mostly within antibiotic groupings, and thus do not affect the treatment outcome.

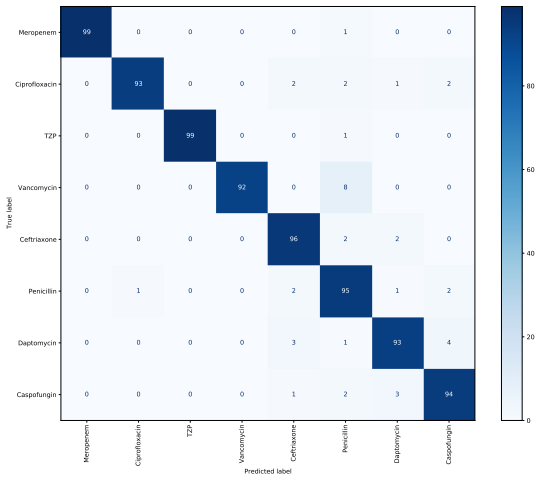


**Figure 7:** Confusion matrix for the empirical antibiotics treatment considering 30-class, with respect to CNN model.

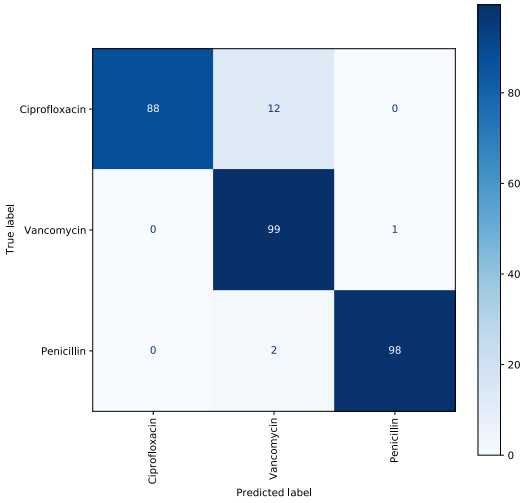


**Figure 8:** Confusion matrix for the empirical antibiotics treatment considering 15-class, with respect to CNN model.

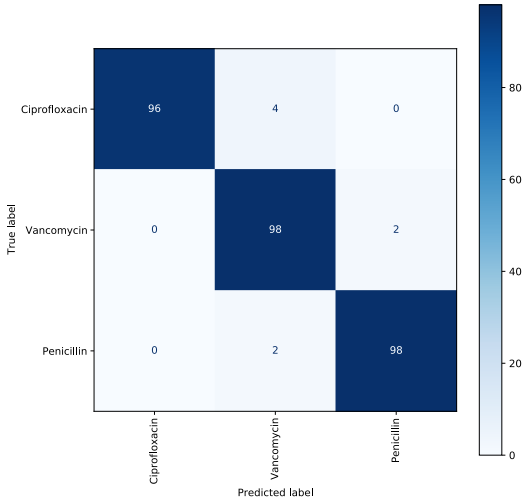


**Figure 9:** Confusion matrix for the empirical antibiotics treatment considering 15-class, with respect to SVM model.

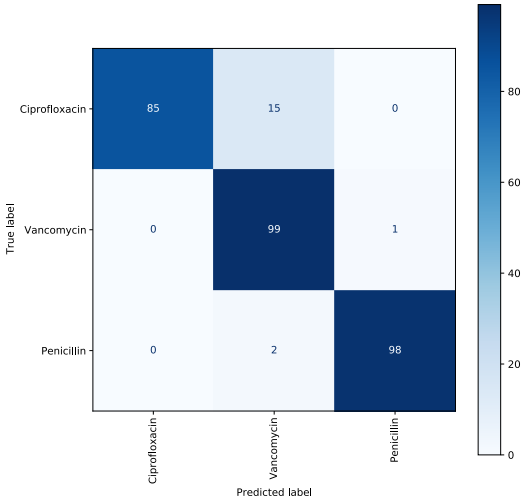


**Figure 10:** Confusion matrix for the empirical antibiotics treatment considering 15-class, with respect to MLP model.

# 6 Conclusions

In this project we apply *deep learning techniques* to Raman spectra, with the intention of identify clinically relevant bacteria and their empiric treatment. Compared to other culture-free methods, Raman spectroscopy has the unique potential to identify phenotypes that does not require specially designed labels, allowing for easy *generalizability*.

The CNN architecture presented here could rapidly identify bacterial pathogens and recommended an antibiotic treatment, achieving comparable performance with a state-of-the-art CNN, although with *less parameters*. In addition, classical machine learning techniques reported good results, even if on a smaller classification task.

We can conclude that such techniques would allow for accurate and targeted treatment of bacterial infections within hours, reducing healthcare costs, antibiotics misuse and improving patient outcomes.

# References


[1] F. Lussier, V. Thibault, B. Charron, G. Q. Wallace, and J.-F. Masson, "Deep learning and artificial intelligence methods for raman and surface-enhanced raman scattering," *TrAC Trends in Analytical Chemistry*, vol. 124, p. 115796, 2020.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[3] C. Fleischmann, A. Scherag, N. K. Adhikari, C. S. Hartog, T. Tsaganos, P. Schlattmann, D. C. Angus, and K. Reinhart, "Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations," *American journal of respiratory and critical care medicine*, vol. 193, no. 3, pp. 259–272, 2016.

[4] R. DeAntonio, J.-P. Yarzabal, J. P. Cruz, J. E. Schmidt, and J. Kleijnen, "Epidemiology of community-acquired pneumonia and implications for vaccination of children living in developing and newly industrialized countries: A systematic literature review," *Human vaccines & immunotherapeutics*, vol. 12, no. 9, pp. 2422–2440, 2016.

[5] C. M. Torio and B. J. Moore, "National inpatient hospital costs: the most expensive conditions by payer, 2013: statistical brief# 204," 2016.

[6] R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke *et al.*, "Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012," *Intensive care medicine*, vol. 39, no. 2, pp. 165–228, 2013.

[7] A. Chaudhuri, P. Martin, P. Kennedy, R. Andrew Seaton, P. Portegies, M. Bojar, I. Steiner, and E. T. Force, "Efns guideline on the management of community-acquired bacterial meningitis: report of an efns task force on acute bacterial meningitis in older children and adults," *European journal of neurology*, vol. 15, no. 7, pp. 649–659, 2008.

[8] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. Saleh, S. Ermon, and J. Dionne, "Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[9] J.-L. Vincent, J. Rello, J. Marshall, E. Silva, A. Anzueto, C. D. Martin, R. Moreno, J. Lipman, C. Gomersall, Y. Sakr *et al.*, "International study of the prevalence and outcomes of infection in intensive care units," *Jama*, vol. 302, no. 21, pp. 2323–2329, 2009.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] X. Fan, W. Ming, H. Zeng, Z. Zhang, and H. Lu, "Deep learning-based component identification for the raman spectra of mixtures," *Analyst*, vol. 144, no. 5, pp. 1789–1798, 2019.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] A. T. Society *et al.*, "Infectious diseases society of america. guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia," *Am J Respir Crit Care Med*, vol. 171, pp. 388–416, 2005.

[14] K. E. Fleming-Dutra, A. L. Hersh, D. J. Shapiro, M. Bartoces, E. A. Enns, T. M. File, J. A. Finkelstein, J. S. Gerber, D. Y. Hyun, J. A. Linder *et al.*, "Prevalence of inappropriate antibiotic prescriptions among us ambulatory care visits, 2010-2011," *Jama*, vol. 315, no. 17, pp. 1864–1873, 2016.