frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 22 June 2022
doi: 10.3389/frai.2022.863261

Spectroscopy Approaches for Food Safety Applications: Improving Data Efficiency Using Active Learning and Semi-supervised Learning

Huanle Zhang[1]*, *Nicharee Wisuthiphaet*[2], *Hemiao Cui*[2], *Nitin Nitin*[2], *Xin Liu*[1] and *Qing Zhao*[3]

[1] Department of Computer Science, University of California, Davis, Davis, CA, United States, [2] Department of Food Science and Technology, University of California, Davis, Davis, CA, United States, [3] School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, United States

The past decade witnessed rapid development in the measurement and monitoring technologies for food science. Among these technologies, spectroscopy has been widely used for the analysis of food quality, safety, and nutritional properties. Due to the complexity of food systems and the lack of comprehensive predictive models, rapid and simple measurements to predict complex properties in food systems are largely missing. Machine Learning (ML) has shown great potential to improve the classification and prediction of these properties. However, the barriers to collecting large datasets for ML applications still persists. In this paper, we explore different approaches of data annotation and model training to improve data efficiency for ML applications. Specifically, we leverage Active Learning (AL) and Semi-Supervised Learning (SSL) and investigate four approaches: baseline passive learning, AL, SSL, and a hybrid of AL and SSL. To evaluate these approaches, we collect two spectroscopy datasets: predicting plasma dosage and detecting foodborne pathogen. Our experimental results show that, compared to the *de facto* passive learning approach, advanced approaches (AL, SSL, and the hybrid) can greatly reduce the number of labeled samples, with some cases decreasing the number of labeled samples by more than half.

Keywords: food science, spectroscopy analysis, machine learning, data efficiency, active learning, semi-supervised learning

OPEN ACCESS

Edited by:
Xiangliang Zhang,
University of Notre Dame,
United States

Reviewed by:
Dan Li,
National University of Singapore,
Singapore
Denis Helic,
Graz University of Technology, Austria

***Correspondence:**
Huanle Zhang
dtczhang@ucdavis.edu

Specialty section:
This article was submitted to AI in Food, Agriculture and Water, a section of the journal Frontiers in Artificial Intelligence

Received: *27 January 2022*
Accepted: *30 May 2022*
Published: *22 June 2022*

Citation:
Zhang H, Wisuthiphaet N, Cui H, Nitin N, Liu X and Zhao Q (2022) Spectroscopy Approaches for Food Safety Applications: Improving Data Efficiency Using Active Learning and Semi-supervised Learning. Front. Artif. Intell. 5:863261. doi: 10.3389/frai.2022.863261

1. INTRODUCTION

Rapid measurement and monitoring technologies are being developed for diverse applications in food science. The goal of these technologies is to develop predictive relationships that can be used to better monitor and enhance the quality, safety and nutritional properties of food. Among these measurement approaches, spectroscopic analysis has been widely used to analyze food properties. To fully realize the great potential of these technologies, several key barriers need to be overcome before their transfer to industrial applications. The discovery of predictive relationships between the measurements and properties of food systems is one of the key limitations. This limitation results from the complexity of food systems and the lack of comprehensive predictive models that can use rapid and simple measurements to predict complex properties in food systems.

ML has shown significant potential to improve the classification and prediction of these properties. However, the barriers to collecting large datasets for ML applications persists.

With the advances in computational capabilities and big data technologies, ML has been applied to a variety of agriculture and food-related fields (Liakos et al., 2018; Tsakanikas et al., 2020; Khullar and Singh, 2021). A prevailing approach is to train an ML model using labeled samples. While unlabeled samples can often be collected at relatively low cost, annotating each sample to create its label is expensive and time-consuming, as it often involves human inspection and in-field experiments. For example, to predict the vineyard yield, robot carrying cameras can be used to collect a large number of unlabeled images, but practitioners still need to manually process each image in order to assign appropriate labels to the images (Ballesteros et al., 2020). The prevailing practice is to randomly select samples and label them *via* costly in-field experiments and human annotation processes, and the ML model is trained only using the labeled samples. This approach results in poor data efficiency (Chapelle et al., 2006; Settles, 2012).

1.1. AL and SSL for Data-Efficient Model Training

To improve data efficiency, we explore two advanced model training techniques: AL (Settles, 2012) and SSL (Chapelle et al., 2006). Instead of randomly selecting unlabeled samples for annotation, AL selects samples for annotation based on how informative these samples are to the currently trained ML model. In comparison, SSL exploits unlabeled samples by assigning pseudo-labels to them and trains the ML model using both labeled and pseudo-labeled samples.

In this paper, we study four approaches of data annotation and model training as illustrated in **Figure 1**. (A) In the baseline passive learning approach, samples are randomly selected from the pool of unlabeled samples for annotation, and the ML model is trained only on the labeled samples. (B) In the AL approach (details in Section 2.1), the selection of unlabeled samples is dependent on the currently trained ML model. Specifically, AL selects the unlabeled samples most useful for training the ML model. Various sampling strategies are explored, which quantify the usefulness of unlabeled samples based on different criteria. The ML model is trained only using the labeled samples. (C) In the SSL approach (details in Section 2.2), unlabeled samples are randomly selected for annotation. Instead of training the ML model using only the labeled samples, SSL assigns pseudo-labels to the unlabeled samples. The ML model is trained using both the labeled and pseudo-labeled samples. Therefore, the number of training samples in SSL equals the total number of samples. The methods of assigning pseudo-labels can be either related to the currently trained ML model (e.g., self-training) or rely only on the relationship among samples (e.g., label spreading). (D) In the hybrid approach that integrates AL and SSL, AL selects an unlabeled sample for annotation, and SSL assigns pseudo-labels to the remaining unlabeled samples in each iteration. The ML model is trained using both labeled and pseudo-labeled samples. In this hybrid approach, AL and SSL interact with each other in the following manner. The

