Machine Learning Analysis of Phage Oxidation for Rapid Verification of Wash Water Sanitation

Hemiao Cui [a], Reza Ovissipour[a†], Xu Yang[a‡], Nitin Nitin[a,b,*]

[a] Department of Food Science and Technology, University of California-Davis, Davis, CA 595616, United States

[b] Department of Agricultural and Biological Engineering, University of California-Davis, Davis, CA 95616, United States

Running Title: Machine learning analysis of phage oxidation for verification of wash water sanitation

* Address correspondence to Nitin Nitin, nnitin@ucdavis.edu

[†]Present address: Reza Ovissipour, Seafood AREC/Department of Food Science and Technology, Virginia Polytechnic Institute and State University, Hampton, VA 23669, USA

‡Current address: Nutrition and Food Science Department, California State Polytechnic University Pomona, Pomona, CA, United States of America

'Declarations of interest: none

Abstract

The current approaches for process verification during sanitation of fresh produce and other minimally processed products are limited to point measurements of sanitizer concentration at discrete locations and lack rapid biological measurements to assess effectiveness of sanitation. To address this gap, this study evaluates immobilized T7 phage on anodisc membrane (phage@anodisc) as a surrogate for process verification. Fourier Transform infrared (FTIR) spectroscopy results suggested that both chlorine and Peracetic acid (PAA) caused phage DNA damage and protein oxidation. The Gradient Boosting algorithm was employed to develop predictive model for sanitizer concentration levels and *Escherichia coli* O157:H7 inactivation. The machine learning model predicted both the effective sanitizer concentration level and bacterial reduction with ROC (receiver operating characteristic) values between 0.86 and 0.93. Overall, this study identified spectral measurement of phage particles in combination with machine learning approach as an effective tool for process verification.

Keywords: Fresh produce, sanitation, rapid verification, phage, vibrational spectroscopy, machine learning

1. Introduction

Fresh produce safety is recognized as one of the challenges by the U.S. Food and Drug Administration. Food contamination by microorganisms may occur at various stages in the food supply chain. Postharvest handling of fresh produce usually involves various cooling and washing steps as well as various mechanical equipment for transportation, storage and packaging of fresh produce. During these handling steps, fresh produce can be contaminated with microbes from wash water or food contact surfaces (Suslow, 1997). Hence, disinfection of wash water and equipment is a critical step to ensure the safety and quality of fresh produce (Suslow, 1997; Gil, Selma, Lopez-Galvez, & Allende, 2009; Cossu, Le, Young, & Nitin, 2017). Monitoring and rapid validation of sanitization efficacy is critical for fresh produce industry to provide safe products, and meet the preventive control requirements of the Food Safety Modernization Act (Brackett, Ocasio, Waters, Barach, & Wan, 2014). The current approaches for validation of sanitation include the standard plate counting methods, water chemistry based on sanitizer concentration, total organic content, oxidation reduction potential (ORP), turbidity and pH of the aqueous phase (Cossu et al., 2017). However, these methods are limited in direct assessment of biological damage induced by sanitizers and can be influenced by complexity of the environment, such as fouling of electrodes and the presence of organic matter (Cossu et al., 2017).

To assess the biological response to sanitizers in wash water, previous studies have explored measurement of changes in cell membrane permeability, enzymatic activity, protein oxidation and DNA damage (Cossu et al., 2017). These measurements suggest the potential for measuring biochemical changes to validate sanitation efficacy. One of the key limitations of this approach is that the biochemical measurements required multiple sample preparation and biochemical reactions steps to assess changes induced by sanitizers. Furthermore, due to the constraints of

introducing live bacteria in food facilities including commensal bacteria, translation of these biological approaches for measurements in food industry is limited. In contrast to these biochemical assays, we recently developed a spectroscopic approach to quantify oxidation of isolated DNA molecules after exposing to different chlorine concentrations using vibrational spectroscopy and chemometrics (Ovissipour, Rai, & Nitin, 2019). Spectroscopic approach significantly reduces the operational complexity of biochemical measurements in cells both by reducing the time and the number of manual steps required for the biochemical assay. Thus, spectroscopic analysis combined with appropriate biological surrogate can provide a platform to rapidly assess sanitation efficacy. The spectroscopic analysis using machine learning approaches can reduce the multidimensional spectroscopic dataset and uncover complex relationship (Zareef et al., 2020). Supervised machine learning models can be used for quantitative analysis to identify biochemical changes in biological surrogates. Predictive quantitative models can be built using proper reference data set such as sanitizers concentrations and bacterial population.

The overall objectives of this study were to evaluate the potential of using phage as a biological surrogate for assessment of sanitation efficacy of wash water using vibrational spectroscopic measurements. Bacteriophage was selected as a biological surrogate as phages are commonly present in the environment, relatively easy amplification procedures to generate phages and simple structural composition (nucleic acid and protein). In addition, phages are commonly used as an indicator organism to evaluate contamination of water and likely to be more widely acceptable by food industry as a surrogate compared to live bacteria. Upon interaction between phages and sanitizers (e.g. chlorine or peracetic acid), the results of phage DNA oxidation and DNA conformational changes induced by wash water with chlorine or peracetic acid were measured using FTIR. Spectra from FTIR were analyzed using both principle

component analysis for classification of the spectral data and the Gradient Boosting Algorithm for quantitative predictive model development. The results of this study illustrate the potential of a novel approach to validate antimicrobial potential of wash water in the fresh produce industry.

2. Materials and methods

Reagents and supplies

Peroxyacetic acid (PAA), NaOCl (10% sodium hypochlorite), citric acid, Dey-Engley neutralizing broth (D-E broth), tryptic soy broth (TSB) and tryptic soy agar (TSA) were obtained from Sigma-Aldrich (St. Louis, MO, USA). Whatman® anodisc inorganic filter membrane (13 mm, 0.02 µm pore size) was obtained from GE Healthcare (Buckinghamshire, UK). Filtration system (a filtering flask and a fritted glass base) was obtained from Fisher Scientific (Pittsburgh, PA, USA). Phosphate buffered saline (PBS) was purchased from Fisher Bioreagents (Fair Lawn, NJ, USA). Milli-Q water was produced by QPAK® 2 purification system (EMD Millipore, Billerica, MA, USA).

Bacterial cultures and phage preparation

Both *E. coli* BL21 (ATCC BAA-1025) and bacteriophage T7 (ATCC BAA-1025-B2) were obtained from the American type culture collection. A shiga-toxin knockout rifampicin-resistant *E. coli* O157:H7 mutant (ATCC 700728) was kindly given by Dr. Linda Harris (University of California, Davis). Both *E. coli* strains were cultured in TSB broth at 37 °C for 16 hours before use.

Bacteriophages were propagated as the following procedure. Bacteriophages were first inoculated into log-phase *E. coli* BL21 culture at the ratio of 1:100 (phage: bacteria). The mixture was incubated at 37 ºC for 15 min for initial infection and then centrifuged at 16100 × *g* for 10 min. Supernatant was discarded and the same volume of TSB was added to resuspend the

pellet, followed by incubating at 37 ºC with 200 rpm shaking until no visible turbidity was observed. Chloroform was then added to the final concentration at 20% (vol/vol) and incubated at 4 ºC overnight. Then, chloroform added mixture was centrifuged at 5,000 × *g* for 10 min and water phase was collected. The water phase which contained free phages at around 10^{10} PFU/ml was used for further filtration and FTIR analysis.

Preparation of sanitizer solutions

Both PAA and NaOCl and their concentrations were selected based on their application in food industry. PAA was prepared at final concentrations of 20, 40, 60 or 80 ppm while NaOCl was prepared at final free NaOCl concentrations of 2, 5, or 10, 15 ppm and adjusted pH to 6.5-7.0 using 0.5 M citric acid.

Preparation of phage@anodisc

Free phage solution (1 ml) prepared as previously described was gently pipetted onto anodisc and filtered using the filtration system. After filtration, 1 ml of sterile MilliQ water was gently added to anodisc to wash and get rid of soluble impurities that generated during phage preparation.

Exposure of phage@anodisc to PAA or NaOCl

Five phage@anodisc were exposed to PAA at 0, 20, 40, 60 or 80 ppm for 2 min at 4 ºC. The time and temperature were selected based on food industry sanitation protocol. Similarly, another five phage@anodisc were also exposed to 0, 2, 5, 10, 15 ppm of NaOCl solution at the same condition. All phage@anodisc were then rinsed with 0.1% sodium thiosulfate to inactivate the NaOCl followed by MilliQ water, before FTIR analysis.

Inactivation of bacteriophage or *E. coli* using PAA or NaOCl

Bacteriophages or *E. coli* O157:H7 cells were inoculated into PAA or NaOCl solutions with different sanitizer concentration levels following the same procedure as described earlier for 2 min at 4 ºC. Survivor population of both phage T7 and *E. coli* O157:H7 were enumerated to quantify inactivation of these microbes as a function of sanitizer concentration.

