Identification of Discriminative Features in Raman Spectra for Antibiotic Susceptibility Testing

Ahmed Shuaibi
Department of Computer Science
Stanford University
ashuaibi@stanford.edu

Amr Saleh
Stanford University
aessawi@stanford.edu

Abstract

The technique of Raman spectroscopy can produce high-dimensional molecular fingerprints of pathogenic bacteria that can potentially be used for culture-free pathogen identification. Raman data can additionally be used to help select the most significant spectral regions as determined by their ability to accurately classify bacterial isolates. Pinpointing the most discriminative features of bacterial raman spectra will help chemists and biologists tie molecular and biological significance of the spectral regions to distinct bacterial isolates, a relationship that is not sufficiently established in general. Here, we first adapt a deep learning model (CNN) to identify 30 bacterial pathogens. Afterwards, we different feature selection techniques to isolate the subset of most significant raman spectra. Using a feature subset one fourth the size of the original input space, we are able to classify the bacteria with 92% accuracy. The feature selection techniques used help pinpoint particular regions of the spectra that may hold important biological significance.

1 Introduction

Bacterial infections a leading cause of death, taking more than 6.7 million lives each year [**2**]. Furthermore, bacterial infections are very time-consuming and costly to diagnose. The process of pathogen identification first requires culturing, a slow process that may span several days. While this process takes place, doctors often prescribe general antibiotics to patients often unnecessarily, thereby harmfully increasing antimicrobial resistance of several common bacterial pathogens[**3**]. With more than 2.8 million antibiotic-resistant infections surface each year in the U.S. alone, it is extremely important that we find less time-consuming ways to carry out pathogen identification and antibiotic susceptibility testing. We aim to use raman spectra to quickly identify bacterial pathogens using deep learning models instead carrying out the process of culturing. Rapid pathogen identification will aid in selecting an antibiotic that will work most effectively for a particular pathogen, otherwise known as antibiotic susceptibility testing. Raman spectroscopy is a light scattering technique that excites and measures the vibrational levels of a molecule. Running raman scattering on a pathogen will produce a vector of intensities over difference wavelengths of excitement for the molecule. This raman spectra of intensity readings serves as a molecular fingerprint by which we can distinguish between bacterial samples. However, raman scattering produces significant noise and thereby makes it difficult to distinguish between spectra by eye or using simple models. Thus, we will utilize a deep learning model (1-dimensional CNN with resnet architecture) adapted from Jean and Ho's [**1**] model to classify pathogens. In addition to classification from the 1000-dimensional input raman spectra, we will apply multiple feature selection techniques to signify which spectral regions are most significant in the classification process. Using filter feature selection techniques, we were able to classify the 30 bacterial isolates using only 250 of the 1000 original features with 91% accuracy. Using a wrapper feature selection technique known as Ant Colony Optimization, we were able to marginally improve upon the filter feature selection approaches, achieving 92% accuracy.

Note: This project is also used for CS231N. This project expands more on the feature selection and the hyperparameter optimization of the CNN model for this particular task while the other project expands more on the CNN model architecture and ablative analysis.

CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

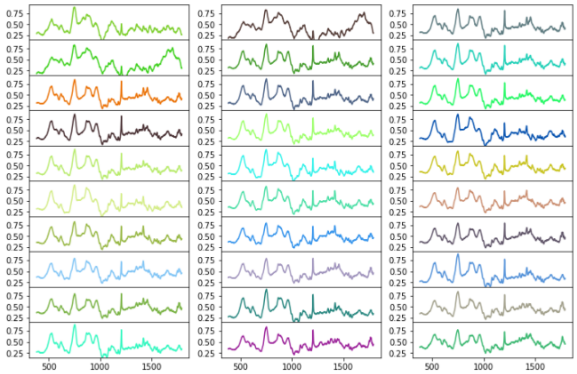


Figure 1: Average of all 2000 spectra for each bacterial isolate yielding 30 total raman spectra. Presented to visualize similarity between all 30 common bacterial isolates.

