PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Classification of Arcobacter species using variational autoencoders

Valery Patsekin, Stephen On, Jennifer Sturgis, Euiwon Bae, Bartek Rajwa, et al.

Valery Patsekin, Stephen On, Jennifer Sturgis, Euiwon Bae, Bartek Rajwa, Aleksandr Patsekin, J. Paul Robinson, "Classification of Arcobacter species using variational autoencoders," Proc. SPIE 11016, Sensing for Agriculture and Food Quality and Safety XI, 1101608 (30 April 2019); doi: 10.1117/12.2521722

SPIE.

Event: SPIE Defense + Commercial Sensing, 2019, Baltimore, Maryland, United States

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Classification of *Arcobacter* species using variational autoencoders

Valery Patsekin[1], Stephen On[2], Jennifer Sturgis[1], Euiwon Bae[3], Bartek Rajwa[4], Aleksandr Patsekin[5], and J. Paul Robinson [1,6]

[1]Department of Basic Medical Sciences, [2]Department of Wine, Food and Molecular Biosciences, Lincoln University, Lincoln, 7647, New Zealand, [3]School of Mechanical Engineering, [4]Bindly Bioscience Center, [5]Computer Information Technology, Purdue Polytechnic Institute, [6]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, Indiana 47907 USA

ABSTRACT

Arcobacter (formerly classified as *Campylobacter* spp.) are curved-to helical, Gram-negative, aerobic/microaerobic bacteria increasingly recognized as human and animal pathogens. In collaboration with Lincoln and Purdue University, we report the first experimental result of laser-based classification method of bacterial colonies of these species. This technology is based on elastic light scatter (ELS) phenomena where incident laser interacts with the whole volume of the colony and generates a unique fingerprint laser pattern. Here we report a novel development and application of deep learning algorithm to classify the scatter patterns of *Arcobacter* species using variational autoencoders (VAE). VAE creates set of normal distributions. Each of these distributions are responsible for certain properties of the original images. We used VAE to identify features in the features space for several hundred images which includes size of the colony based on scatter size, intensity of the image, and, the number of rings within the image, and so on. Thus each sample within our image database can be coded with sets of features that facilitates fast preliminary search for similar images allowing clustering of similar patterns in feature space. In addition, such initial selection could assist in identifying non-bacterial scatter patterns (i.e. bubbles or dust spots in the agar), or doublets where two colonies are overlapping during the acquisition time thus removing non-biological artifacts prior to analysis. An interesting result was that while VAE created far more realistic synthetic images closer to the original image, a simple autonencoder resulted in better cluster separation.

Keywords: Variational autoencoders, A*rcobacter*, classification, unsupervised classification, feature extraction

1. INTRODUCTION

We have previously presented a rapid technology for obtaining preliminary results for identification of bacterial colonies on agar plates based on light-scatter patterns[1]These studies focused on the importance of rapid prescreening methods when it comes to food poisoning or bioterrorism prevention. However, a difficulty with this approach is the stage of classification of the obtained images based on Zernike moments as a feature extracting technique. This allowed us to approach correct classification of around 84%. Further progress in this technology was described by Dundar, Kou, Zhang, He, and Rajwa [2]. Different classification models were applied to compare their accuracy. Surprisingly, K-Means-based representation demonstrated nearly perfect classification accuracy – 97.89% on the dataset of four bacteria classes. As a follow-up study, we proposed application of deep-learning models with large-scale datasets to see if the performance can be maintained or improved when challenged with increased sample variety. However, all these learning techniques are suitable only for end-to-end classification with training in a supervised manner. From a practical point of view, this technology is not capable of identifying novel strains of bacteria due to the nature of current supervised classification models [3]. One of our goals is to construct a statistical learning model for automated analysis and labeling of biological datasets using an pre-trained, unsupervised feature and manifold learning paired with subsequent clustering in order to determine the likely number of biologically meaningful classes. The resultant model should be able to discover relevant features and use the learned dimensionality reduction to identify emerging classes in the data, and in consequence, detect defective or anomalous samples without supervised training or class number knowledge. It can

Sensing for Agriculture and Food Quality and Safety XI, edited by Moon S. Kim, Bryan A. Chin, Byoung-Kwan Cho,
Proc. of SPIE Vol. 11016, 1101608 · © 2019 SPIE · CCC code: 0277-786X/19/$18 · doi: 10.1117/12.2521722

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

serve as a tool for preliminary analysis of provided samples to identify possible novelty and prevent further misclassification by currently utilized methods.