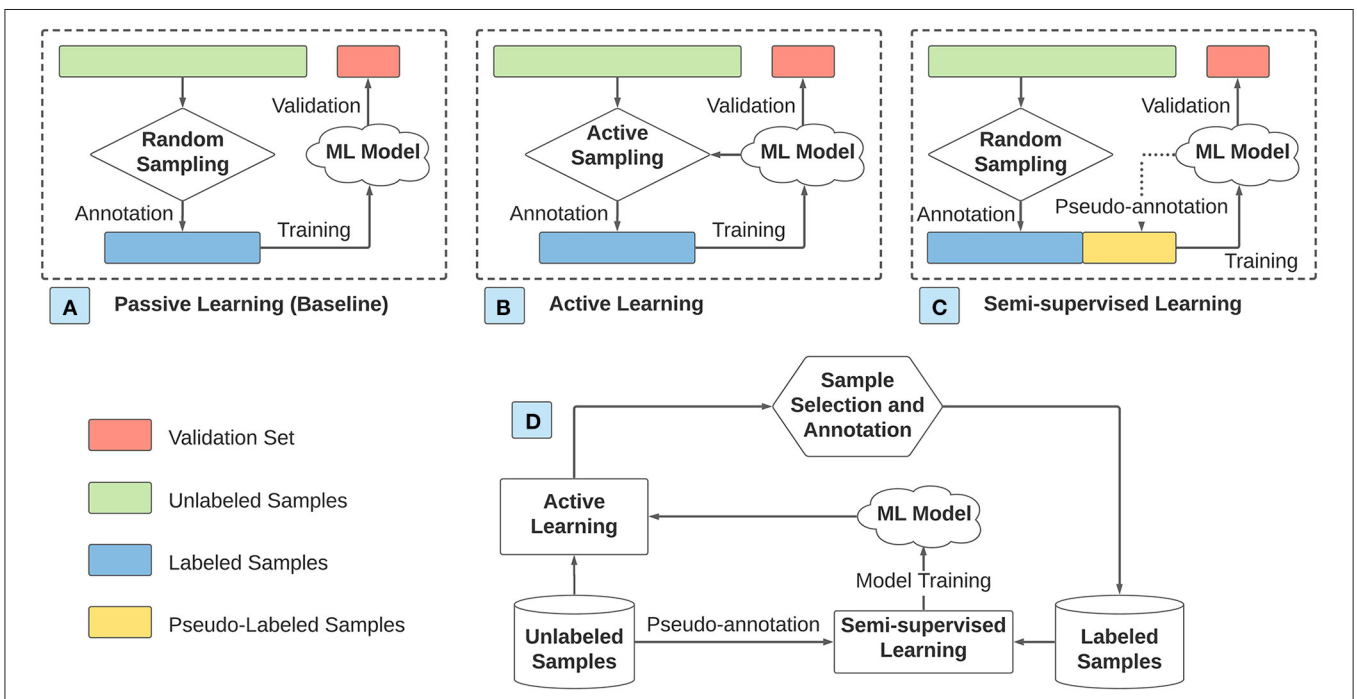

FIGURE 1 | Four approaches of data annotation and model training. **(A)** Passive learning. **(B)** Active learning. **(C)** Semi-supervised learning. **(D)** Hybrid of active learning and semi-supervised learning.

sampling strategy in AL is dependent on the ML model, which is trained using labeled samples from human annotation and pseudo-labeled samples from SSL. Meanwhile, labeled samples from AL affect SSL regarding which samples are required for pseudo-annotation.

To evaluate different approaches for spectroscopic analysis in the food science field, we collect two datasets: plasma dosage classification and foodborne pathogen detection. Atmospheric plasma technologies are being developed as a non-thermal processing technology to improve food safety, reduce the impact on food quality, and improve the sustainability of food processing operations. One of the key challenges in plasma technology applications is the rapid assessment of its efficacy in the sanitation of food contact surfaces and food products. With this motivation, we are developing infra-red spectroscopy methods to aid in assessing the dosimetry of plasma treatment. Similarly, rapid detection of foodborne pathogens is a critical unmet need in food systems. In this direction, we are developing fluorescence spectroscopic methods based on molecular interactions between bacteriophages and their target bacteria to enable specific detection of bacteria.

We apply different ML models for the multi-class classification tasks. Representative methods are considered for AL and SSL. In the experiments, we adopt five-fold cross-validation for both datasets. Our experimental results show that compared to the baseline passive learning approach, the numbers of labeled samples for the plasma dosage classification dataset and the pathogen detection dataset are reduced by more than 50% using the hybrid approach. The promising results demonstrate that AL and SSL based approaches of data annotation and model training effectively improve data efficiency for spectroscopy-based ML applications.

1.2. Related Work of Applying AL and SSL in Food Systems

Both AL and SSL have been successfully applied to many domains, such as drug discovery (Naik et al., 2013; Liu et al., 2020), material science (Lookman et al., 2019; Ma et al., 2019), and systems biology (Tamposis et al., 2019; Wang et al., 2020). Research in food systems has adopted AL (a.k.a. optimal experimental design) to optimize non-ML model parameters, with the goal of reducing the number of experiments. It shows promising results of applying AL to applications such as determining partition coefficient in freeze concentration (Munson-McGee, 2014a), tuning micro-extraction in beer (Leca et al., 2015), identifying a rice drying model (Goujot et al., 2012), and developing uniaxial expression of extracting oil from seeds (Munson-McGee, 2014b). This paper differs from existing AL works in food science as we target general ML models while previous works are designed for specific and analytical models of few parameters. Similarly, SSL have successfully been used in food science problems such as determining tomato maturity (Jiang et al., 2021) and tomato-juice freshness (Hong et al., 2015). To the best of our knowledge, this paper is the first work that extensively evaluates different AL and SSL approaches for the ML models with applications in food systems.

TABLE 1 | Notations used in this paper.

Notation	Meaning
$\mathcal{U}$	The set of unlabeled samples
$\mathcal{L}$	The set of labeled samples
θ	The currently trained ML model
θ^+	An updated ML model by adding a new training sample
C	The number of classes for classification
x	A sample
y	The label of the sample x
x^*	The sample that has the maximum utility measure
y^*	The label of the sample x^*
$P_\theta(y\|x)$	The probability that the sample x belongs to the label y under the model θ
$\hat{y}$	The most likely label for the sample x, i.e., $\hat{y} = \arg\max_y P_\theta(y\|x)$

2. MATERIALS AND METHODS

In this section, we first provide preliminaries for AL in Section 2.1 and SSL in Section 2.2. Then, we explain our two datasets in detail in Section 2.3. Last, our experimental setup is presented in Section 2.4. **Table 1** tabulates the notations that are used in this paper for quick reference.

2.1. Active Learning: A Primer

AL, also known as query learning/optimal experimental design, is a sub-field of machine learning, which studies ML models that improve with experience and training (Settles, 2012). Compared with passive learning, AL considers the "usefulness" of unlabeled samples for the current ML model. It strategically selects unlabeled samples for annotation to improve data efficiency for ML model training.