2 Related Work

One paper we explored was "Identification and characterization of colorectal cancer using Raman spectroscopy and feature selection techniques" by Li et. al. This work aims to pinpoint 5 bands in Raman spectra that are most significant when classifying colorectal cancer. An SVM model was used for the classification and was combined with the technique of Ant Colony Optimization for feature selection. Ant Colony Optimization Band Selection is a technique that selects the optimal subset of features for a classification task through repeatedly subsampling and evaluating distinct collections of features. This paper additionally presents direct links between Raman peaks and molecular and cellular alterations associated with malignant transformations that are ascertained through a deeper analysis of the selected significant feature regions by biology and medical experts.

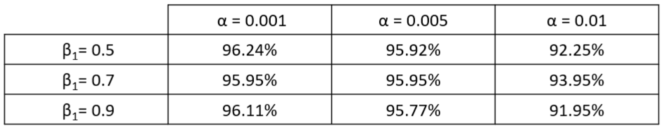
Lastly, we explore an unsupervised feature extraction method for ant colony optimization. In the work "An unsupervised feature selection algorithm based on ant colony optimization", the authors first analyze a variety of feature selection approaches. They explain the distinction between filter feature selection, embedded feature selection, and wrapper feature selection approaches, highlighting the advantages and disadvantages in a variety of contexts. They conclude with an unsupervised implementation of ant colony optimization that can easily be coupled with any classification technique. Their implementation of ant colony optimization effectively learns subsets of features with minimal intra-subset correlation, optimal for tasks with multiple highly correlated features. We opt to implement the ant colony optimization technique explained in this article since we expect it to synergize well with the raman spectra dataset. Consecutive frequencies in the raman data tend to correlate highly with one another. As such, we implemented this unsupervised feature extraction method for ant colony optimization and used it in combination with a CNN for pathogen classification to obtain the best expected results.

3 Dataset and Features

The data is provided by The Dionne Research Group at Stanford and is not currently publicly available. The raw dataset is composed of 62,000 Raman spectra. There are exactly 2,000 spectra for each of 31 bacterial and yeast isolates. Two of these isolates are isogenic and are thereby extremely similar. The classification and analysis in this project will not consider one of isogenic bacteria in order to develop a model that differentiates between the 30 common bacteria for clinical applications. Each of the samples span a spectral range of $381.98-1792.4cm^{-1}$. The Raman device used for data collection utilized spectral resolution of $1.41cm^{-1}$, thus resulting in 1000 intensity readings for each of the spectra.

Two steps were taken to preprocess and standardize this data. First, the spectra were individually normalized to have intensity values ranging from 0 to 1. To correct background signals from autofluourescence in the data collection process, the spectra were individually corrected using a 5th order polynomial to adjust the intensity measurements. For each class of bacteria, I averaged the spectral readings among all 2000 sample and display the visualization below such that we can depict the similarity in spectra among the bacteria.

Given $60,000$ total input samples, we decided to first separate 10% of the data to use as the test set. The test set was thus composed of $6,000$ training samples. From here, we decide to use 5-fold cross validation and thereby segment the remaining $54,000$. Each fold was composed of $10,800$ samples. As such, we trained on $43,200$ samples and evaluated on $10,800$ at a time in the cross-validation process more explicitly outlined in the Methods section below.



	α = 0.001	α = 0.005	α = 0.01
β_1= 0.5	96.24%	95.92%	92.25%
β_1= 0.7	95.95%	95.95%	93.95%
β_1= 0.9	96.11%	95.77%	91.95%

Figure 2: Hyperparamter optimization of different learning rate and β_1 values for the Adam optimizer of the CNN

4 Methods

We aim to find the subset of most significant features and thus apply four feature selection approaches. The first three approaches are known as filter feature selection techniques, in which the features are treated independently and are ranked based off some discriminative score in a manner decoupled from the classification task at hand. The final approach Ant Colony Optimization, and is known as a wrapper feature selection method that ties directly with the classification method. To assess the effectiveness of some subset of features in classification, we apply a 1-D convolutional neural network to classify among the 30 bacterial pathogens. The subset of features that yields the greatest accuracy in classification is deemed to be the most significant and discriminative. The CNN model is adapted from Jean and Ho's model [**1**] that effectively uses all 1000 input features of the raman spectra to use only a subset of the features. We choose to find feature subsets of size 250 such that we exhibit a four-fold decrease in the number of features in the input space. This hyerparameter is later optimized in the ablative analysis section below. This section will detail the 1-D convolutional neural network architecture used for classification of the bacterial pathogens in addition to the feature selection methods used.