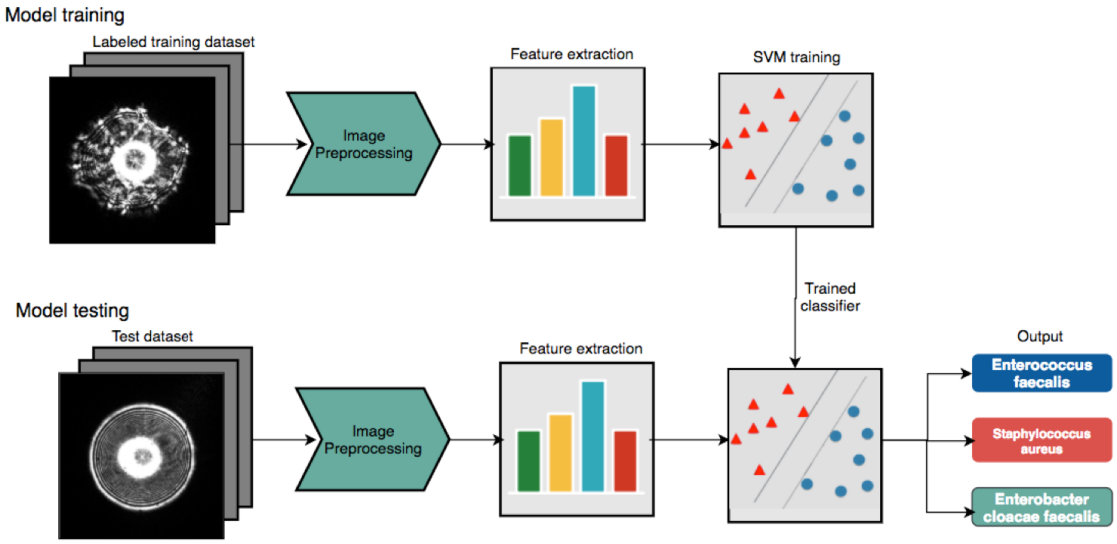

Figure 1. Conventional method for classification of scatter patterns from bacterial colonies is shown. This method is a supervised training since verified training set is required. Feature extraction by Zernike polynomials followed by support vector machine (SVM) based training is used for classification. Data sets are then processed in a similar manner with the SVM creating identified clusters providing organism identification.

An area recently defined has been termed “deep learning”, an approach that is described as a solution to "allow computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts"[4]. According to these authors, the term "deep" comes from the idea of graphs representing this structure which has many layers. This is a kind of machine learning, which in turn is a common type of Artificial Intelligence (AI). A deep learning approach is intended to tackle intuitive tasks from everyday life. The difficulty of tasks such as speech or object recognition is that they cannot be formulated as a set of rigorous abstract rules. These problems are usually translated to the computer field through representation learning. The latter approach is designed to convert raw data, such as images or voice, into a set of descriptive features which are represented as multidimensional vectors [4].

Recently a group from China proposed a method for application of Deep Convolutional Neural Networks (Deep CNN) to classification and segmentation of histopathology images [5]. To overcome the issue of scarce datasets, researchers adopted deep CNN model provided by Cognitive-Vision team described in one of the ImageNet contest-related publications [6]. The model was trained using publically available labeled dataset ImageNet. Despite the fact the model was trained for natural images, scientists were able to apply it to biological imaging.

