ACS OMEGA

Open Access

This article is licensed under CC-BY-NC-ND 4.0

http://pubs.acs.org/journal/acsodf

Article

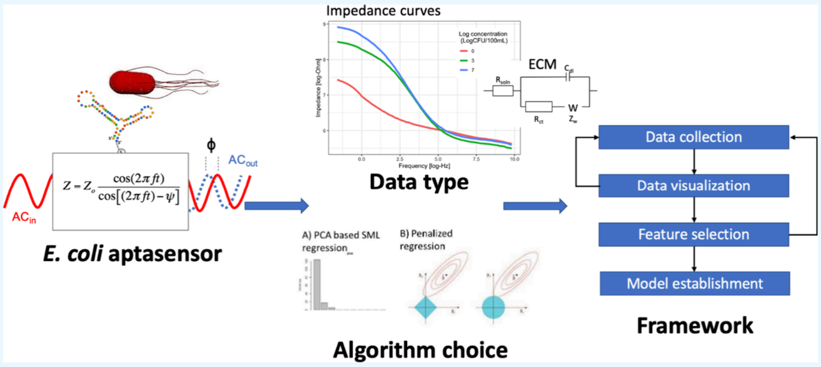
Machine Learning for Improved Detection of Pathogenic *E. coli* in Hydroponic Irrigation Water Using Impedimetric Aptasensors: A Comparative Study

Hanyu Qian, Eric McLamore,* and Nikolay Bliznyuk*

Cite This: *ACS Omega* 2023, 8, 34171−34179

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Reuse of alternative water sources for irrigation (e.g., untreated surface water) is a sustainable approach that has the potential to reduce water gaps, while increasing food production. However, when growing fresh produce, this practice increases the risk of bacterial contamination. Thus, rapid and accurate identification of pathogenic organisms such as Shiga-toxin producing *Escherichia coli* (STEC) is crucial for resource management when using alternative water(s). Although many biosensors exist for monitoring pathogens in food systems, there is an urgent need for data analysis methodologies that can be applied to accurately predict bacteria concentrations in complex matrices such as untreated surface water. In this work, we applied an impedimetric electrochemical aptasensor based on gold interdigitated electrodes for measuring *E. coli O157:H7* in surface water for hydroponic lettuce irrigation. We developed a statistical machine-learning (SML) framework for assessing different existing SML methods to predict the *E. coli O157:H7* concentration. In this study, three classes of statistical models were evaluated for optimizing prediction accuracy. The SML framework developed here facilitates selection of the most appropriate analytical approach for a given application. In the case of *E. coli O157:H7* prediction in untreated surface water, selection of the optimum SML technique led to a reduction of test set RMSE by at least 20% when compared with the classic analytical technique. The statistical framework and code (open source) include a portfolio of SML models, an approach which can be used by other researchers using electrochemical biosensors to measure pathogens in hydroponic irrigation water for rapid decision support.

1. INTRODUCTION

Escherichia coli (*E. coli*) is one of the most diverse and widely distributed organisms on the planet. Pathogenic *E. coli* includes Shiga-toxin-producing *E. coli* (STEC) such as O157:H7, and enterotoxigenic *E. coli*, among others.[1] The existence of STEC in food, water, or in sporadic cases air poses a risk to human health. STEC may enter the human body through food consumption, water ingestion, or inhalation. After entering the human body, STEC may cause damage to intestinal lining, resulting in disruption of homeostasis within the gastrointestinal tract microbiota.[2] In the United States alone, the CDC reports that at least 47.8 million illnesses, 127,000 hospitalization, and more than 3000 deaths are caused by known foodborne pathogens; STEC contributes more than 265,000 illnesses, 2100 hospitalizations, and 30 deaths annually.[3]

The demand for food is increasing with a world population growth. Increasing food demand causes water shortage in many areas, as approximately 70%, perhaps as high as 90% of the freshwater demand is from agriculture in most regions of the world.[4] Agricultural demand for freshwater competes with other needs (e.g., drinking water), creating pressure on the

Received: August 7, 2023
Accepted: August 28, 2023
Published: September 10, 2023

© 2023 The Authors. Published by American Chemical Society

agricultural industry to use alternative water sources for irrigation. However, bacteria (often found in alternative waters) may contaminate fresh produce and exacerbate public health risks.[5] Irrigation water, particularly treated wastewater, is one of the main culprits of fresh produce contamination by bacteria, including STEC.[6] As few as 10 to 100 viable STEC in or on fresh produce can cause human illness and in some cases death.

To reduce the potential for disease burden in the U.S. food supply chain, regulations call for regular monitoring. The regulation for irrigation water applied in fresh produce production is set by the produce safety rule (PSR).[7] According to PSR, agricultural water used during growing activities must have a microbial water quality profile that meets the following:[8] for a rolling four-year data set of water testing, the profile must not exceed ≤126 CFU of generic *E. coli* per 100 mL sample, with a statistical threshold value of ≤410 CFU generic *E. coli* per 100 mL sample. Thus, effective detection of *E. coli* is critical to modern food safety systems.

Biosensors offer rapid detection of bacteria, and most do not require complex instrumentation or operator training.[9] As reviewed by many others,[10] there are numerous types of biosensors for detection of bacterial pathogens in food and water samples. The most common biosensors for food safety utilize various combinations of biomaterials and nanomaterials on a conductive electrode, where binding of bacteria transduces an electrochemical signal.[11] Among the transduction types used in bacteria sensing, impedimetric sensors are common. For example, Soares et al.[12] developed an impedimetric immuno-sensor to detect *S. enterica* in chicken broth. This immunosensor was tested in laboratory conditions, and data analysis employed post hoc equivalent circuit modeling (ECM) to determine charge transfer resistance (R_{ct}) by chi-squared fitting, a classic method of data analysis. Another recent example is the aptasensor by Sidhu et al.[9] This aptasensor was developed to detect *Listeria* spp. in hydroponic lettuce water under flowing conditions using platinum interdigitated microelectrodes biofunctionalized with aptamers.[9] ECM was used to determine R_{ct}, which was used to estimate the target concentration. Expanding on these works, Giacobassi et al.[13] developed a partially automated label-free impedimetric aptasensor based on net impedance (Z) as the response variable. This aptasensor was designed to detect generic *E. coli* in hydroponic water and used the output data (Z) for water pump control based on a priori thresholds.