Briefly, *E. coli* O157:H7 cells treated with PAA were 10-fold serially diluted in D-E broth and allow for incubating at room temperature for at least 10 min before plating to allow recovery of injured cells. In comparison, *E. coli* O157:H7 cells treated with NaOCl were first neutralized using 0.1 M sodium thiosulfate before serial dilution and plating. Enumeration of phage T7 particles was conducted by co-incubation of 10-fold diluted phage suspension and its host *E. coli* BL21 in soft TSA agar (0.75%), followed by gently pouring the mixture in empty petri-dishes. The clear plaque can be observed and enumerated after incubation at room temperature overnight.

Fourier Transform infrared spectroscopy (FTIR)

Phage@anodisc exposed to selected levels of PAA and NaOCl concentrations were dried under the laminar hood for 2 h. FTIR spectra were collected from 4000 to 400 cm^{-1} at a resolution of 2 cm^{-1} from the phage@ anodisc membrane samples (32 interferograms) (IRPrestige-21 FTIR spectrometer, Shimadzu Co., Kyoto, Japan).

FTIR data modelling

The FTIR data was pre-processed with Python programming language to normalize and concatenate FTIR measurements. Principal component analysis PCA was then conducted with the scikit-learn library PCA package using the FTIR data as an input to reduce dimensionality of the data set and identify the principal components for classification of the spectroscopy data set. (Li & Phung, 2014).

The relationships between spectral signals and sanitizers concentrations as well as bacterial inactivation were also modeled using the Gradient Boosting algorithm, specifically the LightGBM. LightGBM is a Gradient Boosting Decision Tree (GBDT) based machine learning algorithm. Given a set of training data $(x_1,y_1),\ldots(x_n,y_n)$ where y refers to sanitizer dosage/bacteria log reduction and x refers to spectral signal at the same treatment condition, the optimization goal of GBDT is to minimize the loss function which quantifies the prediction error. Details of GBDT is shown below:

Algorithm detail for Boosting Regression Trees
1.Set $\widehat{f(x)} = 0$ and $r_{(i)} = y_{(i)}$ for all i in the training set where $r_{(i)}$ represents ith residual 2. For each tree b= 1,2, …, B where B is the total number of trees, repeat: a. Fit a tree $\widehat{f^b}$ with d splits to the training data (X, r) b. Update $\hat{f}$ by adding in a shrunken version of the new tree where λ is the learning rate: $$\widehat{f(x)} \leftarrow \widehat{f(x)} + \lambda \widehat{f^b}$$ c. Update residuals, $$r_{(i)} \leftarrow r_{(i)} - \lambda \widehat{f^b}$$ 3. Output the boosted model,

$$\widehat{f(x)} = \sum_b \lambda \widehat{f^b}$$

LightGBM builds on top of GBDT by applying Gradient-based One-Side Sampling (GOSS) method and exclusive feature bundling (EFB). LightGBM uses GOSS to determine the split point via calculating variance gain, thereby accelerating the training for each level of the tree. It uses EFB to bundle exclusive features into a single feature to reduce the number of features and uses greedy algorithm for approximation of the best split point (Ke et al., 2017). These render LightGBM with ~~was chosen due to its~~ fast speed and the ability to handle a high dimensional ~~large~~ dataset. FTIR data tends to be high dimensional. Thus, LightGBM algorithm was selected to predict sanitizers concentrations as well as bacterial inactivation from FTIR spectra~~from~~ .

~~Specifically, it uses the Gradient-based One-Side Sampling method to exclude a significant proportion of data instances with small gradients, thereby accelerating the training for each level of the tree. In addition, LightGBM bundles mutually exclusive features (rarely take nonzero values simultaneously) to reduce the number of features and uses greedy algorithm for approximation (Ke et al., 2017).~~

LightGBM model was implemented with the scikit-learn library. **Figure 1** showed the prediction pipeline. FTIR data (predictor variables) and bacterial inactivation data (response variables) were concatenated. To investigate the model performance to unseen data set, 5-fold cross-validation was conducted by splitting the combined data set to 5 folds. For each fold, model was trained on training datasets (represented as blue cells in Figure 1) and evaluated on test datasets (represented as red cells in Figure 1). The default loss function (multi Log loss) was used to construct the objective function. Predictions was evaluated based on the Receiver Operating

Characteristic (ROC) curve and Confusion Matrix (CM). Prediction errors for all 5 folds were aggregated to give the final ROC and CM value in the figures.– Sample Python code for the analysis was provided.

