C-1045

Application of Supervised Machine Learning Algorithms for Rapid Identification of MRSA PFGE Strain Types

Michael Bernauer, PharmD Candidate, Jenny Shroba, PharmD Candidate, Jessica Lewis, PharmD Candidate, Renée-Claude Mercier, PharmD, BCPS
[1]University of New Mexico College of Pharmacy; Albuquerque, NM, USA

Michael L. Bernauer, PharmD Candidate
UNM College of Pharmacy
1 University of New Mexico
Albuquerque, NM 87131-0001
Phone: (505) 933-4359
Email: bernauer@salud.unm.edu

Abstract

Introduction: Pulsed-field gel electrophoresis (PFGE) is a molecular typing method used in epidemiologic surveillance of MRSA. PFGE relies on electrophoretic migration of restricted DNA fragments, resulting in characteristic fingerprints which can be used to classify organisms. Current methods of analysis, such as Gel Compar II and BioNumerics rely on unsupervised hierarchical clustering algorithms to group organisms based on pairwise similarity. These methods are labor intensive, often requiring a significant amount of user intervention and oversight. This study presents an automated approach to PFGE typing to reduce the amount of user involvement and time required for analysis.

Methods: A total of 70 gel electrophoresis images were obtained from previous PFGE experiments conducted according to protocol. Lanes were extracted from the raw images using k-means clustering, resulting in 1067 fingerprints. Transformations applied to each of the fingerprints include; normalization, alignment and background subtraction. The training set was created manually by labeling each fingerprint according to PFGE type. Several fingerprints were removed due to the presence of artifacts, resulting in 843 (17% USA100, 72% USA 300, 11% USA400) for the final analysis. Principal component analysis was used to reduce dimensionality by selecting the top 20 components. The support vector machine (SVM) was implemented using 10-fold cross validation. Accuracy of the other classification algorithms was assessed by splitting the data 70:30 training to test. Hyperparameters for k-nearest neighbors (kNN), random forest (RF) and the artificial neural network (aNN) were selected via grid search to yield the highest test set accuracy.

Results: The aNN and SVM were the highest performing algorithms with overall classification accuracy of 0.9004, p = 1.15×10^{-12}, 95% CI [0.8565, 0.9345] and 0.8845, p = 1.33×10^{-5} , 95% CI [0.8383, 0.9212] respectively. Both kNN and RF algorithms performed similarly with classification inaccuracies of 0.8606, p = 5.38×10^{-8} , 95% CI [0.8114, 0.9009] and 0.8685, p = 8.34×10^{-9} , 95% CI [0.8203, 0.9077], respectively.

Conclusions: Supervised learning algorithms such as SVM and aNN provide feasible alternatives to traditional hierarchical clustering methods for determination of MRSA PFGE strain type.

Background

- PFGE is the molecular typing method of choice for distinguishing strains of organisms and monitoring for infectious disease outbreaks.[1]
- Current methods of PFGE analysis rely on the use of hierarchical clustering algorithms to match organisms based on pairwise fingerprint similarity.[2]
- To our knowledge this is the first study that implements a machine learning system for the identification/classification of MRSA PFGE strain type using raw PFGE gel electrophoresis images.

Research Objectives

Primary Objective:

- Assess the performance of various machine learning algorithms including; *k*-nearest neighbors (kNN), random forest (RF), support vector machines (SVM) and artificial neural networks (aNN) for classification of MRSA PFGE strain types.

Methods

Image collection:

- A total of 70 gel electrophoresis images were obtained from previous PFGE experiments conducted according to protocol, as described by McDougal et al.[3]
- Images contained PFGE fingerprints for USA100 (17%), USA300 (72%) as well as USA400(11%) strain types.
- Images were obtained as TIFF files and were 480x640 pixels in size.

Methods Continued

Image processing:

- A series of transformations were applied to the original 8-bit grayscale images to increase signal-to-noise ratio (Figure 1).
- Gamma of each image was adjusted using a gamma value of 3.
- Morphological operations were applied in R v.3.2.1 using the *EBImage* package.
- Pixel intensities were normalized to μ=0 and σ=1.

Lane detection:

- Lane detection was performed using *k*-means clustering.
- Binary images were produced by applying a threshold value of 0.1 to the transformed images.
- After thresholding, *x*-coordinates for each of the band pixels were extracted and clustered to determine the centroid location for each lane.
- Centroids were used to extract the corresponding lane from the original, unadulterated grayscale image (Figure 2).
- Fingerprints were extracted as 21x480 pixel images.