In these approaches above (and other electrochemical techniques common for pathogen detection), data analysis depends on either post hoc ECM or a priori establishment of a response variable threshold. Neither of these techniques is ideal for system operation in a dynamic environment with a complex sample matrix as the prediction accuracy is either low or unknown. The lumped circuit abstraction used for ECM is a powerful tool for analyzing impedimetric data but may oversimplify the complex biosensor system because the ECM assumes a single and ideal circuit system to represent the complex and dynamic iterations within the biosensor system, potentially leading to decreased prediction accuracy.[14] Furthermore, the computational burden associated with ECM can be relatively high, which may restrict its practical application in real-time biosensor monitoring systems. To improve accuracy and working efficiency (particularly for large data sets characterized by high data dimensionality), machine learning has emerged as a powerful analytical tool for a wide range of biosensor applications.[15−17]

Machine learning (ML) models can effectively learn from large biosensor data sets without preprocessing or a priori thresholds, reducing the data dimension and avoiding assumptions made in ECM. ML is capable of partially automating the extraction of new relevant information, in addition to that available in the equivalent circuit summary reduction, to improve the predictive performance (e.g., for bacterial concentrations in water).[18] When applied to impedimetric biosensor data sets where equivalent circuit analysis lacks desired resolution, ML can be a powerful tool for improving prediction of target concentration. Statistical ML (SML) techniques, an important subtype of machine-learning research, are grounded in building and validating, with the help of training data, a statistical predictive model for new (unobserved) data. SML has the potential to provide direct readout in a partially autonomous manner without requiring the usage of ECM to preprocess the data, which is a major improvement over the currently available technique.

The goal of this manuscript was to develop a framework for applying and evaluating existing SML algorithms to analyze electrochemical aptasensor data, which is a critical step toward a real-time decision support for food safety diagnostics. One of the most common aptasensors (gold interdigitated electrodes) were applied for measuring the concentration of *E. coli O157:H7* in lettuce irrigation water. This work builds upon previous aptasensors for irrigation water[9,12,13,19] by applying the framework herein to compare the candidate approaches and, consequently, to identify the optimum analytical method for (out-of-sample) prediction of *E. coli O157:H7* concentration. We investigated the operational characteristics (e.g., test set root mean squared error, RMSE) for a suite of SML approaches[20,21] for predicting *E. coli O157:H7* concentration. The developed framework includes data visualization, feature selection, model validation, and refinement (hyperparameter tuning and model comparison), establishing a systematic approach to analyze aptasensor data for irrigation water. It allowed us to investigate not only the impact of the selection of the ML algorithm but also the relative utility of the data type, i.e., ECM outputs, raw impedance curves, or both. In the example case shown here, the use of SML yielded an improvement over the conventional post hoc ECM used in conventional aptasensing. The SML models were compared with nonlinear models based on the ECM in terms of the prediction ability. The best SML algorithm exhibited a reduction in the test set RMSE of at least 20%. The data and code used to establish the model are open access and may be generalized to other biosensors used in water quality diagnostics and may have application in other domains where pathogen prediction in water is required.

2. METHODOLOGY

2.1. Biosensor Fabrication and Impedance Testing. *E. coli O157:H7* (ATCC 25922; FDA strain 1946) was cultured on tryptic soy agar (TSA) slants at 4 °C and stored at 4 °C when not in use. Before experiments, serial dilutions were made to achieve concentrations from 1 to 10^7 CFU/100 mL. All samples were confirmed using the Colilert Quantitray method.[22]

Aptamer-based biosensors (aptasensors) were fabricated using protocols described in our previous work.[19,23,24] In summary, aptamer P12−55 was thiol-terminated with a C6 spacer for covalent binding to gold interdigitated electrodes (IDE). Aptamers were refolded, decapped, drop cast on gold IDE, and stored in a Petri dish at room temperature for 20 min. Electrochemical impedance spectroscopy (EIS) data were recorded using a portable potentiostat.[25] The analytical sensitivity and limit of detection (LOD) of the aptasensors are listed in Table S9.

A 50 L hydroponic lettuce system was irrigated with pond water using an approach similar to Giacobassi et al.[13] (see supplemental for details). Hydroponic lettuce (*Lactuca saliva*) was germinated in sterile foam cubes and then transferred to CocoTek-lined grow cups (7.6 cm diameter) containing expanded clay pellets (Stacky Hydroponic Center, Lake City, FL, USA). Irrigation water was collected in sterile 5.0 L plastic bottles at a site on the southern end of a reservoir in Gainesville, FL (Lake Alice), with low algal growth. These samples were used to exchange 10% of the irrigation water each day in the 50 L hydroponic system. Samples (100 mL) were taken from the reservoir tank, and each sample was measured by 15 unique (replicate) IDE aptasensors based on Giacobassi et al.[13] An aliquot of cultured cells was added to achieve bacteria concentration (contrived) ranging from 1 to 10^7 CFU/100 mL. After spiking bacteria, aliquots (5 mL) were taken for validation via Quantitray. Aptasensors were directly immersed in a 100 mL sample bottle.

2.2. Overview of Aptasensor Impedance Data.

Eight contrived *E. coli O157:H7* samples were prepared with concentrations from 0 to 7 log CFU/100 mL. After collecting EIS data, ECM was used to determine charge transfer resistance (R_{ct}), solution resistance (R_s), Warburg impedance (Z_w), and double layer capacitance (C_{dl}) from complex plane diagrams based on Sidhu et al. (2020).[9] ECM parameters (R_s, R_{ct}, C_{dl}, and Z_w) were determined by Chi^2 fitting using the Randles Ershler model.

Various combinations of raw EIS data (real and imaginary impedance and net impedance) and/or ECM parameters were input to the statistical models. For each EIS curve, the data dimension was 146, which is the sum of the response variable (either real or imaginary impedance) and the number of unique frequencies for each EIS scan.

As our primary purpose is out-of-sample prediction using a new sensor (sensor ID was a block in the statistical design), we randomly sampled (without replacement) sensor IDs to assign the data (on all eight concentrations for each ID) to the training (80%) and test (20%) sets. We defined the scenario where the training set and test set shared the same concentration range as "observed concentration." We also used another sampling approach where we sampled (without replacement) concentration to assign the data to training set and test set, which was defined as "unobserved concentration."

In order to understand the drivers of predictive performance, we applied three classes of statistical models: nonlinear regressions combined with ECM parameters as inputs, SML models solely relying on aptasensor raw EIS data (the number of predictors is 146), and (hybrid) statistical models utilizing the combined information across ECM parameters and EIS data (see Sections 2.3, 2.4, and 2.5, respectively). Models were fitted to the training data, and performance was evaluated using RMSE on the complementary test set. The response of the models was *E. coli O157:H7* concentration (in log scale, i.e., 0 to 7 log CFU/100 mL), and the test set RMSE was calculated between real concentration and predicted concentration.

In this manuscript, we define the "aggregated concentration data set" as the data set in which the *E. coli O157:H7* concentrations of 4 log CFU/100 mL and above 4 log CFU/100 mL were aggregated to 4 log CFU/100 mL. This represents the data that displayed signal saturation with limited hook effect. The "nonaggregated (original) concentration data set" was defined as the data set before the aggregation process.