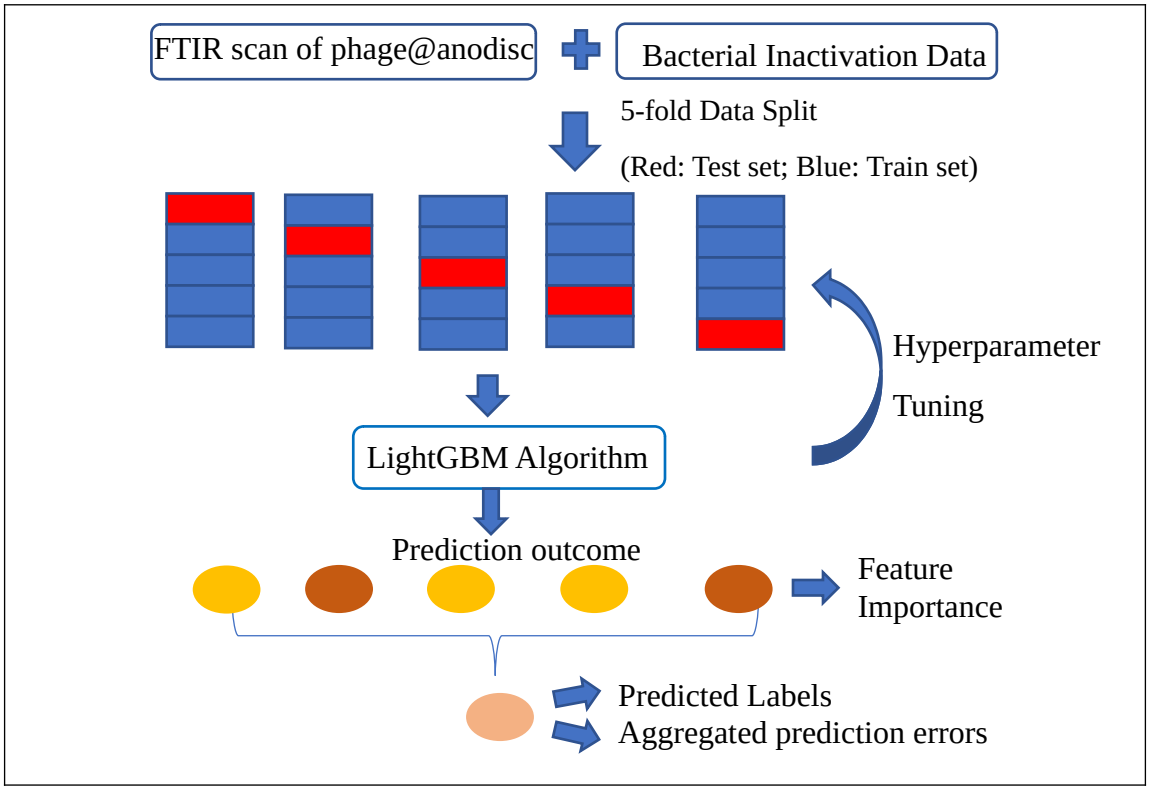


Figure 1 Flowchart of the model development pipeline

Statistical analysis

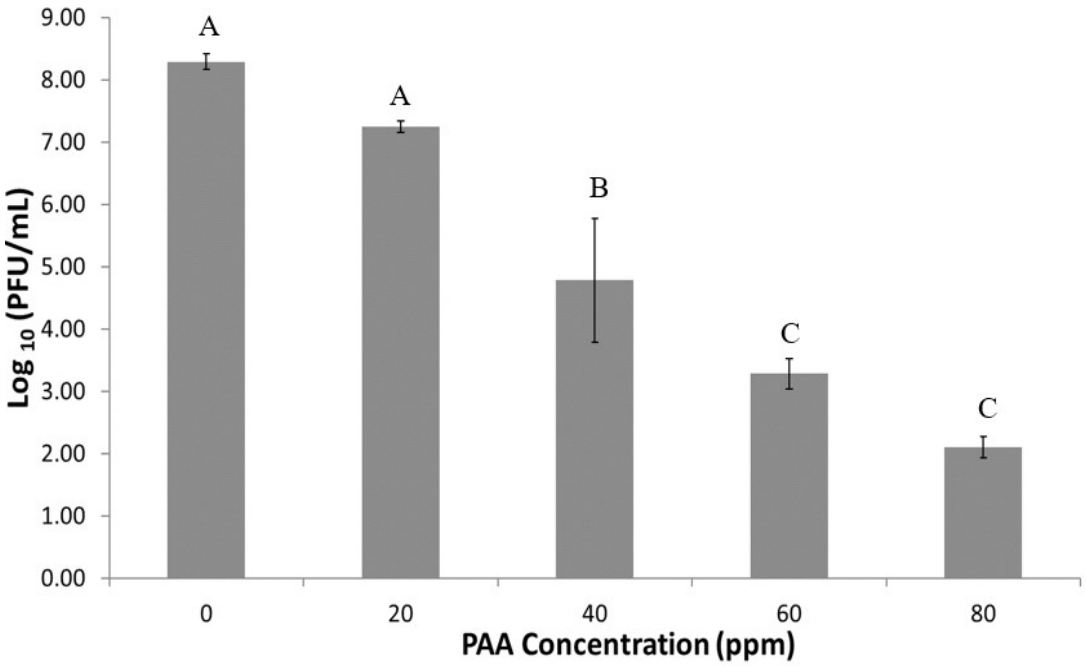
Throughout the study, experiments were conducted in three independent trials. For each trial, measurements were conducted in three replicates. The significant differences between treatments were determined through one-way Analysis of Variance (ANOVA) followed by Tukey's pairwise comparisons and $p<0.05$ is considered as significant.

3. Result and discussion:

Inactivation of phage T7 by selected sanitizers

Figure 2 illustrates the survivor population of phage T7 upon treatment with PAA or NaOCl at varying levels of sanitizer concentration. As shown in **Fig 2A**, inactivation of phage T7 increased with an increase of PAA concentration ($P < 0.05$). PAA at 80 ppm concentration successfully inactivated more than 6-log of phage T7 but despite using high concentration levels, complete inactivation of phages (9 log inoculum level) was not observed as shown in **Fig 2A**. In comparison, complete inactivation of inoculated phages was observed at even the lowest concentration tested (2 ppm of free NaOCl) in this study. The results suggested that viral particles such as bacteriophages may be more susceptible to NaOCl than PAA. Similar results have also been reported by Morin *et al.* that MS2 phages are more resistant to PAA than NaOCl (Morin et al., 2015). This difference may result due to broad reactivity of NaOCl with proteins and DNA molecules, thus damaging both the capsid proteins and the DNA of viral particles.

A

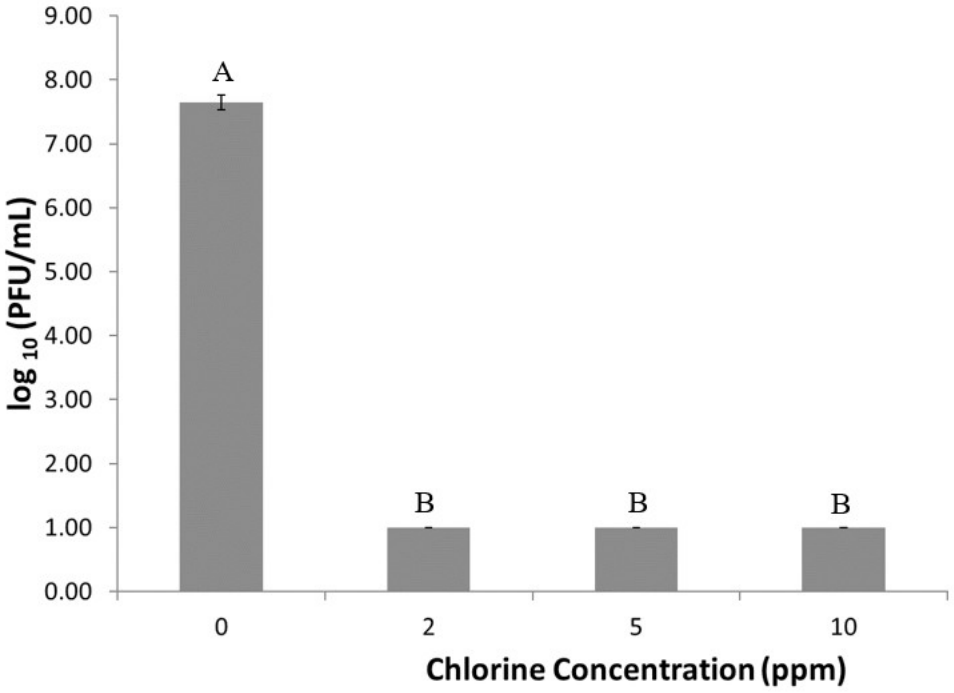


211

212

213

B

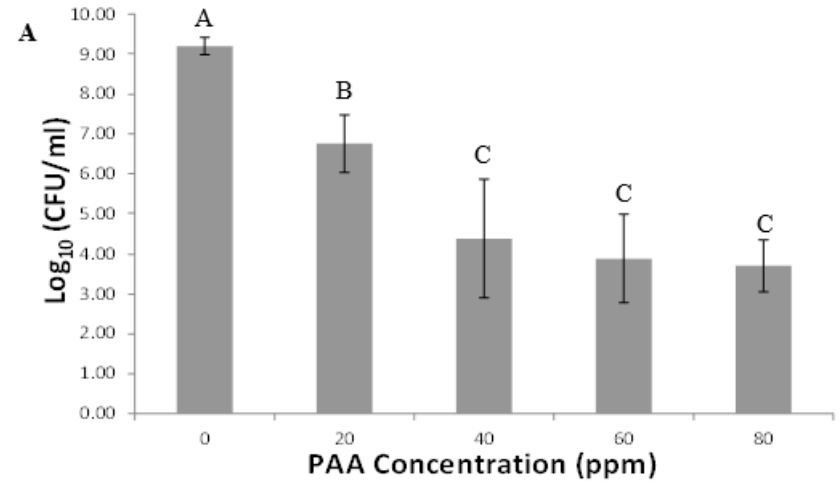


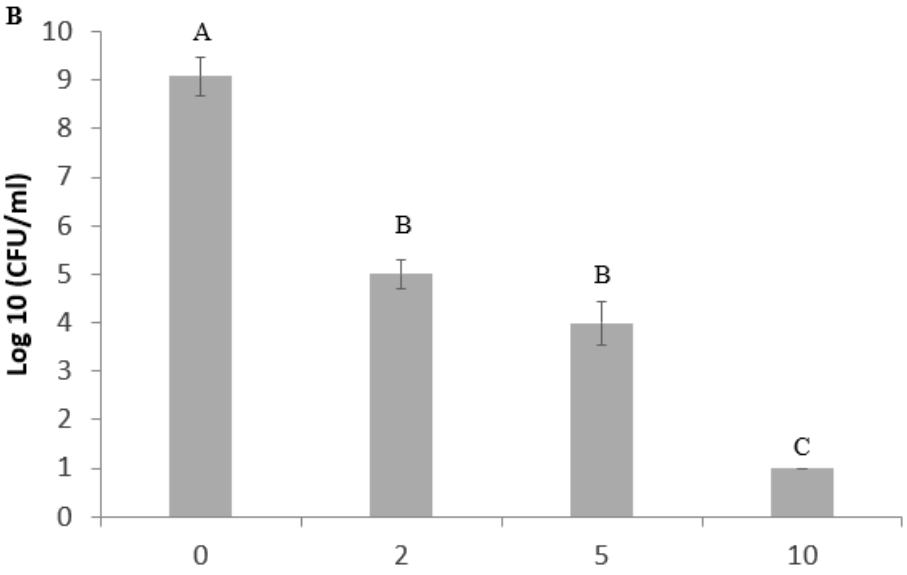
214

Figure 2 Survivor plot of phage treated by PAA (A) or NaOCl (B) at selected concentrations for 2 min at 4 ºC.