Algorithm 1 shows the workflow of the AL-based data annotation. $\mathcal{U}$ and $\mathcal{L}$ is the set of unlabeled samples and labeled samples, respectively (line 1–2). The ML model, denoted by θ, is trained on the labeled sample set $\mathcal{L}$ (line 4). The following unlabeled sample, denoted by x^*, is selected that has the highest utility measure according to the currently trained model θ (line 5). Then, the experiment is conducted for x^* to obtain its ground-truth label, denoted by y^* (line 6). Since the label for x^* is obtained, x^* is removed from the pool of unlabeled samples $\mathcal{U}$, and the sample x^* along with its label y^* is added to the set of labeled samples $\mathcal{L}$ (line 7). The process repeats until the trained model θ has a satisfactory accuracy or does not improve with more labeled samples.

The utility measure is essential for AL algorithms. There are various utility measures to estimate the usefulness of an unlabeled sample to the ML model. They mostly leverage the probability of the current ML model classifying an unlabeled sample x to class y, i.e., $P_\theta(y|x)$. We introduce two categories of utility measure methods of AL: uncertainty-based sampling (Sharma and Bilgic, 2017) and minimizing expected errors (Long et al., 2015).

2.1.1. Uncertainty Sampling Based Active Learning

The premise of uncertainty sampling is that we can avoid annotating samples that the ML model is confident about and

Algorithm 1: Selection of unlabeled samples for annotation in active learning.

1 $\mathcal{U}$: set of unlabeled samples $\{x^{(u)}\}_{u=1}^{U}$
2 $\mathcal{L}$: set of labeled samples $\{\langle x, y\rangle^{(l)}\}_{l=1}^{L}$
3 **for** $t = 1, 2, \ldots$ **do**
4 $\quad \theta = \textbf{train}(\mathcal{L})$
5 $\quad$ select $x^* \in \mathcal{U}$, the unlabeled sample of the highest *utility measure* according to the model θ
6 $\quad$ conduct experiment for x^*, which obtains its label y^*
7 $\quad$ remove x^* from $\mathcal{U}$, and add $\langle x^*, y^*\rangle$ to $\mathcal{L}$
8 **end**

focus instead on the unlabeled samples that confuse the ML model. Least confident and entropy are the most-used metrics for measuring the uncertainty of unlabeled samples.

- *Least Confident*. For an unlabeled sample x, we can apply the currently trained ML model on it, which outputs the probability of the sample belonging to class y, i.e., $P_\theta(y|x)$, where $y = 1, 2, \ldots, C$ and C is the total number of classes for classification. Let's denote by $\hat{y}$, the class that is most likely for the unlabeled sample x, i.e., $\hat{y} = \arg\max_y P_\theta(y|x)$. The confidence of the current model for the sample x is thus $P_\theta(\hat{y}|x)$. The unlabeled sample with the least confidence is selected for annotation, that is,

$$x^*_{LC} = \arg\max_x \; 1 - P_\theta(\hat{y}|x) \tag{1}$$

- *Entropy*. Entropy is an often-used metric for quantifying the uncertainty of a distribution. For an unlabeled sample x, its entropy is calculated as $-\sum_y P_\theta(y|x) \cdot \log P_\theta(y|x)$, which is over the distribution of classes y for x. The entropy-based uncertainty sampling method selects the sample with the maximum entropy, i.e.,

$$x^*_E = \arg\max_x \; -\sum_y P_\theta(y|x) \cdot \log P_\theta(y|x) \tag{2}$$

2.1.2. Minimizing Expected Error Based Active Learning

This category of AL algorithms aims to select samples to directly increase the model accuracy without relying on the assumption between the sampling strategy and the model performance (e.g., the ML model can avoid annotating confident samples as in the uncertainty sampling). Since the ground-truth label of an unlabeled sample x is not available without experiment and annotation, the model accuracy by adding the sample x and its label to the training set is unknown. Nonetheless, we can estimate the expected model accuracy when the sample x is added for training, thus selecting a sample that minimizes the expected error of the current ML model.

Assume that the unlabeled sample x belongs to the class y. Let θ^+ denote the updated ML model by adding the sample x and its imaginary label y to the training set. Then, the expected prediction error of the model θ^+ can be estimated by applying θ^+ to all unlabeled samples in $\mathcal{U}$, i.e., $\sum_{x' \in \mathcal{U}} 1 - P_{\theta^+}(\hat{y}|x')$, where $1 - P_{\theta^+}(\hat{y}|x')$ is the prediction error for the unlabeled sample x'. Since the probability that x belongs to class y is $P_\theta(y|x)$, the expected prediction error of selecting the unlabeled sample x for annotation is $\sum_y P_\theta(y|x)\left[\sum_{x' \in \mathcal{U}} 1 - p_{\theta^+}(\hat{y}|x')\right]$. Correspondingly, Equation (3) formulates the sampling strategy that minimizes the expected prediction error.

$$x^*_{EPR} = \arg\min_x \sum_y P_\theta(y|x)\left[\sum_{x' \in \mathcal{U}} 1 - p_{\theta^+}(\hat{y}|x')\right] \tag{3}$$

An alternative to minimizing the expected prediction error is to minimize the expected log-loss error. Log-loss error is the *de facto* loss function for training classification models. Therefore, minimizing the expected log-loss error has a strong connection with ML model training. By replacing the prediction error $1 - p_{\theta^+}(\hat{y}|x')$ in Equation (3) with the log-loss error $-\sum_{y'} P_{\theta^+}(y'|x') \cdot \log P_{\theta^+}(y'|x')$, Equation (4) formulates the sampling strategy of minimizing the expected log-loss error.

$$x^*_{ELR} = \arg\min_x \; -\sum_y P_\theta(y|x)\left[\sum_{x' \in \mathcal{U}} \sum_{y'} P_{\theta^+}(y'|x') \cdot \log P_{\theta^+}(y'|x')\right] \tag{4}$$

2.2. Semi-supervised Learning: A Primer

SSL can be applied to exploit unlabeled samples. Typically, an SSL algorithm converts unlabeled samples to pseudo-labeled samples and then fine-trains the ML model using both labeled and pseudo-labeled samples. SSL can be categorized into inductive methods and transductive methods (van Engelen and Hoos, 2020). We introduce self-training (Triguero et al., 2015) and label spreading (Liu et al., 2012), which are representative methods from these two categories.

2.2.1. Self-Training Based Semi-supervised Learning

In the self-training based method, the currently trained ML model is applied to pseudo-annotate the unlabeled samples (Triguero et al., 2015). Self-training methods assume that the prediction of the current model tends to be correct, and thus the model can be further fine-trained by leveraging more training samples.