2.3. Nonlinear Regression and Generalized Additive Models (GAMs).

Nonlinear regression, where response y depends on corresponding predictor x via a nonlinear regression function f, was applied to analyze ECM parameters as

$$y_i = f(x_i, \theta) + \epsilon_i$$

where y is the model response−concentration of *E. coli O157:H7* for each observation; x is the predictor (ECM parameters: R_{ct} and C_{dl}); θ is a vector of regression function parameters; ϵ is the error term, and i is the observation ID. The relationship between *E. coli O157:H7* concentration and aptasensor signal was modeled as a parametric nonlinear function, e.g., logistic or asymptotic functions as shown below:

$$f(x_i, \theta) = c + (d - c)\cdot\frac{\exp(\beta x_i + \beta_0)}{1 + \exp(\beta_0 + \beta x_i)}$$

$$f(x_i, \theta) = c + (d - c)\cdot(1 - \beta_0 - \exp(-\frac{x_i}{\beta}))$$

The model parameters c and d are the lower and upper bounds of response; β and β_0 are nonlinear regression coefficients. We used ordinary nonlinear least-squares method to estimate the model parameters.[26] The nonlinear regressions (as noted) were fitted based on the observations detected by sensors in the training set, and then the fitted models were used to predict the observations in the test set. Additionally, the random effect caused by the potential sensor-to-sensor variation was considered in the nonlinear regression. Mixed-effect nonlinear regression model structures are shown in Appendix S13.

GAMs were applied to estimate semiparametric[27,28] smooth nonlinear associations between the response (*E. coli O157:H7* concentration in log scale) and the covariates of interest (ECM parameters). In addition, fixed-effects and mixed-effects (aptasensor-specific random effects, intercepts) models were fit to predict the observation concentration detected by aptasensors with different IDs in test set.

2.4. SML Models Using Impedance Data as Features.

SML models were used to predict the *E. coli O157:H7* concentration based on information from impedance curve features. The hyperparameters in SML models were tuned by block cross-validation (see Appendix S14 for details). Two different approaches were used as summarized in the next sections.

2.4.1. SML Models for Raw EIS Data.

Ridge, partial least-squares (PLS) and gradient boosting decision tree models were fitted, which directly use the raw EIS data as inputs. Ridge regression is a technique to prevent overfitting by adding a penalty term, which is sum of squared coefficient (L_2 penalty), in the cost function.[29] PLS, a supervised alternative to principal components analysis (PCA), constructs a set of linear combinations of predictors x to a set of new features, and then fits a linear model using these new features.[29] We also applied the gradient boosting decision tree and XGBoost decision tree to predict the *E. coli O157:H7* concentration,[30] see Appendix S6 for details.

2.4.2. SML Models for Dimension-Reduced Data. Dimension reduction such as PCA and functional data analysis techniques allows one to control the unstructured variation in high-dimensional covariates (here, raw discretized impedance curves) by constructing a small set (e.g., 2−5) of orthogonal linear combinations of the original covariates that explain a large proportion (e.g., 95%) of the total variation in the covariates. Alternative variable selection approaches for dimensionality reduction[31,32] have been considered but are not evaluated here. We first applied PCA to reduce the data dimension by projecting covariates onto only a few principal components (PCs). This produced lower-dimensional data in which the first two PCs explained 96.5% of the raw data variation. Therefore, the first two PCs were used as inputs of SML models.

Using the first two PCs as the fixed effects, we fit a mixed-effects model with a random intercept dependent on the identifier of the biosensor:

$$y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + Zu + \epsilon$$

where Z is the design matrix for the random effect caused by the sensor-to-sensor variation, u is a vector of random effect, PC_1 and PC_2 are the first two PCs from PCA, and ϵ is the error term. By testing the significance of the random effect, we evaluated the influence of the block effect in our study.

In addition to linear regression, we also applied a support vector machine (SVM), boosted linear regression, and random forest models, using the first two PCs from PCA as predictors. The aim was to predict *E. coli O157:H7* concentration. Compared with linear regression, SVM regression is appropriate for nonlinear behavior between the input variables and response and depends on selection of appropriate kernels.[33,34] In this work, we analyzed a set of kernels, including linear, polynomial, and radial basis function (RBF) kernel. Boosted linear regression, also known as gradient boosting regression, can usually produce more accurate prediction compared with traditional linear regression.[30] Random forest is a tree-based regression which is especially effective in handling noisy data.[35] All of the code was prepared in the R language caret package.

2.5. Combining Information across ECM and Impedance Data. We applied two different approaches to combine the ECM-derived features and the impedance data features: (1) the ECM parameters and 146 features from the impedance curves were used together as predictors in one SML model to predict the *E. coli O157:H7* concentration and (2) an intermediate variable score was first generated from the SML models. Then the ECM parameters were used along with the score as features in a GAM to predict the *E. coli O157:H7* concentration (see Figure 1).

2.6. Model Comparisons. To have an accurate evaluation of predicting *E. coli O157:H7* concentration, we repeated the sampling process of the training set and test set 100 times and calculated the test set RMSE of each model under each sample seed. The mean of the achieved RMSE vector was calculated to estimate the test set RMSE for each model, and the 95% confidence interval of estimation was constructed using the 2.5% and 97.5% sample percentiles. This estimation approach allows us to avoid sampling influence and estimate the uncertainty of prediction error. The paired t test was further conducted to compare the prediction error between two models using the achieved RMSE vectors (see Appendix S15 for more details).

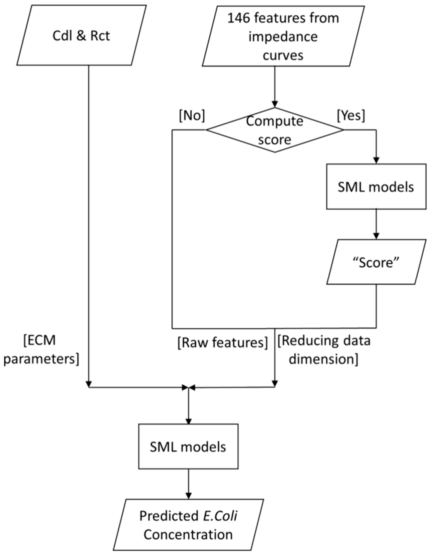

Cdl & Rct
146 features from impedance curves
[No]
Compute score
[Yes]
SML models
"Score"
[ECM parameters]
[Raw features]
[Reducing data dimension]
SML models
Predicted *E.Coli* Concentration

Figure 1. Workflow for ML analysis of impedimetric *E. coli O157:H7* aptasensor data. Two different approaches were used to combine the ECM-derived features with the (raw) EIS data.