Inactivation of *E. coli* O157:H7 by selected sanitizers

Figure 3 illustrates the survivor population of *E. coli* O157:H7 upon treatment with PAA or NaOCl at varying levels of sanitizer concentration. As shown in **Fig 3A**, *E. coli* O157:H7 cells were significantly inactivated (2-log inactivation) by PAA, even at 20 ppm. However, no significant increase in inactivation of *E. coli* O157:H7 was observed with an increase of PAA concentration above 40 ppm. Even at the highest levels of PAA (80 ppm for 2 min) used in this study, only 5 log inoculated bacterial cells were inactivated from the initial inoculum levels of 9 log of bacteria. **Fig 3B** showed *E. coli* O157:H7 reduction upon treatment with different concentration levels of NaOCl after 2 min at 4°C. NaOCl at the levels of 2 and 5 ppm caused 4 and 5 log reduction, and 10 ppm of NaOCl completely inactivated *E. coli.*





229

230

231

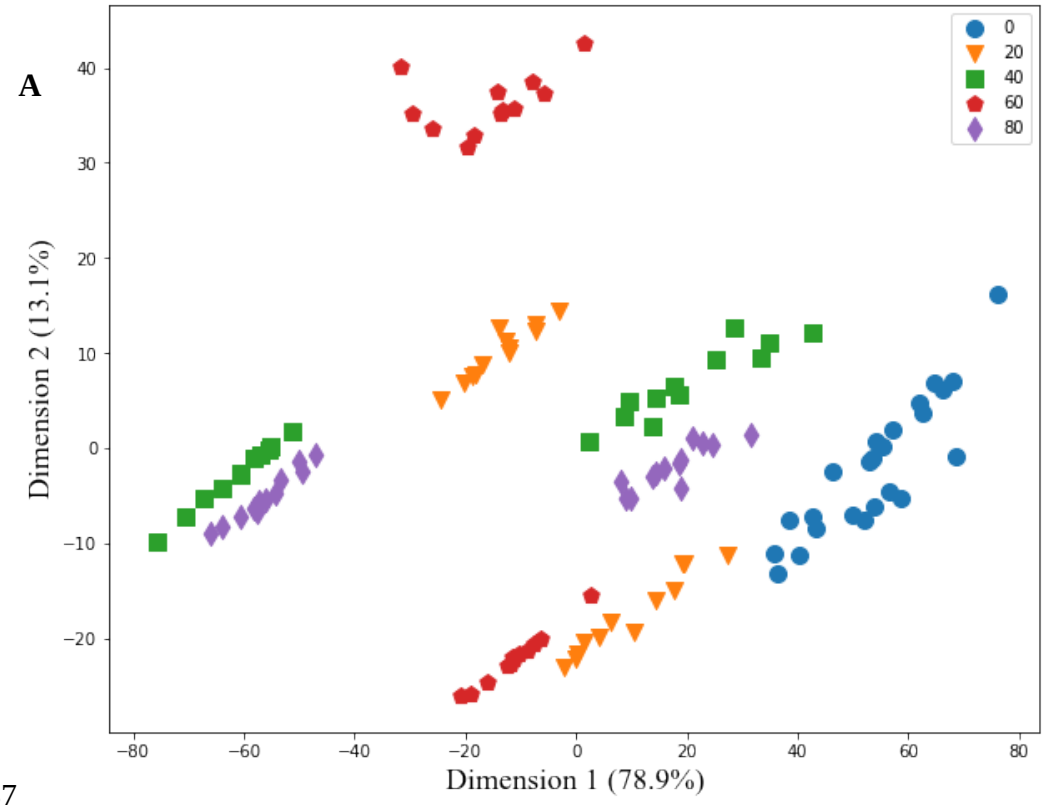
232 **Figure 3** Survivor plot of *E. coli* O157:H7 treated by PAA (A) or NaOCl (B) at selected

233concentrations for 2 min at 4 ºC.

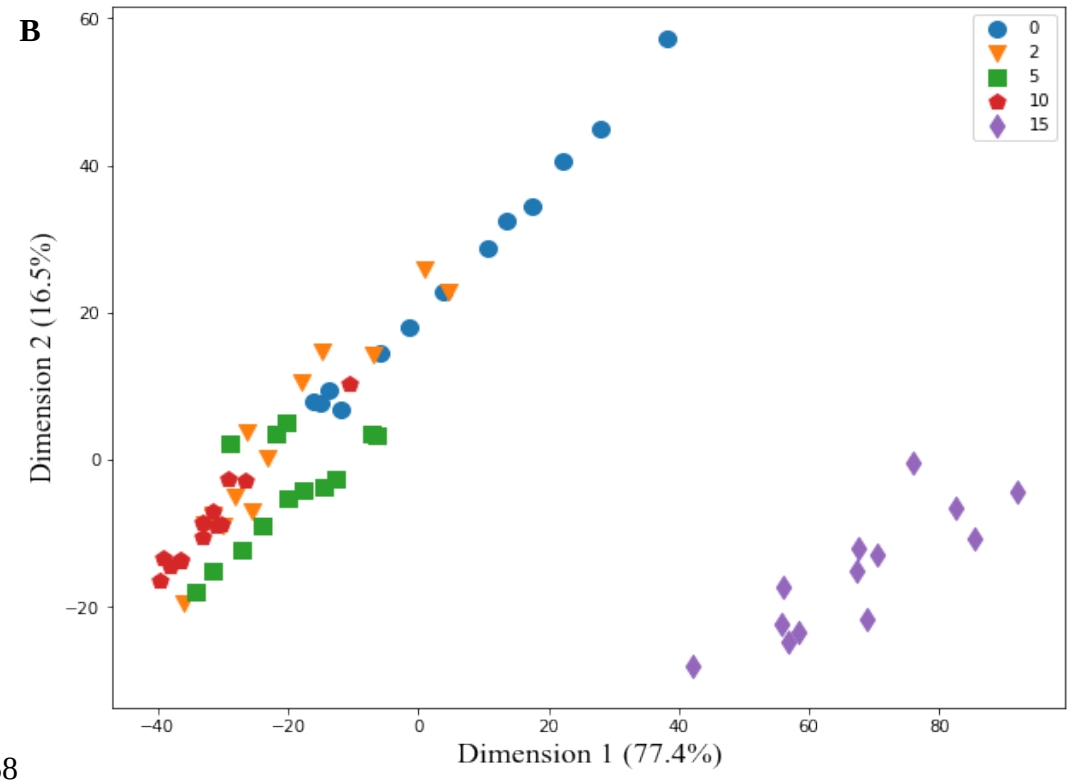
234

235**PCA models**

PCA analysis has been used to describe the variations in evaluating oxidation of DNA in live *E. coli* cells upon exposure to sanitizers (Al-Qadiri, Al-Alami, Al-Holy, & Rasco, 2008). DNA oxidation and changes in DNA-protein interactions have been shown to be target sites for oxidative reactions with NaOCl in the case of phage (Maillard, 1996). The PCA models for NaOCl or PAA treated phage@anodisc are presented in **Figure 4**. The PCA results illustrated that the spectral changes in the DNA region of a phage@anodisc is dose dependent and the PCA model discrimination of spectral changes in NaOCl or PAA treated phage@anodisc was not optimal. In the PCA model for PAA treated phage@anodisc, the PC1 and PC2 components explained 78.9% and 13.1% of the variations in the spectral band corresponding to the DNA region, respectively. In the PCA model of NaOCl treated phage@anodisc, the PC1 and PC2 components explained 77.4% and 16.5% of variations in the spectral band corresponding to the DNA region, respectively. For PAA groups, the same PAA concentration corresponded to mostly two clusters on the PCA visualization plot. For NaOCl, there was a high level of overlapping between treatment dosages on the visualization plot. This was likely due to nonlinear response of phage damage to NaOCl treatment. Various cellular responses to oxidation reactions have been reported to have the nonlinear nature (Kalyanaraman et al., 2012; Neumaier et al., 2012). In contrast, PCA is a linear transformation of the data. PCA analysis is generally not recommended for modeling complex nonlinear relationship (Alanis-Lobato, Cannistraci, Eriksson, Manica, & Ravasi, 2015).


A
0
20
40
60
80
Dimension 2 (13.1%)
Dimension 1 (78.9%)

257


B
0
2
5
10
15
Dimension 2 (16.5%)
Dimension 1 (77.4%)

258

259

Figure 4 2D visualization of Principle Component Analysis of phage@anodisc FTIR spectra (A) PAA: 0, 20, 40, 60, 80 represents 0 ppm, 20 ppm, 40 ppm, 60 ppm, 80 ppm (B) NaOCl: 0, 2, 5, 10, 15 represents 0 ppm, 2 ppm, 5 ppm, 10 ppm, 15 ppm