Fingerprint enhancement:

- Extracted fingerprint images were enhanced to remove noise using the same set of transformations applied to the original grayscale images.

Background subtraction:

- Background noise was subtracted from the fingerprint images using a linear regression method in which the appropriate amount of background to subtract was determined as a function of pixel position (Figure 3).

Fingerprint alignment:

- Fingerprints were aligned by applying an offset to position the first band at the top of the fingerprint image.
- Band positions were determined by summing the pixel intensities across rows and thresholding for those in the bottom 40th percentile.

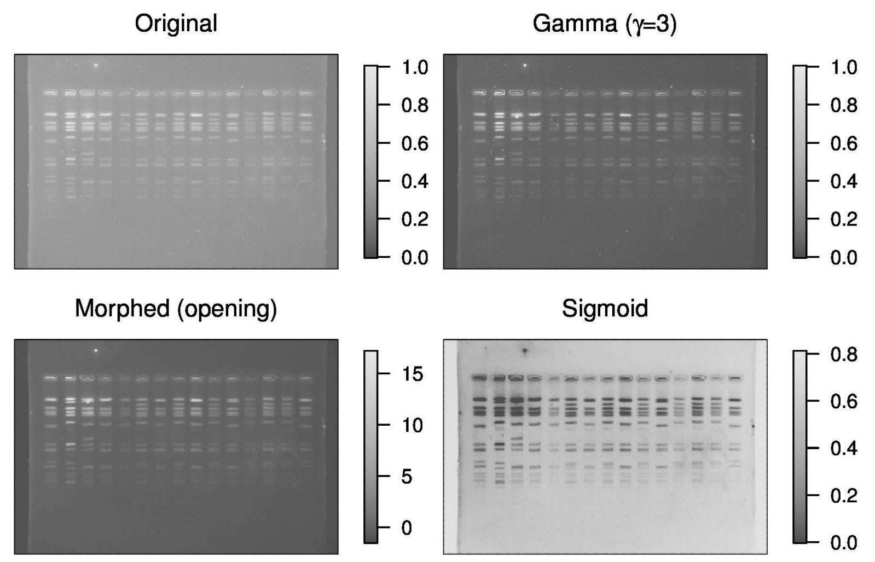


Figure 1: Gel images after applying transformations.

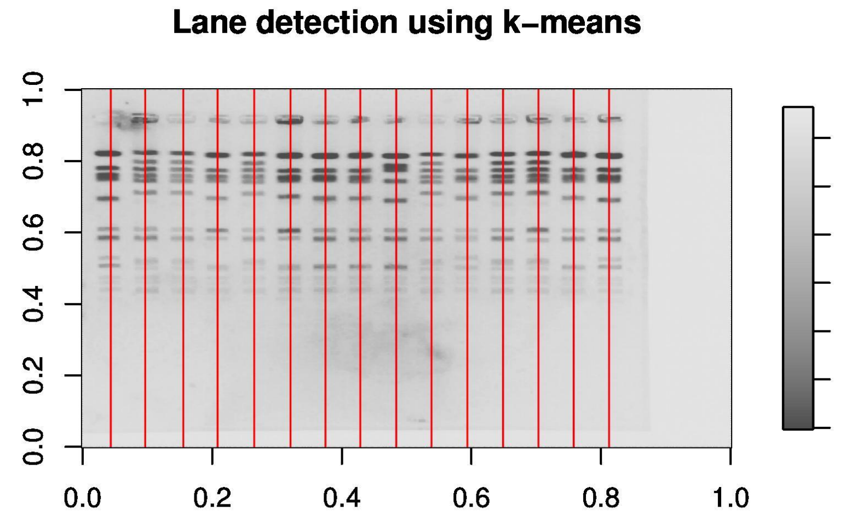


Figure 2: Lane detection using centroids from *k*-means. Red lines indicate lane centers and centroid locations.