Algorithm 2 shows the workflow of self-training based SSL. It differs from the AL workflow (**Algorithm 1**) in two main parts. (1) The unlabeled sample x is pseudo-labeled by the model itself, i.e., $\theta(x^*)$, in self-training, whereas it is human-annotated, i.e., the ground-truth label y^*, in AL (line 6 in **Algorithm 2** vs. line 6 in **Algorithm 1**). (2) The most confident unlabeled sample is selected in self-training. In contrast, the most uncertain unlabeled sample is chosen in AL (line 5 in **Algorithm 2** vs. line 5 in **Algorithm 1**). Self-training selects the most confident unlabeled sample to mitigate the error propagation of pseudo-annotation since the model θ is consecutively trained on the pseudo-labeled

Algorithm 2: Self-training based semi-supervised learning.

1 $\mathcal{U}$: set of unlabeled samples $\{x^{(u)}\}_{u=1}^{U}$
2 $\mathcal{L}$: set of labeled samples $\{\langle x, y\rangle^{(l)}\}_{l=1}^{L}$
3 **while** $\mathcal{U}$ *is not empty* **do**
4 $\quad\theta = \textbf{train}(\mathcal{L})$
5 $\quad$select $x^* \in \mathcal{U}$, the most confident sample according to model θ
6 $\quad$obtain the pseudo-label of x^* by applying the current model θ to it, i.e., $\theta(x^*)$
7 $\quad$remove x^* from $\mathcal{U}$ and add $\langle x^*, \theta(x^*)\rangle$ to $\mathcal{L}$
8 **end**

samples. In contrast, AL selects the model's most confusing unlabeled sample for human annotation so that the model can correctly classify misleading samples. The process of unlabeled sample selection, pseudo-annotation, and model training repeats until all the unlabeled samples are pseudo-annotated (line 3). At that time, the model has been trained on all labeled and pseudo-labeled samples.

2.2.2. Label Spreading Based Semi-supervised Learning

Label spreading based SSL typically adopts a graph structure, where the vertices are the samples (both labeled and unlabeled) and edges exist between neighbor vertices (Liu et al., 2012). The edge weight between two vertices represents the similarity of the corresponding two samples. The goal of label spreading is to assign pseudo-labels to the vertices of unlabeled samples. The labels of x_i and x_j are likely to be the same if the edge weight w_{ij} between them is large. There are two mainstream methods to define the graph structure and the edge weights:

- *Fully Connected Graph* (de Sousa et al., 2013). In this graph, all vertices are connected with other vertices, and the edge weight w_{ij} between x_i and x_j is calculated as

$$w_{ij} = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \tag{5}$$

  where $||x_i - x_j||$ is the Euclidean distance between sample x_i and sample x_j, and σ controls the decreasing rate of the weight over the distance. The edge weight w_{ij} is 1 when $x_i = x_j$, and approximately 0 when x_i is far from x_j. This weight representation is also called Radial Base Function (RBF) kernel or Gaussian kernel.
- *k-Nearest-Neighbor (kNN) Graph* (de Sousa et al., 2013). In the kNN graph, each vertex is connected to its k nearest vertices. If sample x_i and sample x_j are connected, the edge weight w_{ij} is 1; otherwise, w_{ij} is 0. The kNN graph automatically adapts to the density of samples: in dense regions, the radius of the kNN neighborhood is small, while in sparse regions, the radius is large.

The probability of the label for each unlabeled sample is obtained once the graph converges (Zhou et al., 2003). Then, each unlabeled sample is assigned to the label of the highest probability at once. Afterward, the ML model is re-trained using all samples (labeled and pseudo-labeled).

2.3. Dataset

We collect two spectroscopy datasets in food science. One dataset predicts the plasma dosage, and the other detects the foodborne pathogen. Our datasets have representative data structures (1D and 2D spectroscopy), and we use different ML models for these two datasets. Our chosen datasets are representative of food safety research and thus our experimental results are applicable to other food safety applications/datasets. We do not present the details of the data collection process, which is not the focus of this paper.

2.3.1. Dataset 1: Plasma Dosage Classification

As an emerging nonthermal processing technology, plasma is highly effective in inactivating various types of food pathogens in solutions as well as on food contact surfaces (Liao et al., 2017). One of the main goals of our plasma dosage classification dataset is to evaluate whether the plasma dosage can be predicated using the Fourier-Transform Infrared Spectroscopy (FTIR) spectral response of the DNA sample exposed to the plasma treatment. The dataset was obtained by subjecting the substrate to plasma treatment at various dosage levels and characterizing biochemical changes of the substrate with FTIR. The data comprise absorbance intensities at various IR wavenumbers of the substrate under plasma treatment.

Our 1D dosage classification dataset categorizes the plasma dosage into four classes, where classes 1, 2, 3, and 4 represent plasma injection of 0, 2, 5, and 10 min, respectively. In total, we collect 114 samples, where classes 1, 2, 3, and 4 have 27, 27, 30, 30 samples, respectively. To study the sample distribution over the feature domain, we apply K-Means (Arthur and Vassilvitskii, 2007) to group samples into 4 clusters. **Figure 2A** shows the clustering result. It indicates that the class 2 and the class 3 samples are well clustered, while many of the class 4 samples are located within cluster 2 and cluster 3 (a PCA visualization is given in **Figure 3**). **Figure 2B** illustrates a sample for each class. Each sample contains 1,868 values representing the intensity of the IR frequency reflectance in the range of 400–4,000 Hz with a step of 2 Hz.

2.3.2. Dataset 2: Foodborne Pathogen Detection

Detection of the bacterial foodborne pathogen is one of the critical processes in the food and agricultural industry (Velusamy et al., 2010). It ensures the safety of food products before distribution and reduces the risk of a foodborne illness outbreak. Among various types of bacteria, *E. coli* has been used as an indicator of the fecal contamination and poor sanitary quality of food and water (Krumperman, 1983). T7 bacteriophage or T7 phage has been used as a tool for *E. coli* detection as it infects explicitly *E. coli* cells results in bacteria cell lysis and release of the cell components along with the amplified T7 phage progenies (Yang et al., 2020). Therefore, bacterial cell lysis and T7 phage amplification can indicate *E. coli* contamination in