Gradient Boosting model

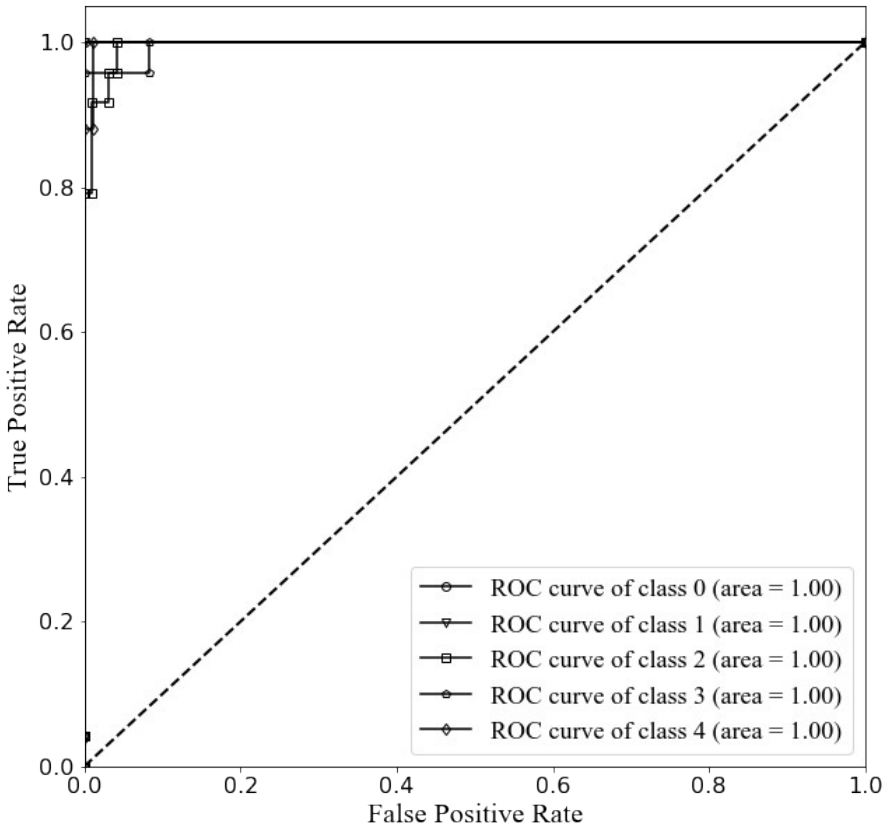
Due to nonlinear nature of the response observed based on the PCA analysis, the LightGBM was used for the analysis of the FTIR data. Using the LightGBM model, both sanitizer dosage as well as bacterial inactivation were predicted.

PAA Dosage

The prediction using the LightGBM model was evaluated based on the Receiver operating characteristic (ROC) curve and confusion matrix. A ROC curve is a plot of sensitivity on the y axis against (1−specificity) on the x axis for varying values of the threshold t (Zou, O'Malley, & Mauri, 2007). Sensitivity is defined as number of true positive samples (TP) / number of true positive or false negative (FP) samples. Specificity is defined as number of true negative samples / number of true negative or false positive samples. Threshold is the cut off probability for defining a positive class. The 45° diagonal line connecting (0,0) to (1,1) in the ROC curve corresponds to a random chance. The area under the ROC curve (AUC) is a summary measure that essentially averages diagnostic accuracy across the spectrum of test values. ROC is a suitable metric for balanced classification problem (Kotsiantis, Kanellopoulos, & Pintelas, 2006). The spectral dataset was balanced among all classes and hence ROC was selected as an evaluation matrix. Confusion matrix on the other hand provides a straightforward view of the number of data samples that have been correctly or incorrectly classified. **Figure 5** showed the

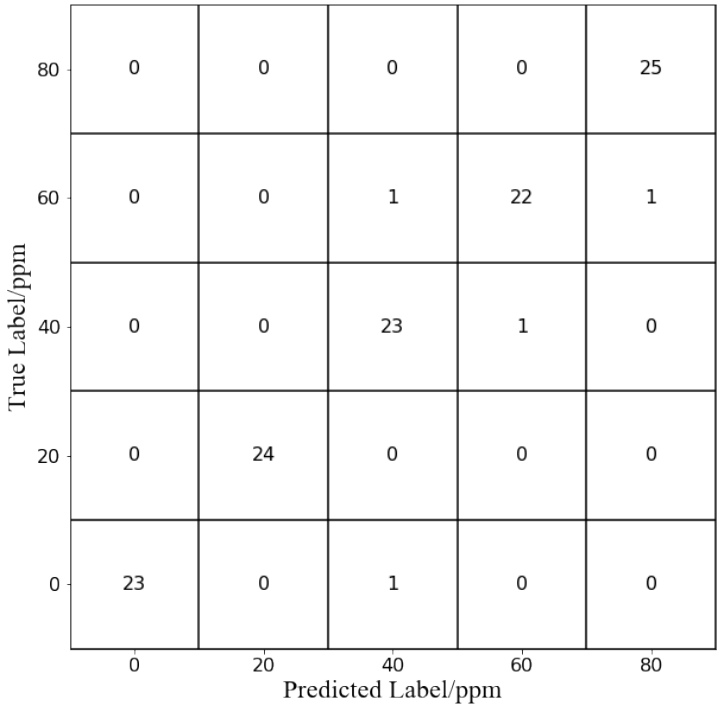
ROC and confusion matrix for predicting levels of PAA treatment, including 0 ppm (class 0), 20 ppm (class 1), 40 ppm (class 2), 60 ppm (class 3) and 80 ppm (class 4). In this result, for all 5 levels, ROC curves oriented towards the top left corner, indicating good prediction accuracy. Area under the curve (AUC) of ROC curve is another indicator for model performance. AUC for ROC curves among all levels reached 1, indicating effective classification of the spectral response. AUC results were consistent with the confusion Matrix. In the confusion matrix, classified samples are located in the diagonal part of the matrix. The total percentage of corrected predicted samples was 97% among a total of 121 samples. Data was also fitted with SVM model using Radial basis function kernel, but the prediction outcome was significantly worse (data not shown). In addition to the performance advantage of LightGBM, it is a very fast algorithm. The model training and prediction are completed within 3 mins for 121 data points. The LightGBM utilizes the Gradient-based One-Side Sampling and Exclusive Feature Bundling to expedite the calculation, which makes it suitable for handling big datasets common in industrial applications (Ke et al., 2017).

A


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 1.00)
ROC curve of class 1 (area = 1.00)
ROC curve of class 2 (area = 1.00)
ROC curve of class 3 (area = 1.00)
ROC curve of class 4 (area = 1.00)

296

B


True Label/ppm
Predicted Label/ppm
0 20 40 60 80
0 0 0 0 25
0 0 1 22 1
0 0 23 1 0
0 24 0 0 0
23 0 1 0 0

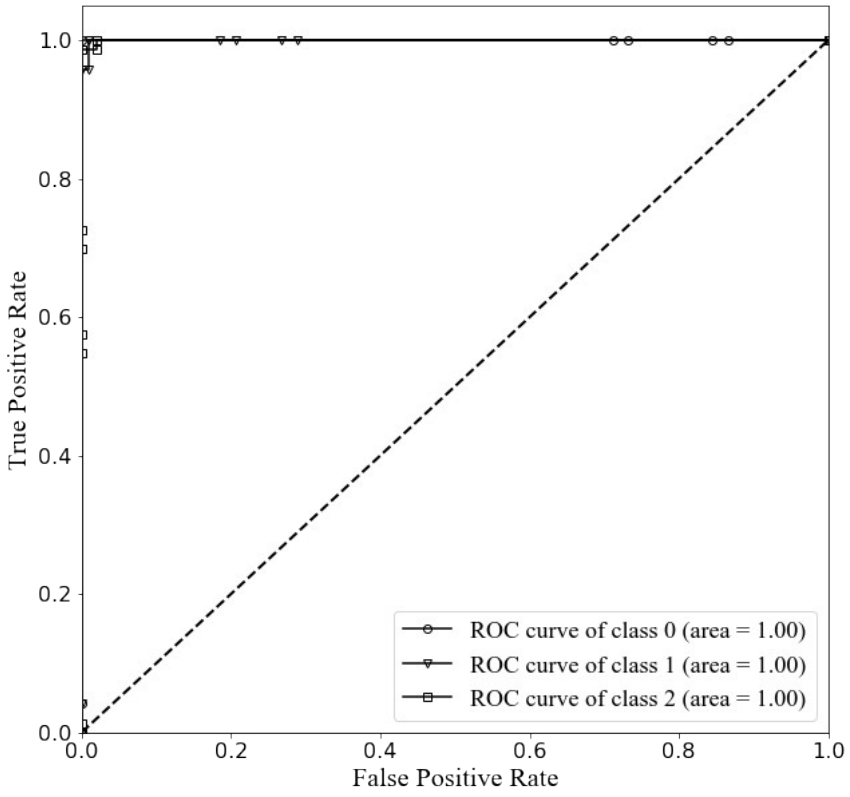
297

Figure 5 (A) Receiver operating characteristic plot (B) Confusion Matrix for prediction of PAA concentration by phage@anodisc

Bacterial Inactivation by PAA

Figure 6 showed the prediction of *E. coli* inactivation with FTIR data for PAA treatment. Data was grouped into 4 classes based on the mean log reduction, namely 0 log reduction (class 0), 2 log reduction (class 1), >4 log reduction (class 2). The categorization was based on statistical significant differences observed in the bacterial inactivation dataset. AUC for all classes were 1.0. According to the confusion matrix, the total percentage of correctly predicted samples was 96% among a total of 114 samples. Various studies have built machine learning models to predict bacterial inactivation under sanitizers treatment including PAA using processing conditions or bacterial internal signals as input variables. For example, Newhart *et al.* utilized Artificial Neural Networks to model bacterial inactivation with physiochemical properties of wastewater (Newhart et al., 2020). Caglar *et al*. predicted bacterial growth from mRNA and protein abundances data (Caglar, Hockenberry, & Wilke, 2018). To the best of our knowledge, this is the first study to predict bacterial inactivation with responses from surrogates using spectroscopy methods.