3. RESULTS AND DISCUSSION

3.1. Conventional ECM-Based Models. Figure 2 shows representative EIS data for baseline (0 CFU/mL) and 10^4 CFU/mL *E. coli O157:H7* in hydroponic water. At low frequency, the phase (Figure 2a) is relatively stable and indicates a capacitive behavior (phase angle of approximately −40° to −50°). The low frequency region of the EIS curve (below 0.1 Hz) coincides with a near-linear decrease in net impedance (Figure 2)b, which is similar to other impedimetric aptasensors in the literature.[19,23,24] A distinct difference is apparent in Nyquist plots that indicates a significant change in R_{ct} and C_{dl} (the semicircular region in Figure 2c), which was consistent across the operating range (0 CFU/100 mL and 10,000 CFU/mL; see supplemental section for all raw curves).

Nonlinear regression and GAM were analyzed using ECM parameters as predictors across the entire operating range. Figure 3a,b shows the relationship between ECM parameters (R_{ct} and C_{dl}) and the *E. coli O157:H7* concentration. The curves follow classic response with a linear region at low concentration and a zero-order region that is indicative of receptor saturation at high concentration. A hook effect occurs at concentrations greater than 10^6 CFU/100 mL, which is common in biosensor calibration curves. The prediction accuracy for the relative high concentration (C > 4 log CFU/100 mL) from the statistical models which use ECM parameters as predictors would be low because there is much less variation of the ECM parameter value when concentration is larger than 4 log CFU/100 mL due to signal saturation. Additionally, one can define high concentrations as the same value in the context of food safety applications, where the food safety threshold (126 CFU/100 mL) is orders of magnitude below the cutoff value (4 log CFU/100 mL). Thereby, we aggregated the concentrations of 4 log CFU/100 mL and above to 4 log CFU/100 mL during the nonlinear regression model fitting process. We also calculated the correlation

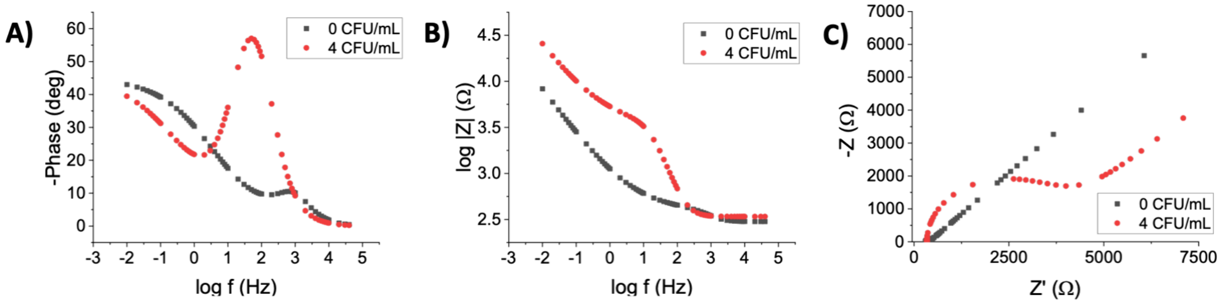


Figure 2. Representative EIS plots for *E. coli O157:H7* aptasensor in hydroponic media at 25 °C. Baseline (0 CFU/100 mL) and upper range (10,000 CFU/100 mL) are shown as a representative example. Representative (A) phase plot, (B) bode plot, and (C) Nyquist plot.

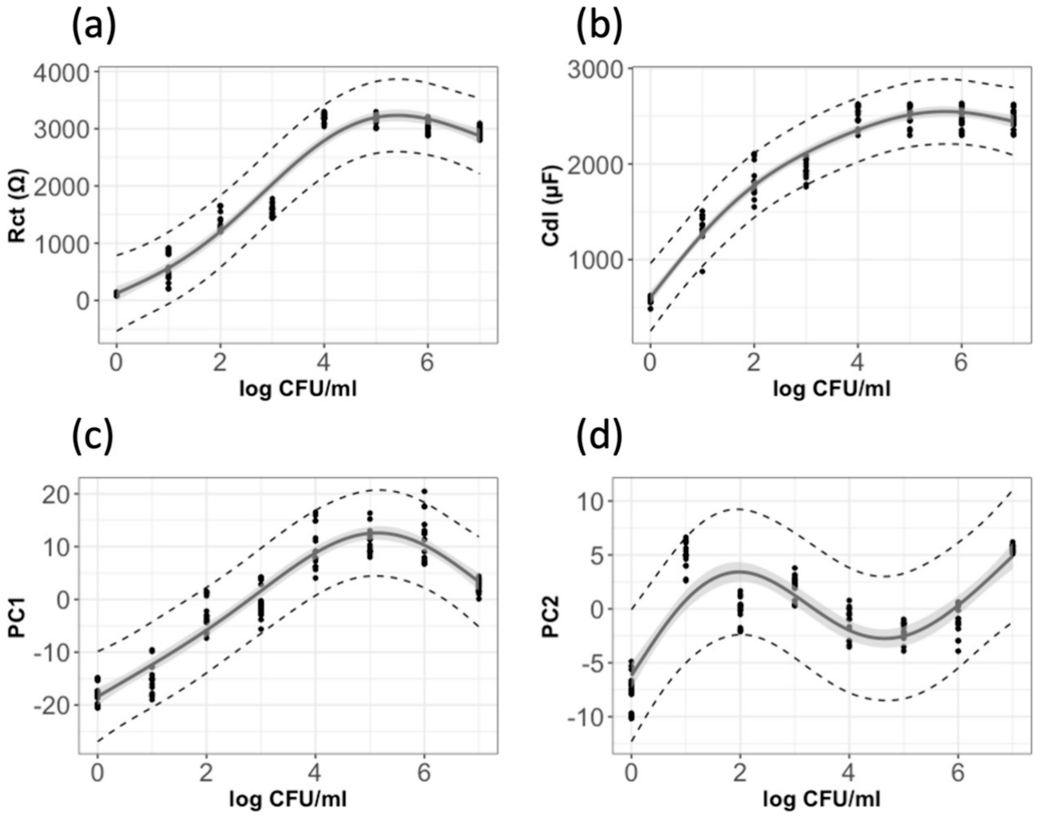


Figure 3. Relationship between ECM parameters ((a) and (b)) and the first two PCs ((c) and (d)) with *E. coli O157:H7* concentration in irrigation water. The symbols in Figure 3 are the observed values, and solid lines are the estimated mean function from the GAMs which used *E. coli O157:H7* concentration as predictors and ECM parameters as response. The gray shaded area represents the 95% pointwise confidence interval, and the pointwise intervals between lower and upper dashed lines are the 95% prediction intervals.

coefficients between R_{ct} and C_{dl}, and the results are presented in appendix Section S4.

Figure 3c,d illustrates the relationship between the first two PCs from analysis of data for *E. coli O157:H7* concentration (0 to 7 log CFU/100 mL). The inputs of PCA were the 146 features of aptasensor impedance curves. Outputs from PCA were used as the predictors in some SML models, as noted, and the results are discussed in Section 3.2. The first PC retains the features observed in raw data calibration curves (nonlinear response at lower concentration with a hook effect at high concentration), however, PC2 does not (oscillating response with high uncertainty).