Similarly, we proposed to enhance our automated analysis by using variational autoencoders (VAE). VAE constrain an encoding network and we tested this to analyze a series of elastic scatter patterns from colonies of pathogenic organisms. VAE creates set of normal distributions and each of these distributions is subsequentially responsible for certain properties of the original images. We used VAE to identify features in the features space for several hundred images. For example one function would be of size of the colony based on scatter size, another one could be intensity of the image. Another could be the number of rings within the image, and so on. Thus each sample within our image database can be coded with set of features that facilitate fast preliminary search for similar images. In addition, such initial selection could assist in identifying bad samples (i.e. bubbles or dust spots in the agar, doublets where two colonies are overlapping, etc.) during the acquisition time thus removing non-biological artifacts prior to analysis. An interesting result was that while VAE created far more realistic synthetic images closer to the original image, a simple autonencoder resulted in better cluster separation. We applied this novel approach to scatter patterns generated from *Arcobacter* species, a group that has been associated with various human and animal diseases [7].

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

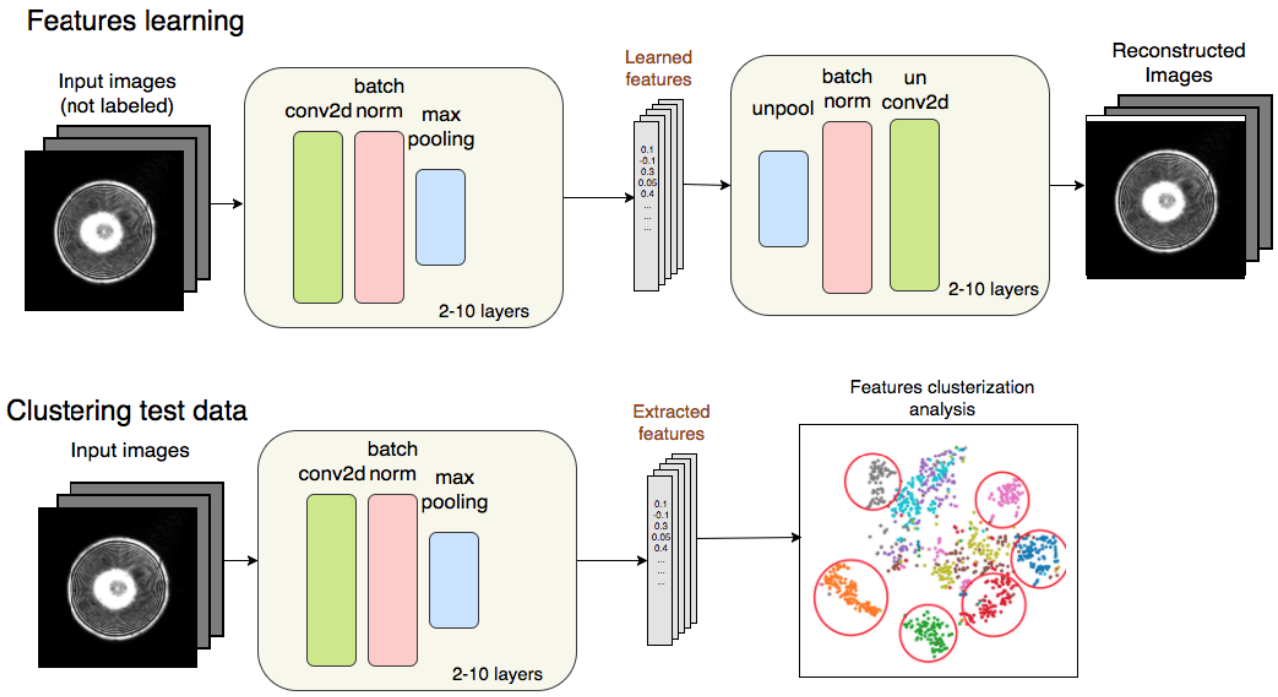

Figure 2. Two applications of convolutional |AE. First application is called the feature learning where unlabeled data is presented to AE and the algorithm automatically find and 'learns' the best feature to the set of input images. This provides insights to the compositional difference of scatter patterns and can reconstruct the representative images of a species from the learning process. A second application is to utilize the extracted features to provide clustering analysis of unlabeled sample images. Clustering unlabeled images has many applications including process monitoring and quality control of microbial based products.