A


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 1.00)
ROC curve of class 1 (area = 1.00)
ROC curve of class 2 (area = 1.00)

B

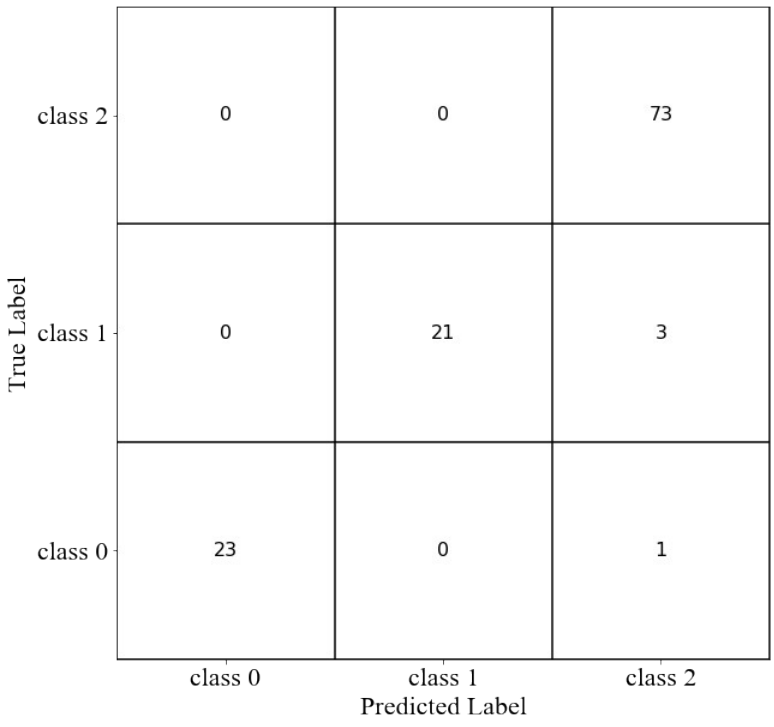

True Label
Predicted Label
class 2
class 1
class 0
0
0
73
0
21
3
23
0
1

Figure 6 (A) Receiver operating characteristic plot (B) Confusion matrix for prediction of *E. coli* O157:H7 inactivation under PAA treatment

NaOCl Dosage

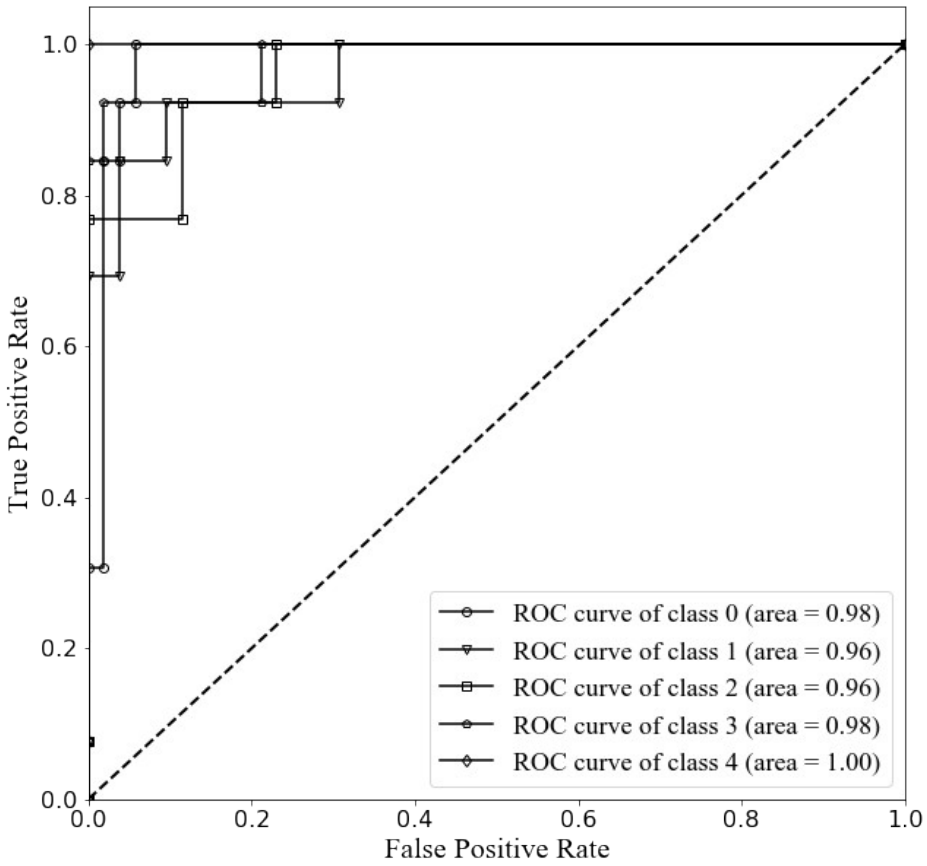
Hyperparameter tuning (HT) was utilized to improve prediction performance of the LightGBM model for NaOCl as the default parameters did not obtain optimal outcome. Grid search approach was used to conduct HT. Specifically, a set of hyperparameter combination was selected as candidates. Models were trained based on all combinations in parallel. Model performances were evaluated on test dataset. The hyperparameter used in the final model was decided based on model performances. In the LightGBM, max tree depth, learning rate, sampling rate and regularization terms are the major hyperparameters that affect the model performance (Veronika, Vasily, & Kruchinin, 2018). **Table 1** showed the effect of Learning rate and Regularization lambda L1 on the overall percentage of correct predicted samples. This metric was chosen as it represents the overall model accuracy. Learning rate controls the step size at each iteration when optimizing the objective function. Regularization lambda shrinks the model coefficients terms to prevent overfitting. The choices of these two hyperparameters significantly affected the total percentage of correct prediction. Based on this, learning rate 0.005 and Lambda l1 0.1 was chosen. Other hyperparameters and their values used in the study were as follows: feature fraction (1), bagging frequency (1) and bagging fraction (0.7083). Feature fraction defines the fraction of features to train each tree. Bagging frequency defines the frequency for resampling from input dataset to build a tree whereas bagging fraction defines the fraction of data to be used for each iteration. **Figure 7** showed the prediction of NaOCl concentration level based on the phage FTIR data using optimal hyperparameters. ROC curve showed the model

prediction performance has an AUC above 0.96 for all classes, namely 0 ppm (class 0), 2 ppm (class 1), 5 ppm (class 2) and 10 ppm (class 3) and 15 ppm (class 4). The confusion matrix also showed good prediction performance of the model. The total percentage of corrected classified samples was 88% among a total of 65 samples.

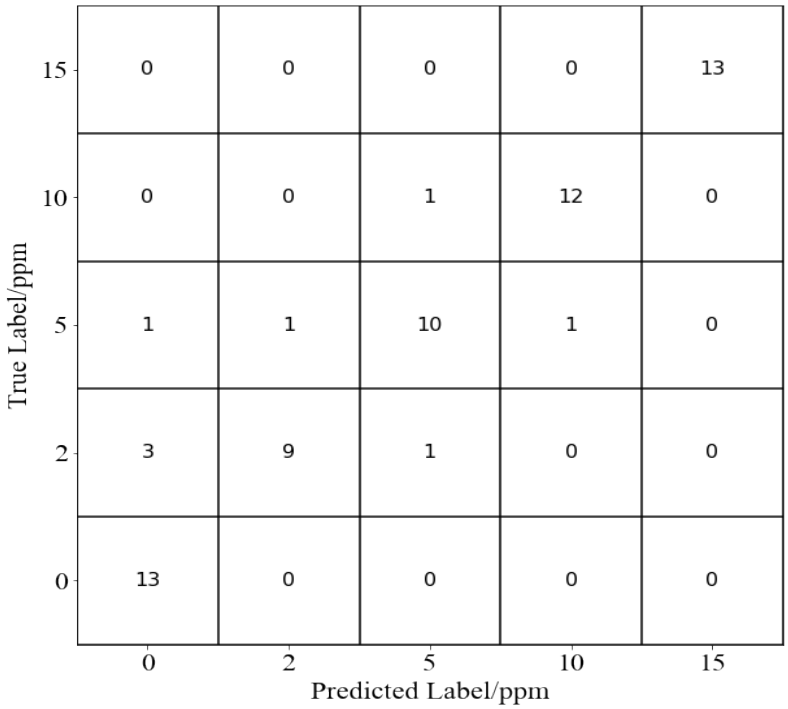
Table 1 Hyperparameter tuning of LightGBM model for predicting NaOCl dosage