Using ECM parameters R_{ct} and C_{dl} as predictors, the performance (for aggregated data set) of some widely used nonlinear regressions is reported in Table 1 (the result for mixed-effect nonlinear regression is reported in Table S7). The average RMSE column represents the average test set RMSE across the 100 simulations (one simulation represents one random split of data into a training set and test set). The CI_{lower} and CI_{upper} are lower and upper bounds of the 95% confidence interval of RMSE for the test set.

Table 1. ECM-Based Nonlinear Regression Average RMSE and 95% Confidence Interval for Aggregated Data Set[a]

nonlinear regression	average RMSE	CI_{lower}	CI_{upper}
logistic function	0.244	0.118	0.360
asymptotic function	0.258	0.155	0.337
GAM	0.260	0.156	0.346

[a] Data with concentration above 4 log CFU/100 mL were aggregated due to signal saturation.

Among the models tested, the logistic function had the lowest prediction error (average RMSE = 0.244). Considering the fact that nonlinear regression with a logistic function is a simple and efficient model, and its results are easy to interpret, we select it as the baseline model to compare with SML models.

3.2. SML Models with Raw EIS Data as Input. Using raw EIS data as inputs in PCA, the first two PCs explain 96.5% of the data variance. Using these two PCs as fixed effects, we established a mixed-effects model with random intercept capturing variation due to the sensor ID. The variance of the random effect (standard deviation = 2.90×10^{-5}), which represents the variability among the sensors with different IDs, is much smaller than the variance of residual term which is defined as the error that cannot be explained by the model (standard deviation = 2.03). The *p*-value of the random effect term is also larger than 0.05, indicating that sensor-to-sensor variation is negligible when PCA is used for these Au-IDE aptasensors.

The performance of select SML models is shown in Figure 4 (see Table S8 for additional details in the tabular form).

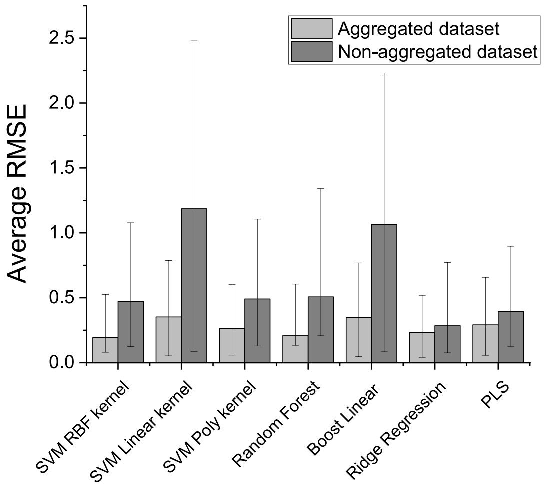


Figure 4. Predictive ability of SML models (aggregated data set for concentrations ≥4 log CFU/100 mL) was compared with nonlinear regression. In addition, the predictive ability for the nonaggregated data set was compared within SML models. The average RMSE column is the average test set RMSE across the 100 simulations, which has the same sampling seed as that of the simulation used inSection 3.1.

Aggregating the data for bacteria concentration higher than 4 log CFU/100 mL significantly reduced RMSE for SML models using the first two PCs as predictors but had a less pronounced effect on SML models based on EIS raw data without dimension reduction (ridge regression and PLS).

The SVM model with the RBF kernel, which uses the first two PCs as the predictors, has the lowest prediction error for the aggregated data set (average RMSE = 0.194). The random forest model has a wide confidence interval, as do XGBoost and gradient boosting tree models, indicating that these models are sensitive to the outliers in the test set under some of the simulations of the test set and training set (see Appendix S6). Compared with the baseline model (logistic function), the best SML models (SVM with RBF kernel, which has the lowest average RMSE) can reduce the average RMSE by 20%. Using the paired *t* test to compare SVM with RBK kernel model test set RMSE vector, obtained from 100 simulations, with the baseline model (ECM-based logistic function), the *p*-value was less than 0.001. This indicates that the SVM model has a significantly lower prediction error.

SML models predictive ability decreases from aggregated data set to nonaggregated data set. This may be due to the noted hook effect at high bacteria concentration. Among the selected SML models, the ridge regression has the best predictive ability for the nonaggregated data set (average RMSE = 0.285). Comparing the performance of SML models that use the original (see Section 2.4.1) and dimension-reduced (see Section 2.4.2) predictors, we found that the increasing rates of average RMSE (from aggregated to nonaggregated data set) for impedance curves data-based SML models (without PCA dimension reduction) are lower than PCA-based SML models. For example, the average RMSE increased 140% from aggregated to nonaggregated data set for SVM, which uses the first two PCs as predictor, while the average RMSE increases 22% for ridge regression, more details are discussed in Appendix S7.

3.3. Hybrid Statistical Method Combining Information from ECM and Raw EIS Data. This section focuses on improving predictive ability of SML models for the aggregated data set. We applied two methods to combine the information between the ECM and aptasensor impedance curves. In the first method, we combined two ECM parameters with 146 features from the raw EIS data. Based on the combined data set (data dimension is 148), we used PCA to reduce data dimension and then used the first two PCs as inputs to fit SVM regression. We also used ridge regression to directly analyze the combine data set. The results of method 1 are shown in Table 2 ("ECM parameters + original EIS data" columns). In the second method, we first generated an intermediate covariate "score" from a ridge or SVM regression. Then we fit a GAM model to predict the *E. coli O157:H7* concentration, in which the inputs were the score and two ECM parameters (see Table 2 "ECM parameters + intermediate score variable" columns). Specific details about these two methods are introduced in Section 2.5.

Table 2. Comparison of Using Two Methods To Improve SML Models Predictive Ability

	ECM parameters + original EIS data			ECM parameters + intermediate score variable		
selected SML models	average RMSE	CI_{lower}	CI_{upper}	average RMSE	CI_{lower}	CI_{upper}
SVM	0.191	0.110	0.327	0.091	0.016	0.241
ridge	0.140	0.100	0.210	0.060	0.020	0.210

The first method, which directly combines two data sets, has a similar average RMSE and 95% confidence interval with the SML models using only impedance curves (i.e., excluding ECM parameters). Using a paired *t* test to compare RMSE (as detailed in Section 2.6), the *p*-value was 0.693, indicating that the first method does not significantly improve the predictive ability for SML models. On the other hand, method 2 (scoring system approach) can significantly decrease the average RMSE, especially for ridge regression. More details about the significance test of each term in GAM are shown in Appendix S5.