2. MATERIAL AND METHODS

2.1 Biological sample preparation

Arcobacter samples were shipped using transport swabs and initially cultured on BHI + 5%SB (sheep blood) (Hemostat Laboratories, defibrinated sheep blood) at 30 C. *Arcobacter* species were subsequently cultured on Brain Heart Infusion Agar (Difco ref#241830) with additional agar to bring the percent agar to 2% and the aseptic addition of filter sterilized ferrous sulfate (Sigma, #F8633) in water for a final of 2.78 g ferrous sulfate/L immediately prior to pouring plates. Strict control of agar volume is critical since the optical characteristics of the scatter detectors are calibrated for the depth of agar determined by this method.

Once plates were tested for sterility, preparation of plates proceeded by depositing 50 microliters of sample dilution onto the center of the plate with distribution using a sterile spreading stick. For those species requiring additional salt, 2% NaCl was added (to blood agar, or BHI agar). For this study, six of representative *Arcobacter* species were selected for analysis (*A. butzleri, A. nitrofigilis, A. aquimarinus, A. defluvii, A. faecis, and A. lanthieri*). Scatter images were collected using an automated incubator, a Cytomat 2C (Thermo) at 30 °C and connected to an elastic light scatter (ELS) system. Images were collected every 90 min between hours 12 to 48 of culture and every 4 hours thereafter.

2.2 Image acquisition

The ELS instrument consisted of a laser source, plate imaging camera, sample handler and scatter camera as described previously[1,8,9]. Using the plate imaging camera, the instrument automatically collects a map of the colony locations

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

based on user defined criteria of colony diameter and circularity. Based on the spatial locations of the colonies, the system implements a traveling salesman-based pathway algorithm to optimize collection time. Data are saved onto the organism database which checks for previously classified organisms. Typically, if the database finds a match the organism is identified, if not, this colony is plated onto a fresh plate to produce at least 50 colonies and is also sent for sequencing and biochemical analysis to determine its identity. Once that identity is made, the database classification can be corrected or updated. This size parameter can be preset and changed between 300 and 1200 microns for the specific laser installed in the instrument. For subsequent data processing, a specialized program called BacLan (see Figure 3 for screenshot) is used which interrogates the database and creates classifications, finds specific colonies, or specific organisms, experiments or plates and produced a variety of analytical outputs as necessary.