4.1 CNN Architecture

The baseline CNN model is adapted from a resnet architecture composed of an initial convolutional layer with 64 filters, six residual layers, and one fully connected layer pictured in figure 3. Each residual layer is composed of four convolutional layers each with 100 filters. Skip connections between the first and last layer of each residual layer are added to mitigate vanishing gradients that were initially encountered in training and thereby allow for better gradient flow overall. An Adam optimizer was used with a learning rate of 0.001 and beta values of 0.5 and 0.999. We experimented with three different values of learning rate and three different values of the β_1 value in the Adam optimizer, yielding 9 different model runs. The learning rates tried were $0.001, 0.005, 0.01$ and the β_1 values tried were $0.5, 0.70, 0.9$. We ran these 9 different models and recored the classification accuracy using the standard raman spectra as input (with all 1000 input features). The accuracies obtained with the CNN model using the different hyperparamters are listed in Figure 2.

Given the results of the hyperparameter optimization, we selected $\beta_1 = 0.5$ and the learning rate $\alpha = 0.001$. Although it was more computationally costly, we opted to run the model using 5-fold cross validation. Given that this task will potentially have significant clinical implications, we need to ensure that we are robust in our evaluation. After separating 10% of the data as the test split, the remaining data was dividing into five folds. One of these folds was selected as the validation set and the remaining four folds acted as the training set. We trained the model on each of the distinct validation folds and thereafter predict the model's accuracy on the test set.

4.2 Filter Feature Selection Methods

These feature selection approaches identify significant features through analyzing inherent properties of the data. We focused on univariate feature selection approaches that assume independence of the input features and measure feature significance through different evaluation criteria [**1**]. We explored three univariate feature selection approaches using ANOVA, χ^2, and mutual information.

The χ^2 test effectively measures the dependence between stochastic variables. Thus, we evaluate the χ^2 statistic for each of the 1000 original spectral features and select the 250 with the highest value since they are more likely to be relevant for pathogen classification.

The ANOVA test assesses if there is significant difference in a spectral feature among the 30 different bacterial classes. We similarly find the 250 features with the greatest F-score after calculating the statistic for each of the 1000 features [**9**].

Mutual information effectively serves as a measurement of the mutual dependence of the bacterial classes on the spectral features. It is defined as the KL divergence between the joint distribution and the product of the marginals [**10**]:

$$I(X;Y) = KL(P_{X,Y} \| P_X \otimes P_Y)$$

We have $I(X;Y) \leq I(X;X) = H(X)$, where $H(X)$ is the entropy of X. Mutual information thus effectively measures the dependency between variables in a manner that we can assess the most significant features as ones with the highest scores.

4.3 Ant Colony Optimization

Ant Colony Optimization is a technique devised by Dorigo et. al to solve computational problems [**1**]. The method is based on the behavior of ants, in which they cooperatively work to find optimal travel paths through substances known as pheromones. In essence, ants deposit the chemical substances of pheromones to communicate with one another that inherently dissipates over time. The intensity of the pheromones in a location signifies to ants the importance or utility of a particular path. Ants tend to follow paths with a greater concentration of pheromones. With regards to feature selection, ants are assigned to different subsets of features and pheromone concentrations are updated based on the significance of a feature subset according to classification accuracy.

To perform ant colony optimization we loop through these steps below:

1. Generate artificial ants and assign some subset of features to each of the ants. Features are originally probabilistically assigned based on what is known as the global pheromone trail.
2. Evaluate the utility and significance of the ants and their component features.
3. Update the global pheromone trail based on the significance evaluation results.

Each artificial ant is assigned to a distinct subset of features from the spectra, determined through the transition probability function below for each spectral feature i:

$$p_i(t) = \frac{(\tau_i(t))^\alpha \eta_i^\beta}{\sum_i (\tau_i(t))^\alpha \eta_i^\beta}$$

In which α and β are weighting factors, $\tau_i(t)$ represents the pheromone trail magnitude at time t for the feature i and η_i represents the local information of feature i. After exploring β values in the range $(0.8, 1.0)$, we utilize a value of 1 due to its greatest performance in selecting an optimal subset of features. The value of $\tau_i(0) = 1$ for all features and pheromone is updated with:

$$\tau_i(t+1) = \rho\tau_i(t) + \nabla\tau_i(t)$$

in which ρ is a constant between 0 and 1 and $\nabla\tau_i(t)$ is related to the classification accuracy of artificial ants. These steps are repeated until convergence to find the optimal subset of features.