3.4. Predictive Ability of SML for Unobserved Concentrations. In previous sections, we discussed the predictive ability for each statistical model using observed concentrations (the training set and test set shared the same concentration set but are based on different sensor IDs). To challenge this, the selected SML models may be used to predict the concentration, which is out of the set of standard solution concentration. Thereby, this section focuses on comparing predictive performance for a new scenario of unobserved concentrations (chosen from the nonaggregated data set), by allocating disjoint sets of the concentrations to the training set and test set. We allocate observations from two of the eight concentrations to the test set, and the remaining observations are grouped into the training set. This sampling process was repeated to produce 100 training/test data sets. Then, representative models were selected to fit the training data and predict the complementary disjoint test set. The average RMSE and 95% confidence interval are shown in Table 3 "C 0−7 log CFU/100 mL" columns.

Table 3. Predictive Ability for Unobserved Concentration in the Test Set

	C 0−7 log CFU/100 mL			C 1−6 log CFU/100 mL		
model	average RMSE	CI_{lower}	CI_{upper}	average RMSE	CI_{lower}	CI_{upper}
SVM	2.299	0.919	3.949	1.459	0.662	2.275
ridge	2.490	0.759	8.756	1.007	0.455	1.421
PLS	1.423	0.259	4.240	0.524	0.259	0.747

The ridge regression model, which has good performance for the observed concentration (average RMSE = 0.285), has a weak predictive ability for the unobserved concentration (average RMSE = 2.490). PLS has the best predictive ability (average RMSE = 1.423), in terms of the unobserved concentration. There are 56 potential combinations if we sampled two of eight concentrations to the test set. We calculate the test set RMSE for each sampling combination to find out why some SML models had large RMSE for unobserved concentration, and the results are shown in the appendix (Table S1). If we allocate the boundary concentrations to the test set, i.e., concentration 0 or 7 log CFU/100 mL, then most models will have a large prediction error. However, in practice, we require that the range of the standard solutions covers the potential sample solution. Thus, we also investigated the model performance when the boundary concentration values have been removed. After removing those outliers, PLS, ridge, and SVM predictive ability would significantly increase, but PLS still had the best performance (see Table 3 "C 1−6 log CFU/100 mL" columns).

3.5. Implications for Water Quality Analysis.

Our study introduced a rapid aptasensor for detection of *E. coli O157:H7* in agricultural (irrigation) water and a series of analytical tools for predicting concentration. Using the SML models to predict the *E. coli O157:H7* concentration, we demonstrated the ability to process EIS data without the use (or need) of ECM modeling to extract parameters, significantly reducing analytical burden while also producing prediction accuracy to inform user(s). The study was based on robust SML models which have been successfully applied in many other studies, which may improve the likelihood that the framework has general applicability in hydroponic water quality and/or food safety applications. Use of the framework here allows future researchers to compare numerous models for optimizing predictive ability. In this study, use of SML models significantly improved accuracy compared with the conventional ECM-based nonlinear regression commonly used in biosensing.

For the Au-IDE aptasensors shown here, the appropriate models for data analysis are ridge, SVM, and PLS to predict the *E. coli O157:H7* concentration in lettuce hydroponic water at concentrations relevant to the food safety modernization act (FSMA). In future studies, the collection of models may be applied to different types of biosensors (e.g., protein-based sensors, phage biosensors, etc.) or different samples using the open-source code and framework. Importantly, the script includes an embedded scoring system, which has numerous potential applications in decision support research.

In addition to the SML models built from the EIS data and nonlinear regressions from ECM data, we developed a new approach that statistically combines the EIS data and ECM parameters. This approach can significantly improve the predictive ability for the aggregated data set in this study of Au-IDE aptasensors compared with the existing method (ECM-based nonlinear regression). However, because of the relatively low amount of data for training and tests, more studies are needed to validate this conclusion.

The choice of data types (raw EIS data, ECM parameters, or both) used in the analysis can significantly influence the prediction accuracy of *E. coli O157:H7* concentrations. As described in Section 3.1, the parameters obtained from ECM (R_{ct} and C_{dl}) were constant when *E. coli O157:H7* concentrations were larger than 4 log CFU/100 mL (due to receptor saturation), and the signal begins to decrease above a concentration of 6 log CFU/100 mL (due to a hook effect). Thus, use of classic ECM alone lacks the ability to analyze high concentrations and may produce false negatives for extremely high concentrations. In contrast, ML algorithms which use raw EIS data as predictors exhibit good prediction accuracy for large concentrations, which could extend the aptasensors working range (see Section 3.2).

In both EIS-based SML and ECM-based nonlinear regression, we examined the influence of the random effect caused by sensor-to-sensor variation in *E. coli's O157:H7* concentration prediction (see Sections 3.1 and 3.2). The results show that the random effect is negligible, indicating excellent device consistency. Additionally, the smooth impedance curve shown in Figure S4 indicates a high signal-to-noise ratio, which enhanced the aptasensor detection ability. This is likely a result of developing the aptamer on Au-IDE commercial electrodes, which is a proven model system in microbial biosensors.[9,36] However, it is common to encounter noisy data with a low signal-to-noise ratio in other devices, such as electrodes fabricated with emerging materials that have not yet been commercialized.[37−39] Biosensor signal noise may originate from operator variation, sensor-to-sensor variation, or environmental influences, which would ultimately be remediated with experimental design techniques (such as principled replication once the sources of variation have been quantified). ECM-based nonlinear regression shows a weaker predictive ability compared with SML methods even though the data sets have a high signal-to-noise ratio. ML algorithms have various techniques to handle noise effectively, such as robust learning, regularization, etc.[40,41] Although the model system shown in this study does not demonstrate the advantages of SML models for noisy data, the framework includes techniques that handle these problematic issues and thus has high translational potential to other biosensors. For example, the L_2 regularization used in ridge regression penalizes the model for using less informative signals as predictors, reducing the impact of noisy features in the data. The open-access code and program that we provide can be easily adapted to projects that face challenges with noisy data for other biosensor systems.

The discrete concentration set used in the analysis posed a limitation in this study as it made it challenging to evaluate the model predictive performance for samples with noninteger log-concentrations. This limitation arose because the research utilized serial dilutions to prepare *E. coli O157:H7* samples (refer to Section 2.1). In our future work, we aim to address these limitations by focusing on two main areas. First, we will validate the developed SML frameworks for additional samples using noninteger concentrations. Second, we plan to apply the SML framework to different biosensors.

4. CONCLUSIONS

Use of environmental surface water for irrigation is an important agricultural practice to ensure food and water security. However, the potential for contamination by microbial pathogens is a major risk, requiring real-time analytical tools for the rapid screening of water quality. Nonlinear equivalent circuit models produce multiple circuit parameters that may not be associated with the underlying physics of the biorecognition or transduction components of the sensor, especially when the bacteria concentration is large (due to common problems, such as the hook effect). Here, we developed a SML model framework for optimizing model selection and demonstrated the approach for impedimetric aptasensor impedance curves to predict *E. coli O157:H7* concentration. In this case study, the SML model had higher accuracy compared with classic equivalent circuit models based on nonlinear regression. When coupled with appropriate SML techniques, the impedimetric aptasensors allow one to improve (over the conventional ECM-based analysis) the prediction accuracy for rapid detection of *E. coli O157:H7* in agricultural waters. This study is a critical step toward real- time decision support for food safety diagnostics in sustainable water reuse.