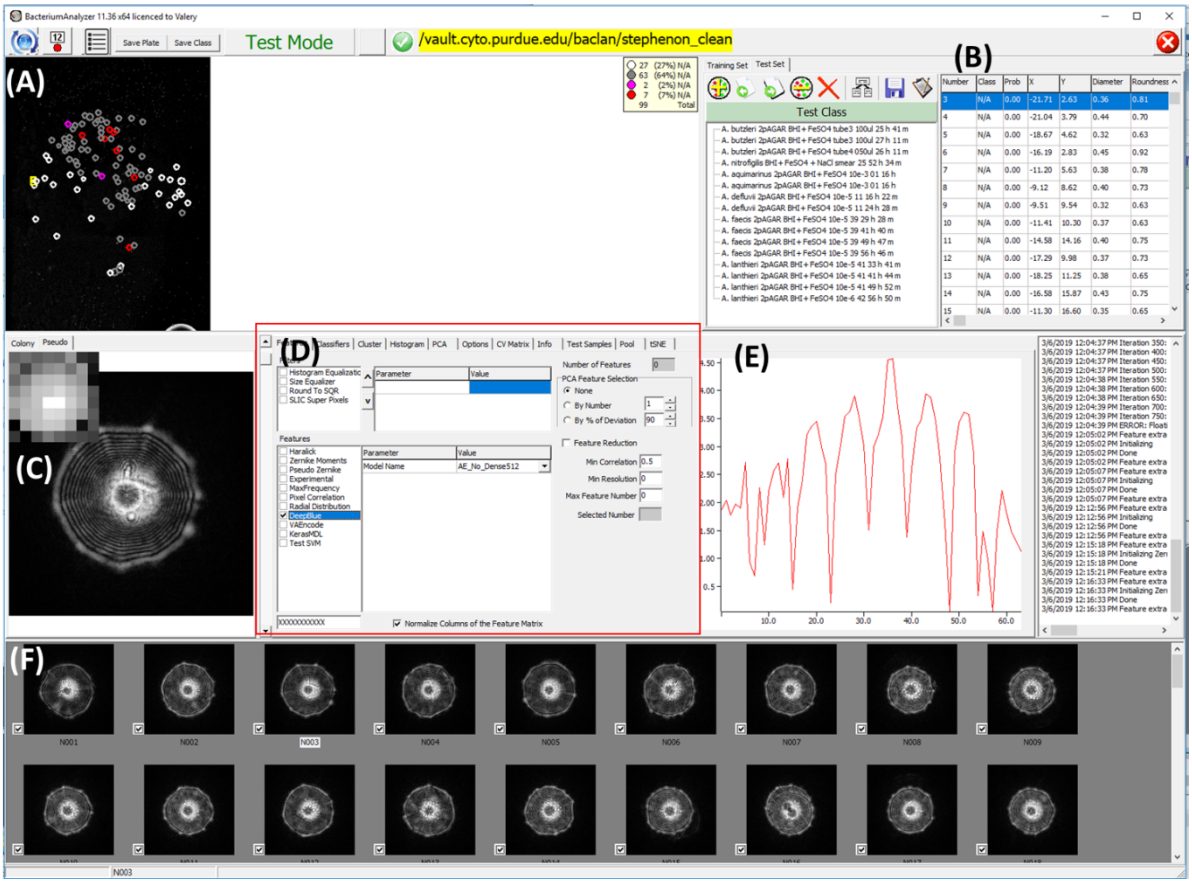

Figure 3. Screenshot of the scatter pattern analysis software (Baclan) developed by Purdue University. This software suite can perform both conventional supervised classification (Zernike-SVM) and unsupervised classification (clustering with K-means and AE). (A) is the plate image of the bacterial colonies. The current example shows color coded clustering result after AE is applied. (B) shows the statistical parameters of each colonies. (C) Thumbnail image of the selected colony. (D) panel for both supervised and unsupervised parameters can be selected here. (E) Screen displaying the feature variations. Here AE features were displayed. (F) Thumbnail images (only some shown) of all of the scatter patterns from one plate.

2.3 Classification method

Conventional image classification method can be grouped as supervised and unsupervised learning method. Supervised learning requires image samples where their labels are pre-checked by different methods. This ensures the algorithm to use these proven labels is used for the input classes and drives the cost function to maximize the separation distance among the classes. Our previous reports regarding light scattering image classification have relied on this method [10, 11]. All of the training library images were captured from bacterial species where their identity had been confirmed. The additional image classification does not require a known label in an unsupervised analysis. Whatever characteristics the

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

images display, an unsupervised method will identify the best separation of the overall images and provide the best estimation of the number of unique groups within the image database. This method is described in Figure 1. Autoencoders are a type of neural network which translates an input data into a feature vector and then tries to restore an original data with minimal loss [12]. Due to its nature, the autoencoder is trained in an unsupervised manner: the labeled data is not required, since the loss function express how fine restored data matches the original one (Figure 2).

Utilization of the unsupervised classification has many applications. Here we present two of the many applications of convolutional AE. The first application is called feature learning where unlabeled data are presented to AE and the algorithm automatically finds and 'learns' the best features to the set of input images. This provides insights to the compositional difference of scatter patterns and can reconstruct the representative images of a species from the learning process. The second application is to utilize the extracted features to provide clustering analysis of unlabeled sample images. Clustering unlabeled images has many applications including process monitoring and quality control of microbial based products.