5 Experiments/Results/Discussion

We perform four distinct experiments in line with the filter feature selection approaches and the convolutional model above. The first three approaches include applying the univariate feature selection techniques of Anova, χ^2, and Mutual Information to input raman spectra and selecting the corresponding top 250 features using each approach. Thereafter, we train the CNN model three distinct times using the 3 different subsets of 250 features and evaluate their accuracies on the test set along with depicting the regions of importance visually on the input spectra. The fourth model consists of implementing ant colony optimization tied with the convolutional neural network model as described above and finding the 250 most discriminative features for classification.

5.1 Filter Feature Selection Results

Features of the input spectra were ranked based off their scores from these statistical techniques of χ^2, Anova, and Mutual Information and we visually depict the 250 most significant features of each approach in Figure 3.

With these subsets of selected features, I ran the CNN model to evaluate their classification accuracies. The classification accuracy when using all 1000 original features was **96%**. Comparatively, the classification accuracies using the subsets of 250 significant features from each of the above univariate statistical test approaches were:

- χ^2: **90.2%**
- Anova: **91.1%**
- Mutual Information: **90.9%**

Among these univariate approaches, the Anova tests and the correlating subset of signficant features produced the best results in retaining classification accuracy.

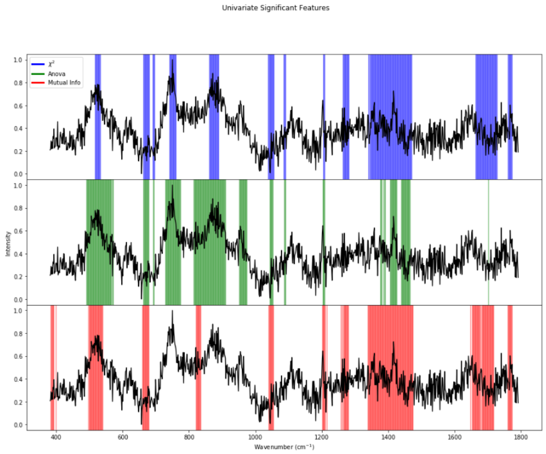


Figure 3: Top 250 features selected using the χ^2, Anova, and Mutual Information univariate statistical tests.

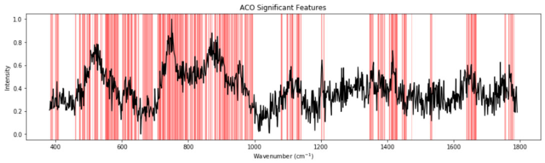


Figure 4: Visualization of most significant features obtained through ACO-CNN

5.2 Ant Colony Optimization Feature Selection Results

After performing the Ant Colony Optimization technique outlined above, we are left with every feature and its final pheromone value. We thereby select the 250 features with the greatest pheromone values and visually depict them in Figure 8. Relative to the filter feature selection approaches, we notice that ACO yields more distributed features. Specifically, we see less bands of consecutive features ascertained to be significant. Since consecutive features in the Raman spectra tend to be slightly correlated, ACO effectively obtains subsets of features with minimized correlation.

With this subset of selected features, I ran the CNN model to evaluate its classification accuracy. Using the 250 ACO-CNN selected features, the model yielded **92.0%** accuracy. A less correlated discriminative subset of the original features thereby yields better performance relative to filter feature selection approaches.

Finally, we produce a confusion matrix depicting the ACO-CNN classifications. This confusion matrix reveals that most misclassifications occur between the strains spanning from E. coli. to S. marcescens in the matrix. We notice that same antibiotic is utilized against these strains. While we may not be able to discern between the strains on an isolate level, we can properly classify this group of isolates from others at the antibiotic level confidently.

We additionally created a confusion matrix for the model that uses all the input features. We notice that this original model does not misclassify the section of strains spanning from E. coli. to S. marcescens as greatly as the model that utilizes a subset of the features.