■ ASSOCIATED CONTENT

Data Availability Statement

All data generated or analyzed during this study are included in this published article.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c05797.

Details on experimental set up, raw data visualization, and modeling details (PDF)

Project data (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Eric McLamore − *Department of Agricultural Sciences, College of Agriculture, Forestry and Life Sciences, Clemson University, Clemson, South Carolina 29634, United States;* orcid.org/0000-0002-1662-7372; Email: emclamo@clemson.edu

Nikolay Bliznyuk − *Department of Agricultural and Biological Engineering and Departments of Statistics, Biostatistics and Electrical & Computer Engineering, University of Florida, Gainesville, Florida 32611, United States;* orcid.org/0000-0001-7118-7907; Email: nbliznyuk@ufl.edu

Author

Hanyu Qian − *Department of Agricultural and Biological Engineering, University of Florida, Gainesville, Florida 32611, United States;* orcid.org/0000-0002-9730-4902

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c05797

Author Contributions

H.Q.: formal analysis, methodology, and writing—original draft; E.M.: conceptualization, supervision, funding acquisition, and writing, review, and editing; and N.B.: conceptualization, data-analytic strategy, supervision, and writing, review, and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2018-67016-27578 awarded as a Center or Excellence from the USDA National Institute of Food and Agriculture. This work was supported by University of Florida Agricultural and Biological Engineering Department Pathfinder fellowship program. The authors thank Dr. Diana Vanegas for key insights and assistance with manuscript review.

■ REFERENCES

(1) Kai, A.; Konishi, N.; Obata, H. [Diarrheagenic Escherichia Coli] Nippon rinsho. *Jpn. J. Clin. Med.* **2010**, *68* (1), 203−207.

(2) Fischer Walker, C. L.; Applegate, J. A.; Black, R. E. Haemolytic-Uraemic Syndrome as a Sequela of Diarrhoeal Disease. *J. Health Popul. Nutr.* **2012**, *30* (3), 257−261.

(3) Center for Disease Control and Prevention. *National Enteric Disease Surveillance: STEC Surveillance Overview Surveillance System Overview: National Shiga Toxin-Producing Escherichia Coli (STEC) Surveillance*, 2012, 3, DOI: 10.4414/smw.2012.13719.

(4) Hoekstra, A. Y.; Mekonnen, M. M. The Water Footprint of Humanity. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (9), 3232−3237.

(5) Rani, A.; Ravindran, V. B.; Surapaneni, A.; Mantri, N.; Ball, A. S. Review: Trends in Point-of-Care Diagnosis for Escherichia Coli O157:H7 in Food and Water. *Int. J. Food Microbiol.* **2021**, *349*, No. 109233.

(6) Araújo, S.; Silva, I. A. T.; Tacão, M.; Patinha, C.; Alves, A.; Henriques, I. Characterization of Antibiotic Resistant and Pathogenic Escherichia Coli in Irrigation Water and Vegetables in Household Farms. *Int. J. Food Microbiol.* **2017**, *257*, 192−200.

(7) Summerlin, H. N.; Pola, C. C.; McLamore, E. S.; Gentry, T.; Karthikeyan, R.; Gomes, C. L. Prevalence of Escherichia Coli and Antibiotic-Resistant Bacteria During Fresh Produce Production (Romaine Lettuce) Using Municipal Wastewater Effluents. *Front. Microbiol.* **2021**, *12*, No. 660047.

(8) Stoeckel, D.; Ph, D.; Clements, D.; Fisk, C.; Ph, D.; Wall, G.; Bihn, B.; Ph, D. FSMA Produce Safety Rule Water Requirements: Insights to Get You Organized !, 2019, No May 1−4.

(9) Sidhu, R. K.; Cavallaro, N. D.; Pola, C. C.; Danyluk, M. D.; McLamore, E. S.; Gomes, C. L. Planar Interdigitated Aptasensor for Flow-Through Detection of Listeria Spp. in Hydroponic Lettuce Growth Media. *Sensors* **2020**, *20* (20), 5773.

(10) Vanegas, D. C.; Gomes, C. L.; Cavallaro, N. D.; Giraldo-Escobar, D.; McLamore, E. S. Emerging biorecognition and Transduction Schemes for Rapid Detection of Pathogenic Bacteria in Food. *Compr. Rev. Food Sci. Food Saf.* **2017**, *16* (6), 1188−1205.

(11) McLamore, E. S.; Alocilja, E.; Gomes, C.; Gunasekaran, S.; Jenkins, D.; Datta, S. P. A.; Li, Y.; Mao, Y. Jessie; Nugen, S. R.; Reyes-De-Corcuera, J. I.; Takhistov, P.; Tsyusko, O.; Cochran, J. P.; Tzeng, T. R. Jeremy; Yoon, J. Y.; Yu, C.; Zhou, A. FEAST of Biosensors: Food, Environmental and Agricultural Sensing Technologies

(FEAST) in North America. *Biosens. Bioelectron.* **2021**, *178*, No. 113011.

(12) Soares, R. R. A.; Hjort, R. G.; Pola, C. C.; Parate, K.; Reis, E. L.; Soares, N. F. F.; Mclamore, E. S.; Claussen, J. C.; Gomes, C. L. Laser-Induced Graphene Electrochemical Immunosensors for Rapid and Label-Free Monitoring of Salmonella enterica in Chicken Broth. *ACS Sens.* **2020**, *5* (7), 1900−1911.

(13) Giacobassi, C. A.; Oliveira, D. A.; Pola, C. C.; Xiang, D.; Tang, Y.; Datta, S. P. A.; McLamore, E. S.; Gomes, C. L. Sense−Analyze−Respond−Actuate (SARA) Paradigm: Proof of Concept System Spanning Nanoscale and Macroscale Actuation for Detection of Escherichia Coli in Aqueous Media. *Actuators* **2021**, *10* (1), 2.

(14) Randviir, E. P.; Banks, C. E. A Review of Electrochemical Impedance Spectroscopy for Bioanalytical Sensors. *Anal. Methods* **2022**, *14* (45), 4602−4624.

(15) Cai, W.; Lesnik, K. L.; Wade, M. J.; Heidrich, E. S.; Wang, Y.; Liu, H. Incorporating Microbial Community Data with Machine Learning Techniques to Predict Feed Substrates in Microbial Fuel Cells. *Biosens. Bioelectron.* **2019**, *133*, 64−71.

(16) Vashistha, R.; Dangi, A. K.; Kumar, A.; Chhabra, D.; Shukla, P. Futuristic Biosensors for Cardiac Health Care: An Artificial Intelligence Approach. *3 Biotech* **2018**, *8* (8), 358.