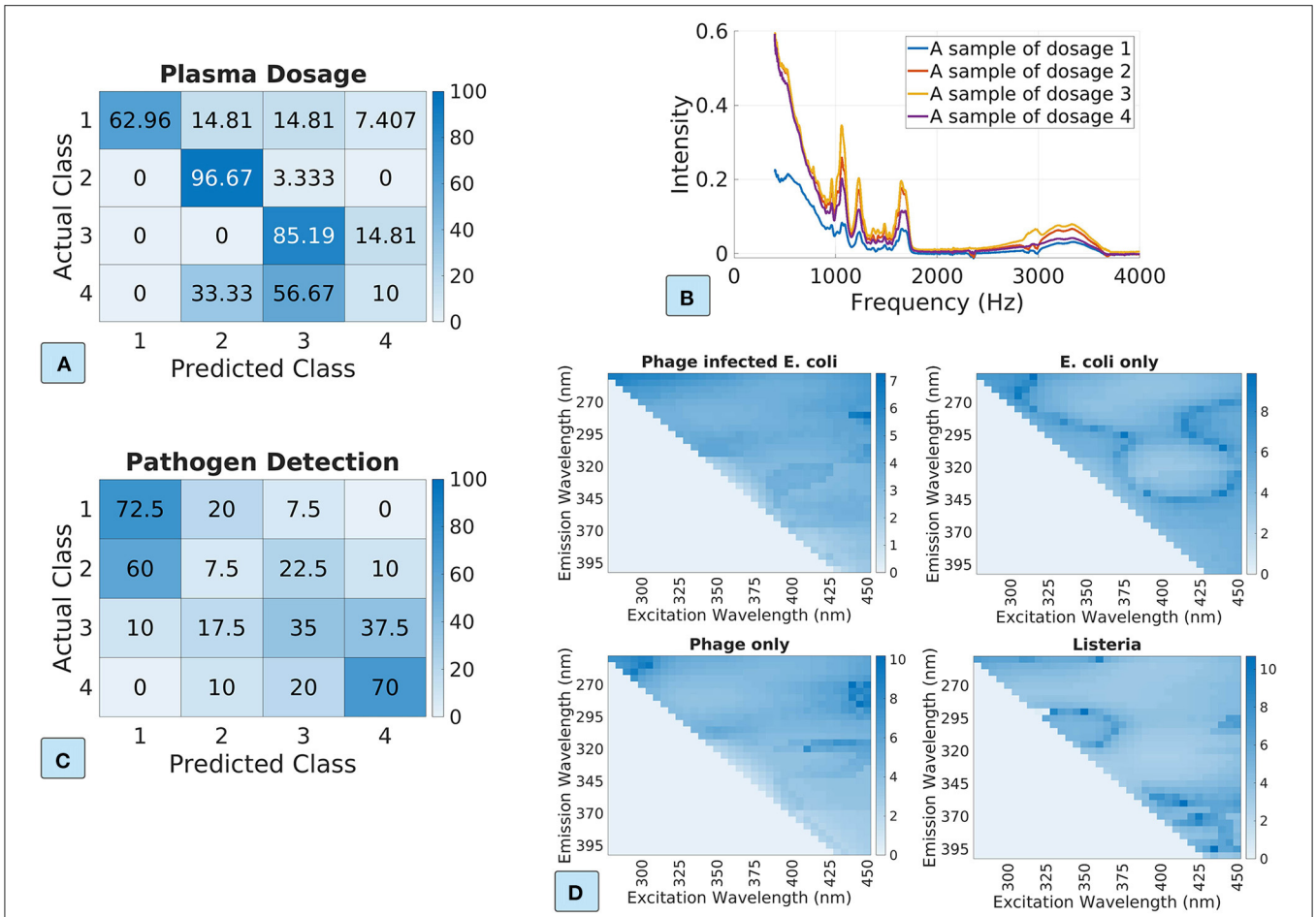

FIGURE 2 | Two datasets are used in our experiments. The top row shows the plasma dosage classification dataset where **(A)** is the clustering result and **(B)** illustrates a sample for each class. The bottom row shows the pathogen detection dataset where **(C)** is the clustering result and **(D)** depicts a sample for each class.

the samples. Fluorescence EEM spectroscopy is an analytical technique providing multi-dimensional information (Li et al., 2020) that has been used to detect organic materials, including bacterial cells (Nakar et al., 2020).

We explore whether 2D EEM spectroscopy can detect the change of the physical and chemical properties of the samples due to the phage infection of *E. coli*. We use classes 1, 2, 3, and 4 to represent phage infected *E. coli*, *E. coli* only, phage only, and *Listeria* solutions, respectively. The EEM spectra of the resuspended samples were collected with the wavelength range of 250–400 nm for excitation and 260–450 nm for emission with 5-nm increments. **Figure 2C** shows the result of grouping the samples into 4 clusters using K-Means. It indicates that most *E. coli* only and the phage only samples are wrongly categorized. **Figure 2D** visualizes one sample for each class. In the heat maps, we take the logarithm of the frequency intensity at each frequency pair of the excitation and emission wavelength. In total, we collect 160 samples, and each class has 40 samples. Each sample includes 744 numbers representing the fluorescence response for the excitation-emission pairs.

2.4. Experimental Setup

Given the same number of labeled samples, we compare the ML model accuracy when different data annotation and model training approaches are applied. We use the LightGBM multi-class classification model (Ke et al., 2017) for the plasma dosage dataset, and the logistic regression classifier (Dreiseitl and Ohno-Machado, 2002) for the pathogen detection dataset. We select these ML models because they show promising results in our related projects. We adopt five-fold cross-validation for both datasets. Specifically, at each cross-validation round, the training and validation sets comprise the four-fold samples and the remaining one-fold samples, respectively. We use the training set for annotating a given number of samples and apply the trained ML model to predict the validation set. Note that the training set is assumed to be unlabeled at the beginning of each cross-validation, and the model is trained with different approaches (refer to **Figure 1**). The predictions for all samples are obtained by aggregating the predictions of the ML model for each validation set from the five cross-validations.