Learning rate	Lambda l1	Total corrected prediction percentage
0.0001	0.1	0.49
0.0001	0.3	0.32
0.0001	0.5	0.31
0.0005	0.1	0.82
0.0005	0.3	0.32
0.0005	0.5	0.31
0.005	0.1	0.88
0.005	0.3	0.32
0.005	0.5	0.31

A


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.98)
ROC curve of class 1 (area = 0.96)
ROC curve of class 2 (area = 0.96)
ROC curve of class 3 (area = 0.98)
ROC curve of class 4 (area = 1.00)

B



356

357**Figure 7** (A) Receiver operating characteristic plot (B) Confusion matrix for prediction of
358NaOCl concentration by phage@anodisc

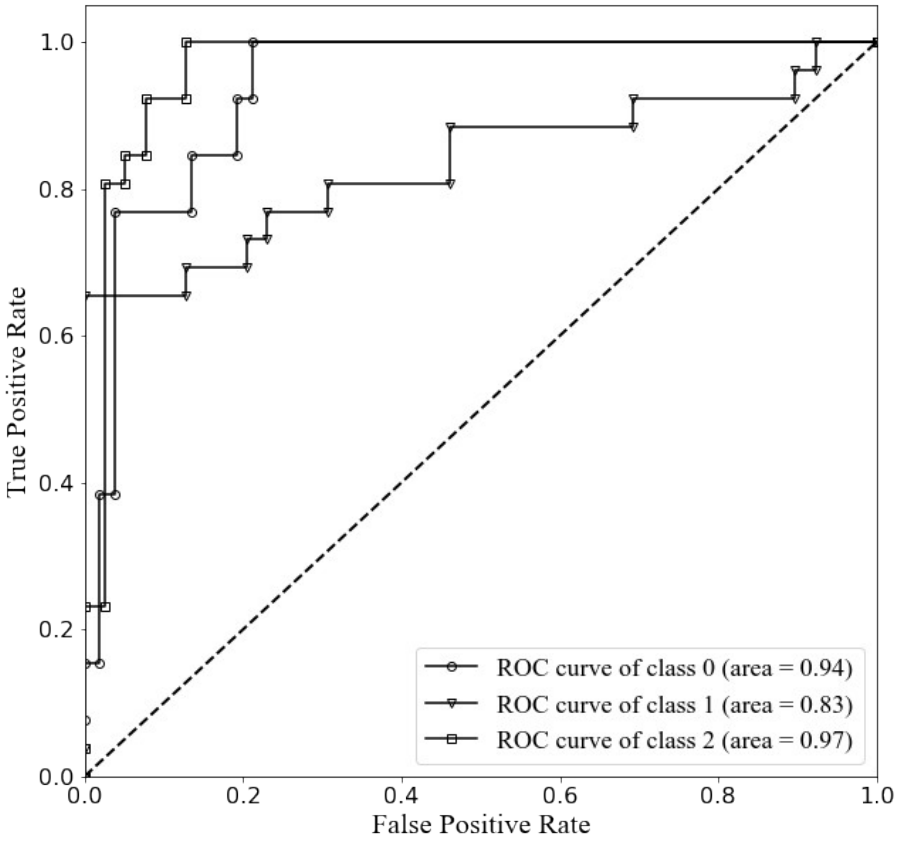
359

360*Bacterial Inactivation by NaOCl*

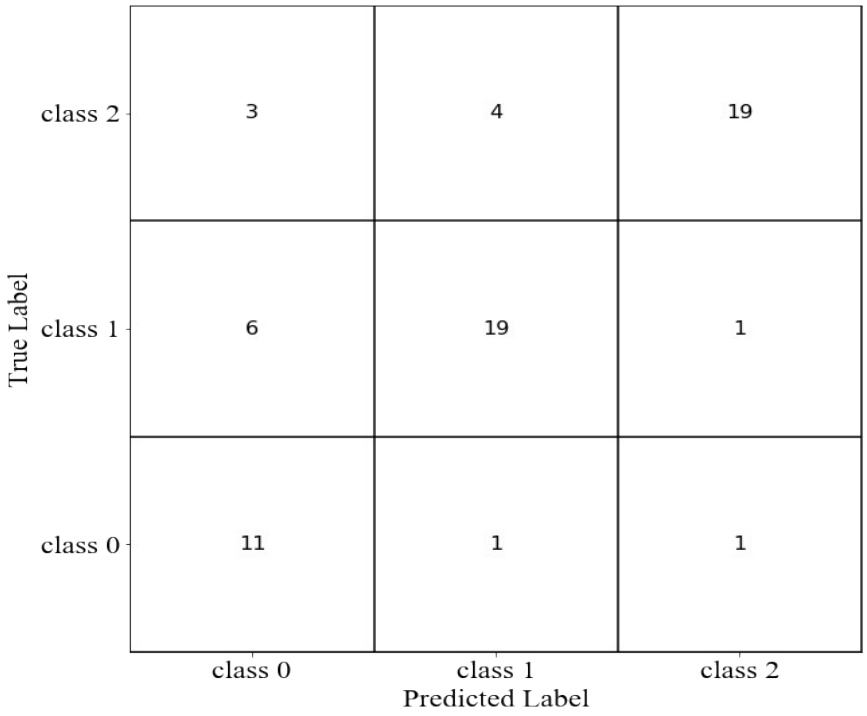
361**Figure 8** showed the prediction of *E. coli* O157:H7 inactivation with FTIR data. *E. coli*
362inactivation was grouped into 3 classes, namely 0 log reduction (class 0), 3 log reduction (class
3631) and >=8 log reduction (class 2). The categorization was also based on the statistical significant
364differences observed in the bacterial inactivation data with NaOCl treatment. AUC for all classes
365were at least 0.8. Among the three classes, class 2 had the smallest AUC (0.85), which could be
366attributed to large sampling variations within this class. According to the confusion matrix, the
367total percentage of true classification was 81% among a total of 65 samples. It is likely that the
368better model performance for predicting bacterial inactivation by PAA than NaOCl was due to a
369relatively large sample size in PAA. The performance of machine learning algorithm can be

highly impacted by the data size. Sordo *et al.* have shown that both Support Vector Machine (SVM) and Decision Trees show a substantial improvement in performance as the number of training samples increase (Sordo & Zeng, 2005).

A



B



375

376**Figure 8** (A) Receiver operating characteristic plot (B) Confusion matrix for prediction of *E. coli*

377O157:H7 inactivation under NaOCl treatment

378

379**Feature Importance Plot**

380Feature importance plot was used to reveal the contributions of each individual spectra

381wavenumber to the prediction model. It was obtained by randomly shuffling each feature to

382determine the increases in prediction error and then conducting ranking among all features

383(Breiman, 2001). The feature importance of LightGBM model between 900 cm^{-1} and 1700 cm^{-1}

384was revealed as shown in **Figure 9**. The importance plot was extracted from the model based on

385predicting bacterial inactivation under PAA treatment.

386

387

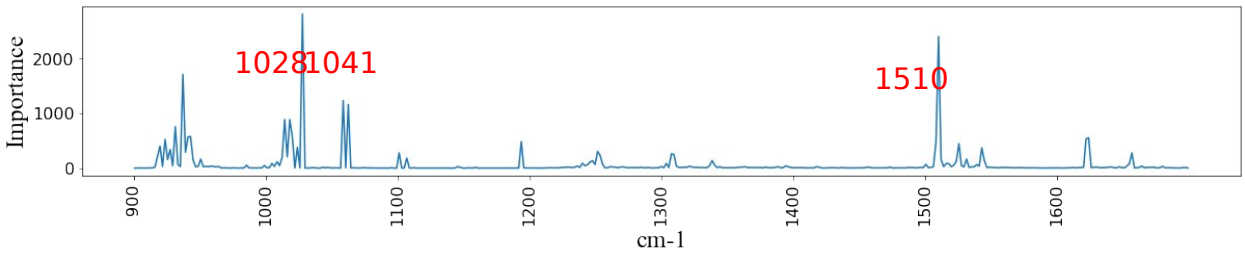


Figure 9 Feature Importance plot for the prediction of *E. coli* O157:H7 inactivation under PAA treatment (x axis: wavenumber/cm^{-1}; y axis: counts of the wavenumber being used to split the tree)