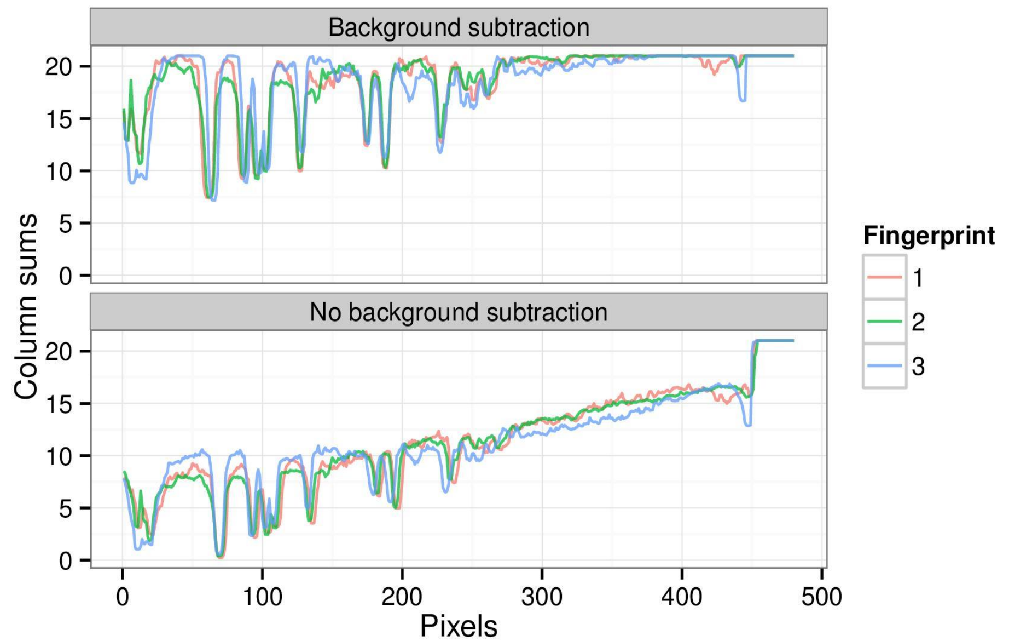


Figure 3: Lane profiles before background subtraction (bottom) and after (top).

Training data:

- Training data was constructed by unrolling each of the 480x21 element fingerprint matrices into an 10080 element vector resulting in a 843x10080 element matrix.
- PCA was used to reduce dimensionality (Figure 4). The top 20 components were selected resulting in an 843x20 element matrix as the final training set.

Model training:

- The training data was split 70:30 with n=590 in the training set and n=253 in the test set.
- The kNN algorithm was trained using two-fold cross validation. A *k*-value of 5 was selected according to highest test set accuracy (86.06%).
- Hyperparameters for RF were determined using two-fold cross validation. Final number of trees used in analysis was 580 with a test-set accuracy of 88.05%.
- Hyperparameters for both SVM and aNN algorithms were derived using grid-search.

Results

- As seen in Table 1, SVM and aNN outperformed other classifiers with respect to overall accuracy, specificity and sensitivity.
- The aNN algorithm had the greatest performance with an overall classification accuracy of 90.40%, 95% CI [85.65, 93.45].
- The SVM performed slightly worse with an overall classification accuracy of 88.45%, 95% CI [83.83, 92.12].
- Both classifiers struggled with the classification of USA400 strain types with reported sensitivities of 50.0% and 57.1% respectively.
- Specificity between the SVM and aNN classifiers remained modest with respect to each strain type.