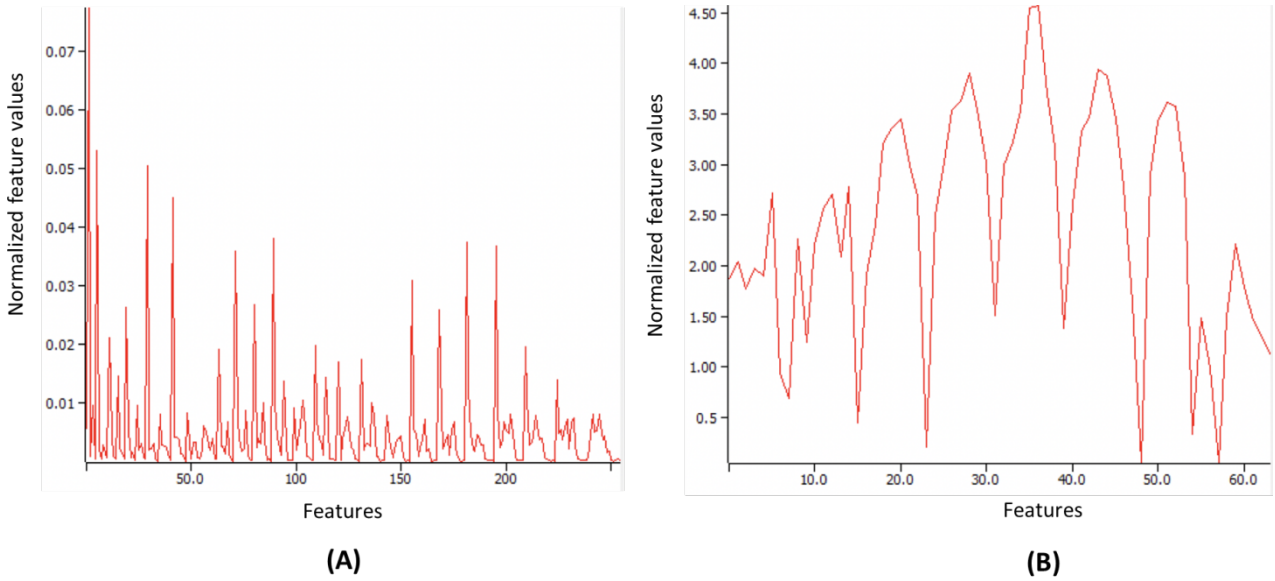

Figure 4. Represents the features extracted from images using Zernike (left) and AE (Right). In the present study we utilized 256 features for Zernike and 64 for VAE. The insert shows the Baclan software analysis that identifies the analysis used in the present study

2.4 Image analysis software

ELS based bacterial scatter pattern analysis software called Baclan version 11.36 was used for this study. This program has a built-in supervised and unsupervised classification module. When using a supervised learning method, any combination of feature extraction method (Zernike, pseudo Zernike, Fourier transform, and Haralick) can be used to retrieve hundreds of interesting features from single ELS image. Following feature extraction, a support vector machine (SVM) algorithm is implemented for classification of input images for purpose of creating training sets. Performance of the trained classifier is reported in terms of a cross-validation matrix. For testing of new samples, scanned ELS images were challenged against the pre-trained database and classification accuracy then reported. For unsupervised methods, K-means, and hierarchical clustering algorithm is already implemented. In this study, the proposed VAE is additionally implemented using a deep neutral network and the general approach is shown in the screenshot of Figure 3.

3. RESULTS

Using the Baclan software suite with six species of *Arcobacter* samples, the performance of the conventional Zernike method and proposed VAE method was compared. Figure 4 shows the screenshot from the Baclan software package when 256 features of Zernike (Figure 4(A)) was selected or 64 features of VAE (Figure 4(B)). Even from the same ELS

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

image, different features can be extracted with different magnitudes. For all six species of *Arcobacter* samples, three different feature extraction methods (Zernike, VAE(64), VAE(256)) were coupled with four different dimensionality reduction methods: Principle component analysis (PCA), t-distributed stochastic neighbor embedding (tSNE) [13], uniform manifold approximation and projection (uMAP) [14], and VAE's native reduction method.