(17) Gonzalez-Navarro, F. F.; Stilianova-Stoytcheva, M.; Renteria-Gutierrez, L.; Belanche-Muñoz, L. A.; Flores-Rios, B. L.; Ibarra-Esquer, J. E. Glucose Oxidase Biosensor Modeling and Predictors Optimization by Machine Learning Methods. *Sensors* **2016**, *16* (11), 1483.

(18) Cui, F.; Yue, Y.; Zhang, Y.; Zhang, Z.; Zhou, H. S. Advancing Biosensors with Machine Learning. *ACS Sens.* **2020**, *5* (11), 3346−3364.

(19) Castillo-Torres, K. Y.; McLamore, E. S.; Arnold, D. P. A High-Throughput Microfluidic Magnetic Separation (MFMS) Platform for Water Quality Monitoring. *Micromachines* **2020**, *11* (1), 16.

(20) Duerr, I.; Merrill, H. R.; Wang, C.; Bai, R.; Boyer, M.; Dukes, M. D.; Bliznyuk, N. Forecasting Urban Household Water Demand with Statistical and Machine Learning Methods Using Large Space-Time Data: A Comparative Study. *Environ. Model. Software* **2018**, *102*, 29−38.

(21) Colantonio, V.; Ferrão, L. F. V.; Tieman, D. M.; Bliznyuk, N.; Sims, C.; Klee, H. J.; Munoz, P.; Resende, M. F. R. Metabolomic Selection for Enhanced Fruit Flavor. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (7), No. e2115865119.

(22) Aulenbach, B. T. Bacteria Holding Times for Fecal Coliform by MFC Agar Method and Total Coliform and Escherichia Coli by Colilert®-18 Quanti-Tray® Method. *Environ. Monit. Assess* **2010**, *161* (1−4), 147−159.

(23) Castillo-Torres, K. Y.; Arnold, D. P.; McLamore, E. S. Rapid Isolation of Escherichia Coli from Water Samples Using Magnetic Microdiscs. *Sens. Actuators, B* **2019**, *291*, 58−66.

(24) Burrs, S. L.; Bhargava, M.; Sidhu, R.; Kiernan-Lewis, J.; Gomes, C.; Claussen, J. C.; McLamore, E. S. A Paper Based Graphene-Nanocauliflower Hybrid Composite for Point of Care Biosensing. *Biosens. Bioelectron.* **2016**, *85*, 479−487.

(25) Jenkins, D. M.; Lee, B. E.; Jun, S.; Reyes-De-Corcuera, J.; McLamore, E. S. ABE-Stat, a Fully Open-Source and Versatile Wireless Potentiostat Project Including Electrochemical Impedance Spectroscopy. *J. Electrochem. Soc.* **2019**, *166* (9), B3056−B3065.

(26) Johnson, M. L.; Frasier, S. G. [16] Nonlinear Least-Squares Analysis. In *Enzyme Structure Part J.*; Methods in Enzymology; Academic Press, 1985; Vol. *117*, pp 301−342.

(27) Merrill, H. R.; Grunwald, S.; Bliznyuk, N. Semiparametric Regression Models for Spatial Prediction and Uncertainty Quantification of Soil Attributes. *Stochastic Environ. Res. Risk Assess.* **2017**, *31* (10), 2691−2703.

(28) Wood, S. N. *Generalized Additive Models: An Introduction with R, 2nd ed.*; CRC Press, 2017.

(29) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Linear Model Selection and Regularization. In *An Introduction to Statistical Learning: with Applications in R*; James, G.; Witten, D.; Hastie, T.; Tibshirani, R., Eds.; Springer US: New York, NY, 2021; pp 225−288.

(30) Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer, 2009; Vol. *2*.

(31) Taylor-Rodriguez, D.; Womack, A.; Bliznyuk, N. Bayesian Variable Selection on Model Spaces Constrained by Heredity Conditions. *J. Comput. Graphical Stat.* **2016**, *25* (2), 515−535.

(32) Merrill, H. R.; Tang, X.; Bliznyuk, N. Spatio-Temporal Additive Regression Model Selection for Urban Water Demand. *Stochastic Environ. Res. Risk Assess.* **2019**, *33* (4−6), 1075−1087.

(33) Drucker, H.; Surges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. *Adv. Neural Inf. Process. Syst.* **1997**, *1*, 155−161.

(34) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Support Vector Machines. In *An Introduction to Statistical Learning: with Applications in R*; James, G.; Witten, D.; Hastie, T.; Tibshirani, R., Eds.; Springer US: New York, NY, 2021; pp 367−402.

(35) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Tree-Based Methods. In *An Introduction to Statistical Learning: with Applications in R*; James, G.; Witten, D.; Hastie, T.; Tibshirani, R., Eds.; Springer US: New York, NY, 2021; pp 327−365.

(36) Li, H. Dielectrophoretic Separation and Manipulation of Live and Heat-Treated Cells of Listeria on Microfabricated Devices with Interdigitated Electrodes. *Sens. Actuators, B* **2002**, *86* (2−3), 215−221.

(37) Pola, C. C.; Rangnekar, S. V.; Sheets, R.; Szydłowska, B. M.; Downing, J. R.; Parate, K. W.; Wallace, S. G.; Tsai, D.; Hersam, M. C.; Gomes, C. L.; Claussen, J. C. Aerosol-Jet-Printed Graphene Electrochemical Immunosensors for Rapid and Label-Free Detection of SARS-CoV-2 in Saliva. *2D Mater.* **2022**, *9* (3), No. 035016.

(38) Soares, R. R. A.; Hjort, R. G.; Pola, C. C.; Parate, K.; Reis, E. L.; Soares, N. F. F.; McLamore, E. S.; Claussen, J. C.; Gomes, C. L. Laser-Induced Graphene Electrochemical Immunosensors for Rapid and Label-Free Monitoring of *Salmonella enterica* in Chicken Broth. *ACS Sens.* **2020**, *5* (7), 1900−1911.

(39) Moreira, G.; Casso-Hartmann, L.; Datta, S. P. A.; Dean, D.; McLamore, E.; Vanegas, D. Development of a Biosensor Based on Angiotensin-Converting Enzyme II for Severe Acute Respiratory Syndrome Coronavirus 2 Detection in Human Saliva. *Front. Sens.* **2022**, *3*, No. 917380.

(40) Abdullah, F.; Wu, Z.; Christofides, P. D. Handling Noisy Data in Sparse Model Identification Using Subsampling and Co-Teaching. *Comput. Chem. Eng.* **2022**, *157*, No. 107628.

(41) Gupta, S.; Gupta, A. Dealing with Noise Problem in Machine Learning Data-Sets: A Systematic Review. *Procedia Comput. Sci.* **2019**, *161*, 466−474.

https://doi.org/10.1021/acsomega.3c05797