To determine the hyper-parameters (e.g., learning rate, regularization) of ML models for a given number of labeled

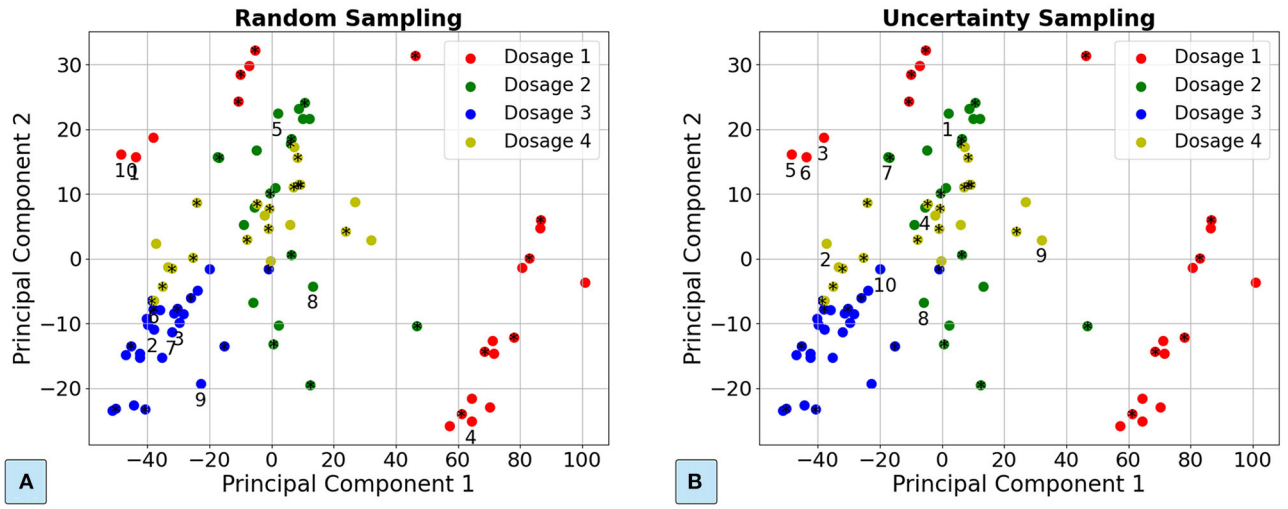

FIGURE 3 | Demonstration of the samples selected in the passive learning approach and the active learning approach for the plasma dosage classification dataset. **(A)** Passive learning approach. **(B)** Active learning approach with the entropy-based uncertainty sampling. The 40 randomly selected samples for training an initial ML model are marked with "*". The next 10 samples are tagged with the order number.

samples, we apply Optuna (Akiba et al., 2019) to the passive learning approach and adopt the same hyper-parameters for all approaches. Therefore, the hyper-parameters favor the passive learning approach in our experiments. Using Optuna, we only need to provide reasonable ranges for hyper-parameters, and Optuna can identify the optimal hyper-parameters. We find that hyper-parameters from Optuna tend to have higher model accuracy than the default hyper-parameters.

3. RESULTS

We evaluate the four approaches of data annotation and model training using our two datasets. We warm-start a LightGBM model using 40 randomly selected labeled samples for the plasma dosage dataset and then apply different approaches. Likewise, we first train a logistic regression classification model using 25 randomly selected labeled samples for the pathogen detection dataset before applying different approaches. We run 10 experiments to average the results. Since we adopt the five-fold cross-validation, the maximum number of samples for training in each cross-validation is 90 for the plasma dosage task and 127 for the pathogen detection task.

3.1. Results of the Active Learning Approach

We first demonstrate the order of samples selected for annotation in the passive learning and AL approaches. Principal Component Analysis (PCA) (Wold et al., 1987) is applied to project the high-dimensional samples into two components for visualization. **Figure 3** illustrates a trace for the passive learning approach and the AL approach for the dosage classification dataset, respectively. As **Figure 3A** illustrates, the selected samples in the passive learning can be close to each other and also near the center of the class (e.g., samples 2, 3, and 7), which are less likely to improve the ML model accuracy. In comparison, as **Figure 3B** shows, the selected samples in the AL approach are close to the boundaries of different classes. In these boundary areas, the ML model is difficult to classify. Thus, training with the samples in the boundary areas is more likely to increase the ML model accuracy.

In the rest of the experiments for the AL approach, we consider (1) uncertainty sampling methods based on least confident and maximum entropy, and (2) minimizing the expected prediction error and the expected log-loss error. **Figure 4** plots the ML model accuracy vs. the number of labeled samples. The first row and the second row show the model performance for the dosage classification and the pathogen detection datasets, respectively. We show the standard deviation in shaded colors. We can see that AL approach achieves better performance than the passive learning approach, especially for the foodborne pathogen dataset. For example, the uncertainty-based AL approach reduces the number of labeled to only 40 for the pathogen detection dataset, compared to 80 in the passive learning approach, achieving 50% label data reduction. The sampling strategies of minimizing expected errors depend on the current ML model accuracy to calculate the expected errors, which does not perform well if the current ML model has low prediction accuracy (e.g., for the plasma dosage classification dataset). Nonetheless, we do not see degradation of data efficiency.

3.2. Results of the Semi-supervised Learning Approach

SSL exploits unlabeled samples by assigning pseudo-labels to them and then trains the ML model using both labeled and pseudo-labeled samples. In the evaluation of the SSL approach,

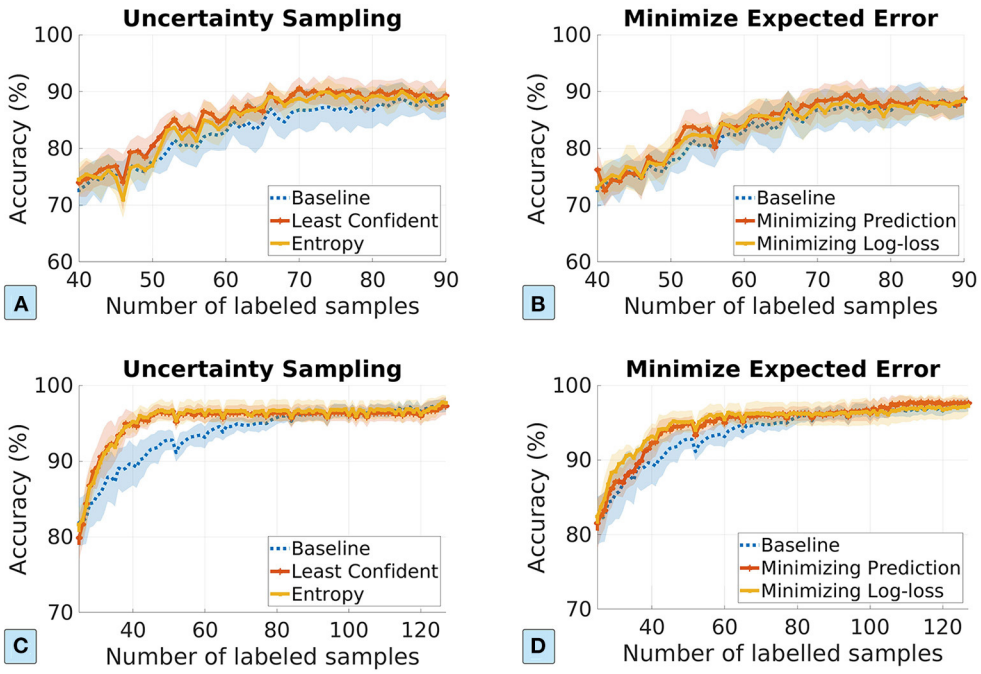

FIGURE 4 | The ML model performance vs. the number of labeled samples in the AL approach. The first row and the second row show the results for the dosage classification dataset and the pathogen detection dataset, respectively. **(A,C)** Uncertainty sampling based on least confident and entropy; **(B,D)** minimizing expected prediction error and expected log-loss error based sampling.

we consider self-training and label spreading. We set the number of neighbors in the kNN kernel to 7 and the σ in the RBF kernel to 0.1 for both datasets.