The x axis showed the spectra wavelength while the y axis showed the accumulated importance measure among all samples. Interestingly, the most important peaks were concentrated around 1000 cm^{-1} as well as 1500 cm^{-1}, which respectively correspond to the DNA and protein region in the FTIR spectra (Sahu et al., 2004; Simonova & Karamancheva, 2013). This coincided with the mechanisms of the biocidal action of PAA which involved both protein and DNA molecules (Alanis-Lobato et al., 2015; Kitis, 2004). Specifically, spectra at 1000 cm^{-1} was assigned to

conformational changes in DNA due to single stranded DNA formation. The 1083 cm^{-1} peak was assigned to symmetric phosphate groups in the DNA backbone also illustrating DNA fragmentation. Peaks at 970, 1265 and 1041 cm^{-1} have been assigned to symmetric phosphate group, asymmetric phosphate group, stretching C-O ribose, and phosphate group, respectively indicating DNA fragmentation and changes in the deoxyribose structure (Oldenhof, Schütze, Wolkers, & Sieme, 2016; Ovissipour, Rai, & Nitin, 2019; Pascolo et al., 2016). The vibrational bands at 1,044 and 1,113 cm^{-1} were also the characteristic markers of methionine oxidation (Ravi, Hills, Cerasoli, Rakowska, & Ryadnov, 2011). Amide I and amide II bands are the two major bands of the protein infrared spectrum. Peaks around protein amide I and II regions were also identified as key changes in the spectral features based on the prediction model. The protein amide I band (between 1600 and 1700cm^{-1}) was related to the protein backbone conformation. Amide II (between 1510 and 1580 cm^{-1}) was associated with the N-H bending vibration and C-N stretching vibration (Barth, 2007). Peaks in this region showed high importance in predicting bacterial inactivation. Thus, the result indicated that both changes in structural conformation and oxidation related to DNA and protein were the key features used in the prediction model.

Conclusions

The overall goal of this study was to develop a rapid sanitation process verification based on spectroscopic measurement of immobilized phage oxidation and chemometric analysis to predict concentration of the selected sanitizers and bacterial inactivation. The results of this study indicate that vibrational spectroscopy coupled with machine learning models can be used for measuring, and quantifying phage responses to chlorine and PAA, and developing model for predicting the bacterial (*E. coli* O157:H7 as the reference) reduction, and sanitizers

concentrations. Overall, this study demonstrates immobilized phage as a surrogate for verifying the sanitation process in fresh produce industry.

Acknowledgement: This project was supported by grant no. 2015-68003- 23411 from the USDA-NIFA Program Enhancing Food Safety through Improved Processing Technologies (A4131) and by the USDA-AI Institute in Food Systems.

References

Al-Qadiri, H. M., Al-Alami, N. I., Al-Holy, M. A., & Rasco, B. A. (2008). Using Fourier transform infrared (FT-IR) absorbance spectroscopy and multivariate analysis to study the effect of NaOCl-induced bacterial injury in water. *Journal of Agricultural and Food Chemistry*. https://doi.org/10.1021/jf801604p

Alanis-Lobato, G., Cannistraci, C. V., Eriksson, A., Manica, A., & Ravasi, T. (2015). Highlighting nonlinear patterns in population genetics datasets. *Scientific Reports*. https://doi.org/10.1038/srep08140

Barth, A. (2007). Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta - Bioenergetics*. https://doi.org/10.1016/j.bbabio.2007.06.004

Brackett, R. E., Ocasio, W., Waters, K., Barach, J., & Wan, J. (2014). Validation and verification: A practical, industry-driven framework developed to support the requirements of the food safety modernization Act (FSMA) of 2011. *Food Protection Trends*, *34*, 410–425.

Breiman, L. (2001). Random forests. *Machine Learning*.

https://doi.org/10.1023/A:1010933404324

Caglar, M. U., Hockenberry, A. J., & Wilke, C. O. (2018). Predicting bacterial growth conditions from mRNA and protein abundances. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0206634

Cossu, A., Le, P., Young, G. M., & Nitin, N. (2017). Assessment of sanitation efficacy against *Escherichia coli* O157: H7 by rapid measurement of intracellular oxidative stress, membrane damage or glucose active uptake. *Food Control, 71*, 293–300.

Gil, M. I., Selma, M. V., Lopez-Galvez, F., & Allende, A. (2009). Fresh-cut product sanitation and wash water disinfection: Problems and solutions. *International Journal of Food Microbiology, 134*, 37–45.

Kalyanaraman, B., Darley-Usmar, V., Davies, K. J. A., Dennery, P. A., Forman, H. J., Grisham, M. B., … Ischiropoulos, H. (2012). Measuring reactive oxygen and nitrogen species with fluorescent probes: Challenges and limitations. *Free Radical Biology and Medicine*. https://doi.org/10.1016/j.freeradbiomed.2011.09.030

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 2017-December*(Nips), 3147–3155.

Kitis, M. (2004). Disinfection of wastewater with peracetic acid: A review. *Environment International*. https://doi.org/10.1016/S0160-4120(03)00147-8

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets : A review. *Science*.

Li, H., & Phung, D. (2014). Journal of Machine Learning Research: Preface. *Journal of Machine*

476 *Learning Research*, *39*(2014), i–ii.

477Maillard, J. Y. (1996). Damage to pseudomonas aeruginosa PAO1 bacteriophage F116 DNA by
478 biocides. *Journal of Applied Bacteriology*. https://doi.org/10.1111/j.1365-
479 2672.1996.tb03254.x

480Morin, T., Martin, H., Soumet, C., Fresnel, R., Lamaudière, S., Le Sauvage, A. L., … Maris, P.
481 (2015). Comparison of the virucidal efficacy of peracetic acid, potassium monopersulphate
482 and NaOClon bacteriophages P001 and MS2. *Journal of Applied Microbiology*.
483 https://doi.org/10.1111/jam.12870

484

485Neumaier, T., Swenson, J., Pham, C., Polyzos, A., Lo, A. T., Yang, P. A., … Costes, S. V.
486 (2012). Evidence for formation of DNA repair centers and dose-response nonlinearity in
487 human cells. *Proceedings of the National Academy of Sciences of the United States of*
488 *America*. https://doi.org/10.1073/pnas.1117849108

489Newhart, K. B., Goldman-Torres, J. E., Freedman, D. E., Wisdom, K. B., Hering, A. S., & Cath,
490 T. Y. (2020). Prediction of Peracetic Acid Disinfection Performance for Secondary
491 Municipal Wastewater Treatment Using Artificial Neural Networks. *ACS ES&T Water*.
492 https://doi.org/10.1021/acsestwater.0c00095

493Oldenhof, H., Schütze, S., Wolkers, W. F., & Sieme, H. (2016). Fourier transform infrared
494 spectroscopic analysis of sperm chromatin structure and DNA stability. *Andrology*.
495 https://doi.org/10.1111/andr.12166

496Ovissipour, M., Rai, R., & Nitin, N. (2019). DNA-based surrogate indicator for sanitation

verification and predict inactivation of Escherichia coli O157:H7 using vibrational spectroscopy (FTIR). *Food Control*. https://doi.org/10.1016/j.foodcont.2018.12.017

Pascolo, L., Bedolla, D. E., Vaccari, L., Venturin, I., Cammisuli, F., Gianoncelli, A., … Ricci, G. (2016). Pitfalls and promises in FTIR spectromicroscopy analyses to monitor iron-mediated DNA damage in sperm. *Reproductive Toxicology*. https://doi.org/10.1016/j.reprotox.2016.02.011

Ravi, J., Hills, A. E., Cerasoli, E., Rakowska, P. D., & Ryadnov, M. G. (2011). FTIR markers of methionine oxidation for early detection of oxidized protein therapeutics. *European Biophysics Journal*. https://doi.org/10.1007/s00249-010-0656-1

Sahu, R. K., Argov, S., Salman, A., Huleihel, M., Grossman, N., Hammody, Z., … Mordechai, S. (2004). Characteristic absorbance of nucleic acids in the Mid-IR region as possible common biomarkers for diagnosis of malignancy. *Technology in Cancer Research and Treatment*. https://doi.org/10.1177/153303460400300613

Simonova, D., & Karamancheva, I. (2013). Application of Fourier transform infrared spectroscopy for tumor diagnosis. *Biotechnology and Biotechnological Equipment*. https://doi.org/10.5504/BBEQ.2013.0106

Sordo, M., & Zeng, Q. (2005). On sample size and classification accuracy: A performance comparison. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/11573067_20

Suslow, T. (1997). *Postharvest Chlorination: Basic Properties & Key Points for Effective*

Distribution. Postharvest Chlorination: Basic Properties & Key Points for Effective Distribution. https://doi.org/10.3733/ucanr.8003

Veronika, A., Vasily, D., & Kruchinin, E. D. (2018). Why every GBDT speed benchmark is wrong. *ArXiv*.

Zareef, M., Chen, Q., Hassan, M. M., Arslan, M., Hashim, M. M., Ahmad, W., … Agyekum, A. A. (2020). An Overview on the Applications of Typical Non-linear Algorithms Coupled With NIR Spectroscopy in Food Analysis. *Food Engineering Reviews*. https://doi.org/10.1007/s12393-020-09210-7

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. https://doi.org/10.1161/CIRCULATIONAHA.105.594929