6 Conclusion/Future Work

A deep learning model allowed for the accurate classification of 30 common bacterial isolates, yet left much to be desired with the model's interpretability. We applied filter and wrapper feature selection techniques to obtain an interpretable subset of features that were deemed most discriminative and significant in bacterial pathogen classification. Relative to the accuracy of 96% using all original spectral features in classification, a subset of 250 features produced classification accuracy of 92%. These results promise a rapid acceleration in the process of pathogen identification and antibiotic susceptibility testing due to a lower dimensional and more targeted input space. Additionally, these selected features allow us to tie the spectral ranges of significant features back to

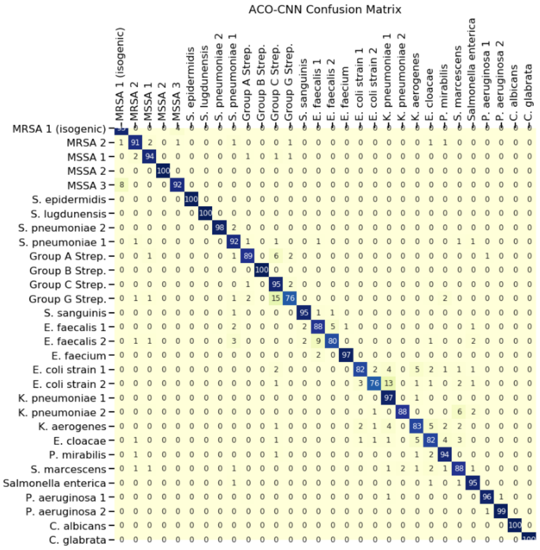


Figure 5: Confusion matrix for model that uses 250 most significant ACO-CNN selected features.

biological and molecular significance in a manner than can accelerate classification and antibiotic susceptibility testing for clinical use.

7 Contributions

This project was done in collaboration with The Dionne Group at Stanford. Chi-Sing Ho and Neal Jean produced the baseline CNN model. Chi-Sing Ho collected the dataset used. Amr Saleh and Jennifer Dionne produced the original idea. Amr Saleh advised and supervised the project. Ahmed Shuaibi adapted and fine-tuned the CNN model for use with feature subsets, implemented and applied the feature selection approaches, and wrote the paper.

References

[1] Ho, Chi-Sing and Jean, Neal and Hogan, Catherine A and Blackmon, Lena and Jeffrey, Stefanie S and Holodniy, Mark and Banaei, Niaz and Saleh, Amr AE and Ermon, Stefano and Dionne, Jennifer *Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning*. Nature Communications, 2019.

[2] *Biggest Threats and Data* Centers for Disease Control and Prevention, 2020. https://www.cdc.gov/drugresistance/biggest-threats.html

[3] Fleming-Dutra, Katherine E and Hersh, Adam L and Shapiro, Daniel J and Bartoces, Monina and Enns, Eva A and File, Thomas M and Finkelstein, Jonathan A and Gerber, Jeffrey S and Hyun, David Y and Linder, Jeffrey A and others *Prevalence of inappropriate antibiotic prescriptions among US ambulatory care visits, 2010-2011*. Jama, 2016.

[4] M. Gevrey, I. Dimopoulos, S. Lek. *Review and comparison of methods to study the contribution of variables in artificial neural network models*. Ecological Modelling, 2003, v.160, pp. 249-264

[5] Garson, G.D., 1991. *Interpreting neural network connection weights*. Artificial Intelligence Expert 6, 47/51

[6] M. Dorigo and L. M. Gambardella *Ant colony system: a cooperative learning approach to the traveling salesman problem*. IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 53-66, April 1997.

[7] Tabakhi, Sina and Moradi, Parham and Akhlaghian, Fardin *An unsupervised feature selection algorithm based on ant colony optimization* Engineering Applications of Artificial Intelligence, 2014.

[8] Chi-Square Statistic: How to Calculate It - Distribution *https://www.statisticshowto.com/probability-and-statistics/chi-square/*

[9] F-Test *https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/f-test/*

[10] Sklearn Mutual Information Classification *https://scikit-learn.org/stable/modules/generated/sklearn feature_selectionmutual_info_classif.html*