Figure 5 plots the ML model accuracy vs. the number of labeled samples in the SSL approach. The first row and the second row show the model performance for the plasma and pathogen datasets. We visualize the standard deviation in shaded colors. We can see that the SSL approach can improve data efficiency in most experiment scenarios. The exception is the label-spreading for the pathogen detection dataset, which shows much-degraded performance. It is probably because the pathogen detection dataset is poorly clustered (see **Figure 2C**), and thus the pseudo-labels from label-spreading are mostly incorrect. In the cases of label spreading for the plasma dataset and self-training for the pathogen dataset, SSL approach improves the model accuracy by about 10% when the label samples are few.

3.3. Results of the Hybrid Approach

We measure the performance of the hybrid approach using two combinations: random sampling based self-training and RBF-based self-labeling for SSL, both with entropy-based uncertainty sampling for AL. The σ value in the label spreading is set to 0.01 for both datasets.

Figure 6 shows the ML model accuracy for the hybrid approach, where the top row is for the plasma dosage dataset and the bottom row is for the pathogen detection dataset. As we can see, the hybrid approach dramatically improves the data efficiency for ML model training. For example, **Figure 6B** shows that the hybrid approach reaches the maximum model accuracy with only 45 labeled samples, in comparison to 85 in the baseline, and thus reduces the amount of labeled data by about 50%. The pathogen detection dataset performs even better with the hybrid approach: 50 labeled samples achieve the same maximum model accuracy as 120 labeled samples in the baseline, reducing the number of labeled samples by about 60%. Even though the self-labeling based SSL works poorly for the pathogen dataset (**Figure 5D**), the hybrid approach still works better than the baseline: 70 labeled samples vs. 90 labeled samples. Overall, the hybrid approach is promising in improving data efficiency.

4. DISCUSSION

We explore different approaches of data annotation and model training to improve data efficiency for ML applications. Our datasets are general in food safety research. Therefore, our approaches can be applied to general food safety research. In this section, we discuss practical considerations in applying these advanced approaches.

4.1. Practicability of Advanced Approaches

The passive learning approach has been adopted for the majority of projects. An essential question is whether more advanced approaches can be effectively and efficiently applied for new projects, considering the overheads of implementing these approaches. We argue that when the annotation process is costly, the benefits of the reduced number of labeled samples overweigh the implementation overheads. In fact, AL and SSL have been successfully applied to many domains (Lookman et al., 2019; Ma

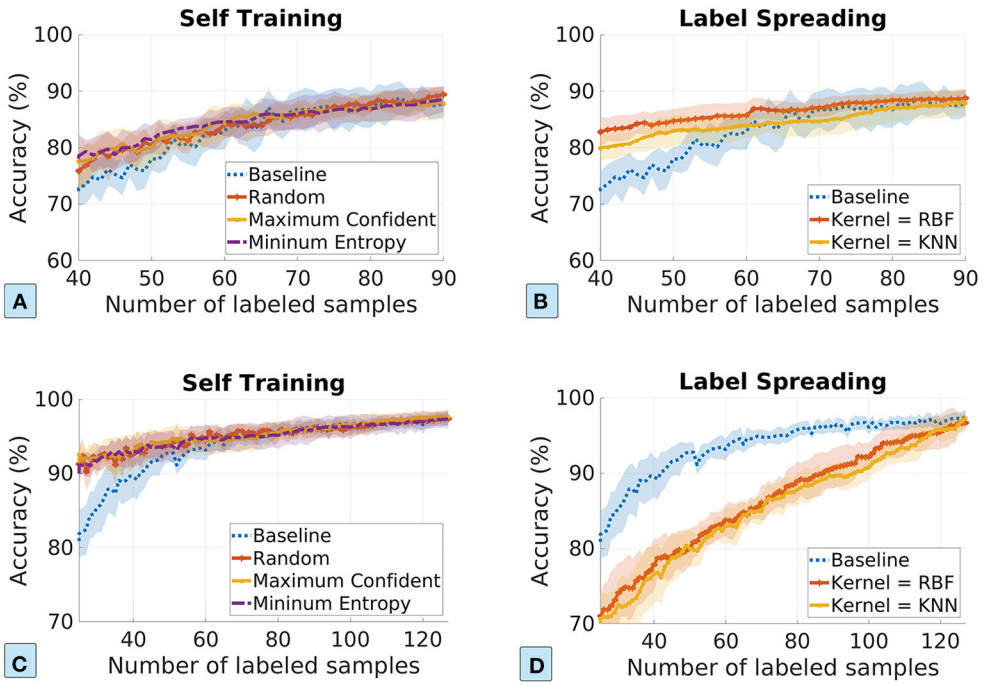

FIGURE 5 | The ML model accuracy vs. the number of labeled samples in the semi-supervised learning approach. The first row and the second row show the results for the dosage classification dataset and the pathogen detection dataset, respectively. **(A,C)** Self-training based on maximum confident and minimum entropy; **(B,D)** label spreading based on kNN kernel and RBF kernel.

et al., 2019; Tamposis et al., 2019; Liu et al., 2020; Wang et al., 2020). For example, the medicine industry applies AL to discover antagonists for dopamine D_4 (Reutlinger et al., 2014) and CXC-chemokine (Reker et al., 2016), among the estimated range of 10^{30}-10^{60} drug-like molecules (Naik et al., 2013). In food systems, development of AL and SSL methods can help address challenges of obtaining labeled samples for ML models as labeling in many food safety applications is time-consuming and labor-intensive.

4.2. Determining the Optimal Approach

We introduce three advanced data annotation and model training approaches and compare their performance with the passive learning approach. Our evaluation results show that the hybrid approach can dramatically improve data efficiency. Therefore, we suggest the hybrid approach that leverages both AL and SSL. However, the sampling strategy in AL and the pseudo-labeling method in SSL are crucial for the hybrid approach's performance.