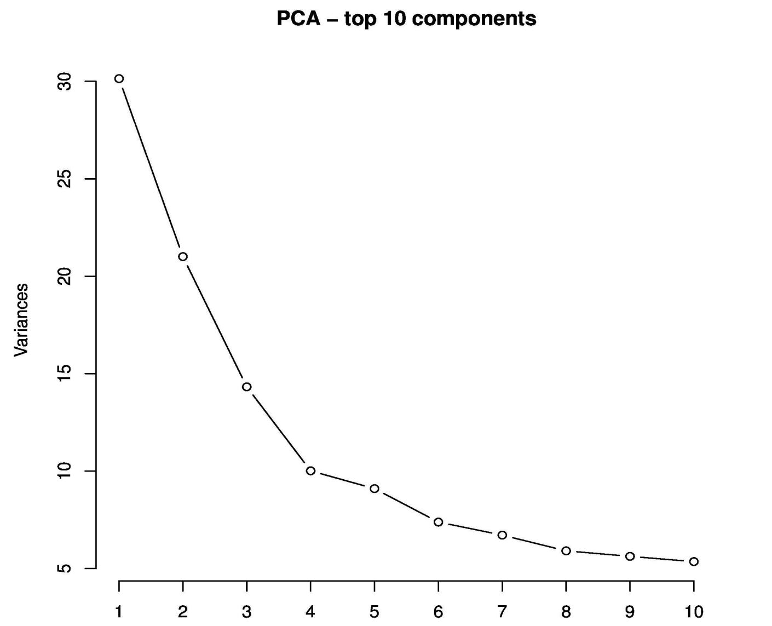
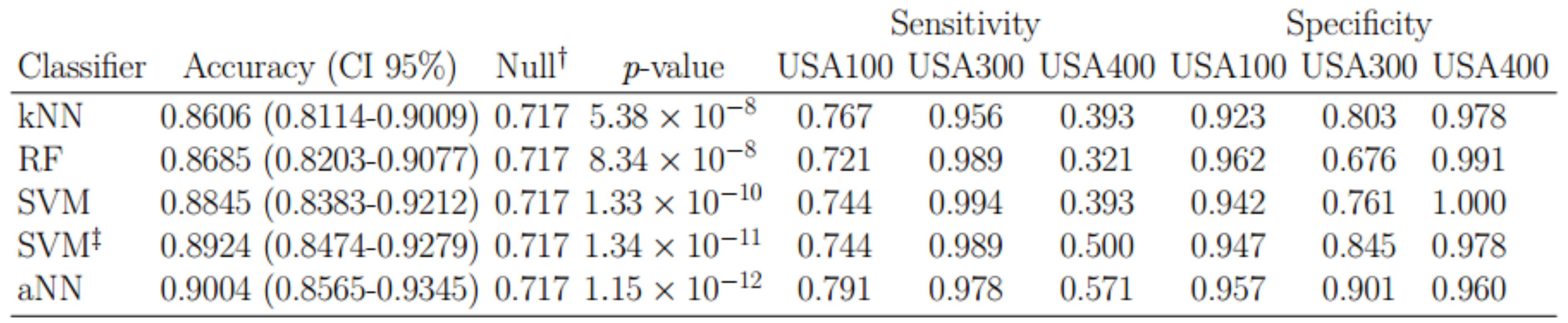


Figure 4: Variance explained by the top ten components of the training set.



				Sensitivity			Specificity		
Classifier	Accuracy (CI 95%)	Null†	*p*-value	USA100	USA300	USA400	USA100	USA300	USA400
kNN	0.8606 (0.8114-0.9009)	0.717	5.38×10^{-8}	0.767	0.956	0.393	0.923	0.803	0.978
RF	0.8685 (0.8203-0.9077)	0.717	8.34×10^{-8}	0.721	0.989	0.321	0.962	0.676	0.991
SVM	0.8845 (0.8383-0.9212)	0.717	1.33×10^{-10}	0.744	0.994	0.393	0.942	0.761	1.000
SVM‡	0.8924 (0.8474-0.9279)	0.717	1.34×10^{-11}	0.744	0.989	0.500	0.947	0.845	0.978
aNN	0.9004 (0.8565-0.9345)	0.717	1.15×10^{-12}	0.791	0.978	0.571	0.957	0.901	0.960

Table 1: Results from classification algorithms. †Null information rate. ‡Performance from retuned SVM classifier using different hyperparameters that was generated after abstract was submitted.

Limitations

- Relatively small number of training examples (n=843).
- Unbalanced dataset; USA100 (17%), USA300 (72%), USA400 (11%).
- Retrospective study design.

Discussion and Conclusion

- Classification performance appeared to suffer with respect to USA100 and USA400 strain types.
- Performance of classification SVM and aNN classification algorithms is expected to improve with increasing training examples.
- In spite of the above limitations, the results from this study suggest that algorithms such as SVM and aNN may be used to rapidly classify infectious organisms based on PFGE images.

References and Disclosure

Authors of this presentation have no disclosures concerning possible financial or personal relationships with commercial entities that may have a direct or indirect interest in the subject matter of this presentation.

1. T.J. Barrett, P. Gerner-Smidt, and B. Swaminathan. Interpretation of pulsed-field gel electrophoresis patterns in foodborne disease investigations and surveillance. *Foodborne pathogens and disease* 2006;3(1):20-31.
2. W. Zou, H. Tang, W. Zhao, J. Meehan, S.L. Foley, W.J. Lin, H.C. Chen, H. Fang, R. Nayak, and J.J. Chen. Data mining tools for salmonella characterization: application to gel-based fingerprint analysis. *BMC Bioinformatics* 2013;14(Suppl 14):S15.
3. L.K. McDougal, C.D. Steward, G.E. Killgore, J.M. Chaitram, S.K. McAllister, F.C. Tenover. Pulsed-field gel electrophoresis typing of oxacillin-resistant *Staphylococcus aureus* isolates from the United States: establishing a national database. *J Clin Microbiol.* 2003;41:5113-5120.