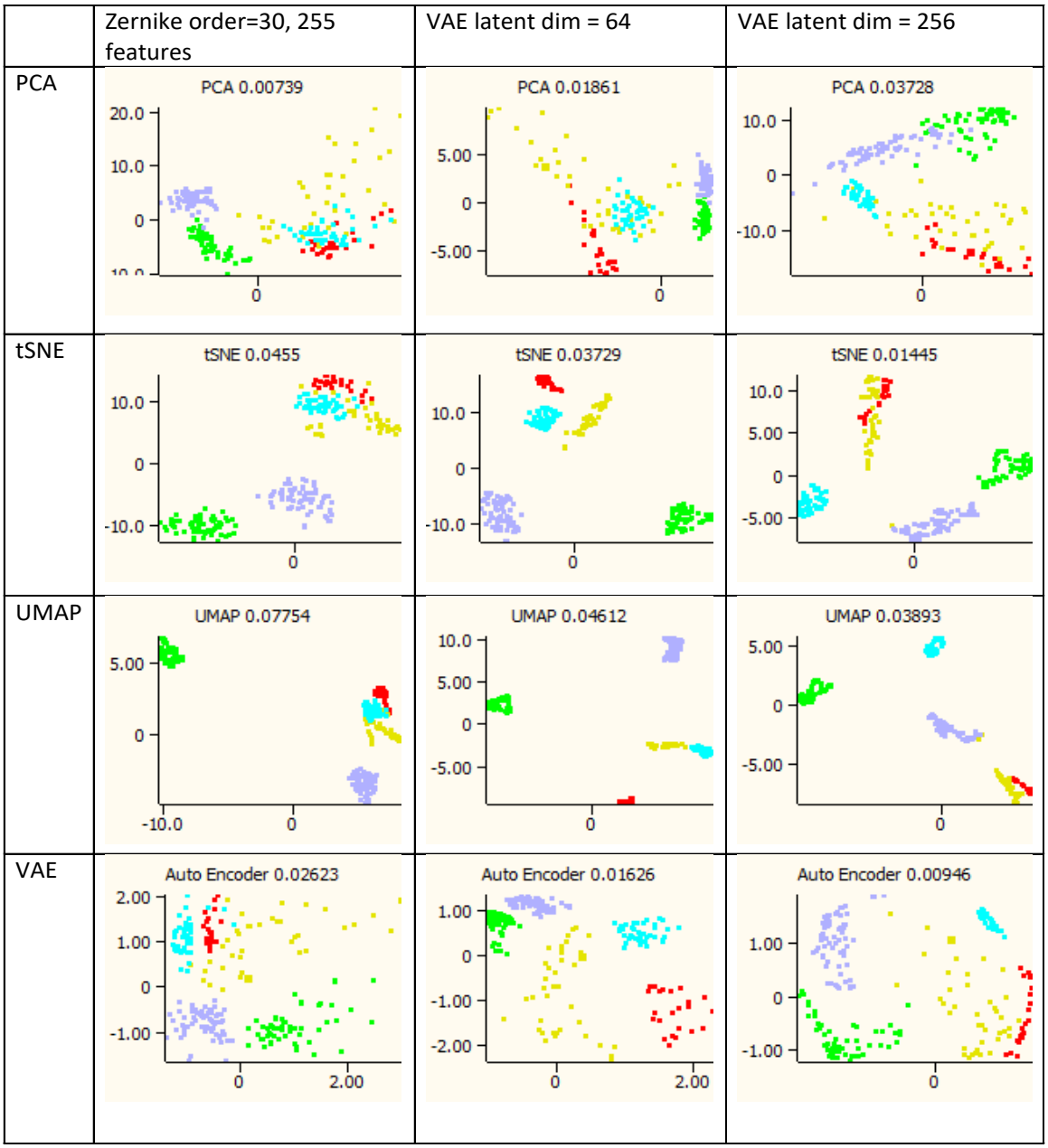

Figure 5. Table summarizing the comparison between different feature extraction and dimensionality reduction and transformation methods. For feature extraction, Zernike and VAE was used while dimensionality reduction, principle component analysis (PCA), tSNE (t-distributed stochastic neighbor embedding), UMAP (uniform manifold approximation and projection), and VAE's native reduction method. Input data sets were the 6 different species of *Arcobacter* and each transformation is labeled with their performance by Kullback-Leibler (KL) divergence.

To accurately provide the quality of clustering a separation characteristic after transformation, Kullback-Leibler (KL) divergence method was implemented for each combination. Figure 5 displays the visual comparison of the 12 cases.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Based on the comparison of the KL divergence, in most cases, Zernike (256) features resulted in better classification than the particular AE used in this study (64 features). Zernike is very fast and the main advantage of this feature extraction is that it works best with images with rotational invariance (circular images). Meanwhile, in terms of computational speed, AE is fast, but not as fast as Zernike. However, AE is particularly useful in reconstruction of image structure for checking the quality or integrity of the original image to determine if it is the accurate representation of scatter patterns. This representative scatter pattern has an important role in dissecting the correlation between genotypic/phenotypic characteristics to the specific feature of the scatter patterns.

Another perspective of the feature extraction and dimensionality reduction combination is that each pair provides different results. For example, Combining AE with tSNE provided excellent class separation while this combination requires intensive computation resources. Meanwhile, AE with uMAP maintains the similar separation while the result can be delivered in a short period of time. An additional advantage of AE is to allow faster searching of the data based because of the reduced features. Thus searching for matching feature space using an AE based approach is extremely fast even for a very large database

4. CONCLUSION

New insight into an unsupervised classification algorithm called VAE was introduced and the effectiveness of the approach was evaluated with six species of *Arcobacter* samples. ELS image of these species provided a model dataset to assess the clustering efficiency of the conventional feature extraction method (Zernike) and the alternative VAE approach. 12 different combinations of the feature extraction and dimensionality reduction approaches produced a map of optimal analysis methods for best separation of ELS patterns among the tested *Arcobacter* species. Future work will include validating this unsupervised classification into the supervised learning method and evaluation of the effectiveness of the classification ratios.

5. ACKNOWLEDGEMENT

This material is based upon work supported by the U.S. Department of Agriculture, Agricultural Research Service, under Agreement No. 59-8072-6-001; and the Royal Society of New Zealand "Catalyst" Fund, grant no. 17-LIU-003-CSG. Any opinions, findings, conclusion, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture or Royal Society of New Zealand.

REFERENCES

[1] B. Bayraktar, P. P. Banada, E. D. Hirleman *et al.*, "Bacterial phenotype identification using Zernike moment invariants," Proceedings of SPIE*,* 6080, 60800V (2006).

[2] M. Dundar, Q. Kou, B. Zhang *et al.*, "Simplicity of Kmeans Versus Deepness of Deep Learning: A Case of Unsupervised Feature Learning with Limited Data." 883-888.

[3] C. Sommer, R. Hoefler, M. Samwer *et al.*, "A deep learning and novelty detection framework for rapid phenotyping in high-content screening," Molecular Biology of the Cell*,* 28(23), 3428-3436 (2017).

[4] I. Goodfellow, Y. Bengio, and A. Courville, [Deep Learning] MIT Press, (2016).

[5] Y. Xu, Z. Jia, Y. Ai *et al.*, "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation." 947–951.

[6] O. Russakovsky, J. Deng, H. Su *et al.*, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision*,* 115(3), 211-252 (2015).

[7] L. Collado, and M. J, Figueras. "Taxonomy, epidemiology, and clinical relevance of the genus *Arcobacter*." Clin. Microbiol. Rev. 24(1), 174-92. (2011).

[8] E. Bae, A. Aroonnual, A. K. Bhunia *et al.*, "System automation for a bacterial colony detection and identification instrument via forward scattering," Measurement Science and Technology*,* 20, 015802 (2009).

[9] E. Bae, N. Bai, A. Aroonnual *et al.*, "Modeling light propagation through bacterial colonies and its correlation with forward scattering patterns," J.Biomed.Opt.*,* 15(4), 045001 (2010).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

[10] P. P. Banada, S. Guo, B. Bayraktar *et al.*, "Optical forward-scattering for detection of Listeria monocytogenes and other Listeria species," Biosens.Bioelectron., 22(8), 1664-1671 (2007).
[11] B. Rajwa, B. Bayraktar, P. P. Banada *et al.*, "Noninvasive forward-scattering system for rapid detection, characterization, and identification of Listeria colonies: image-processing and data analysis," Proceedings of SPIE, 6381, 638105 (2006).
[12] Y. Bengio, "Learning Deep Architectures for AI," Foundations and Trends® in Machine Learning, 2(1), 1-127 (2009).
[13] L. van der Maaten, and G. E. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, 9, 2579-2605 (2008).
[14] E. Becht, L. McInnes, J. Healy *et al.*, "Dimensionality reduction for visualizing single-cell data using UMAP," Nature Biotechnology, 37, 38 (2018).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 16 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use