4.3. Determining the Optimal Sampling Strategy in Active Learning

There are a great variety of sampling strategies available in AL (Settles, 2012). Although it is not possible to obtain a universally good AL sampling strategy (Dasgupta, 2004), we have demonstrated that by using simple sampling strategies such as uncertainty-based sampling, we achieve better data efficiency than the passive learning approach. Simple sampling strategies based on uncertainty and disagreement are recommended if the domain knowledge about the samples and the problems are not available (Settles, 2012). On the other hand, incorporating domain knowledge into AL can further improve data efficiency. For example, Liang et al. (2020) proposes an expert-in-the-loop AL framework that utilizes language explanations from domain experts to iteratively distinguish misleading breeds of birds, which outperform baseline models that are trained with 40–100% more training samples. Automatic selection of sampling strategies has also been extensively studied, (e.g., Hsu and Lin, 2015; Konyushkova et al., 2017).

In our experiments, we warm-start the initial ML model by randomly selecting and training 40 samples for the dosage classification dataset and 25 samples for the pathogen detection dataset. Another common practice is to switch between random sampling (for exploration) and AL sampling strategy (for exploitation). For example, Wang et al. (2020) train a predictive ML model of gene expression with 44% fewer data by consecutive switching between random sampling and mutual information based AL sampling strategy.

4.4. Determining the Optimal Pseudo-Labeling Method in Semi-supervised Learning

In addition to the self-training and the label-spreading methods, semi-supervised learning includes many other methods such as co-training, boosting, and perturbation-based (van Engelen and

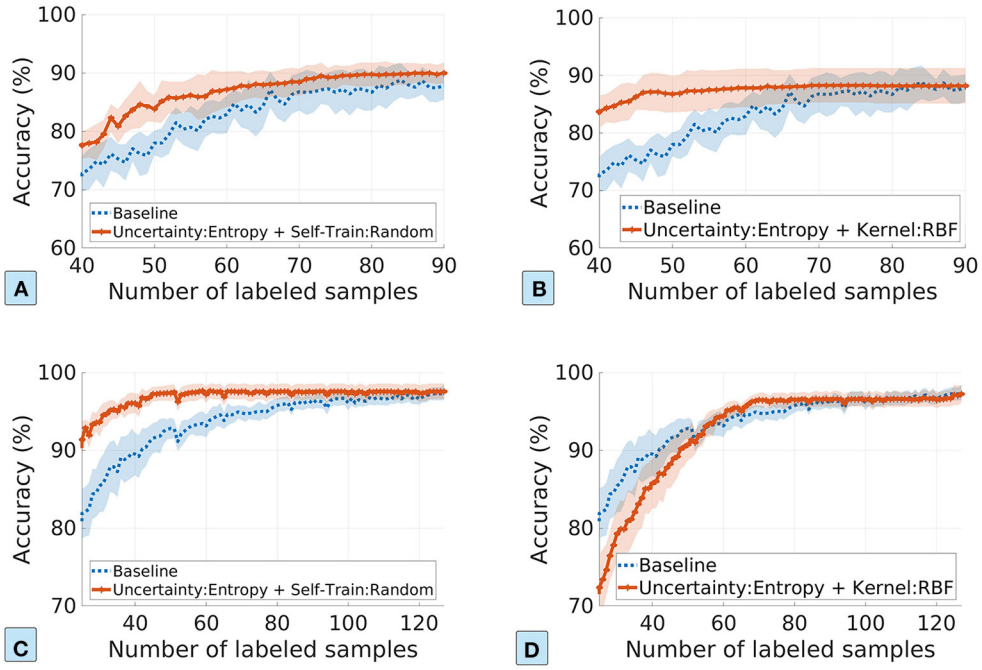

FIGURE 6 | The ML model accuracy vs. the number of labeled samples in the hybrid approach. The top row and the bottom row show the results for the plasma dosage dataset and the foodborne pathogen detection dataset. **(A,C)** Uncertainty-based AL with self-training based SSL. **(B,D)** Uncertainty-based AL with self-labeling based SSL.

Hoos, 2020). Similar to AL, SSL does not guarantee better data efficiency than the passive learning approach (Li and Zhou, 2011). Nonetheless, our experiments demonstrate that simple pseudo-labeling methods such as self-training can greatly improve the model accuracy when the underlying samples can be well-clustered and the number of labeled samples are small. Recent advances in SSL show promising results of the perturbation-based semi-supervised neural networks, which empirically and consistently outperforms the passive learning approach (van Engelen and Hoos, 2020). We leave it as future work to evaluate the perturbation-based methods.

4.5. Complexity Analysis

The computation capability is also a factor in deciding the approach of data annotation and model training. We formulate the computation complexity as the number of ML model training and ML model inference, as they are more computation demanding than other operations (e.g., graph construction in label-spreading). We denote the total number of samples by T, the number of the initial labeled samples by B, and the number of final labeled samples by F. **Table 2** tabulates the overheads in different approaches, whereas we ignore the model validation overheads as they are the same for all approaches. If the model is heavy and slow to execute, then the computation-intensive methods, e.g., minimizing expected errors, should be avoided.

4.6. Extension to Regression Problems

In this paper, we use classification to illustrate different approaches of data annotation and model training. We expect our main conclusion to remain valid for regression problems; that is, advanced approaches can help improve data efficiency for model training compared to the passive learning approach. There are many existing works solving regression problems in AL and/or SSL (Georgios et al., 2018; Wu et al., 2019), which we will explore in our future work.

4.7. Applicability for Diverse Applications in Food Quality, Traceability, and Safety

The spectroscopy measurement approaches selected for this study represent two distinct spectral methods, namely IR spectroscopy and fluorescence spectroscopy. IR spectroscopy has been proposed for diverse applications in food quality (van de Voort, 1992), traceability (Hennessy et al., 2009) and food safety (Bagcioglu et al., 2019). These diversity of applications are enabled by the unique ability of IR spectroscopy to detect molecular composition of food materials using non-destructive sampling and rapid data collection. ML approaches can complement the current chemometric methods used for the analysis of IR spectroscopy data. Complementary to IR spectroscopy, fluorescence spectroscopy approaches rely on the photo-active properties of fluorophores in food systems and their relationship with changes in food quality and traceability (Hassoun et al., 2019). Similarly, fluorescence properties have also been used to monitor the presence of bacteria in water samples (Cumberland et al., 2012). However, due to significant interference between the food materials and bacterial components there has been limited

TABLE 2 | Computation overheads of different approaches of data annotation and model training.

Approach	Method	No of model training	No of model inference
Passive	-	$F-B$	0
AL	Uncertainty-based	$F-B$	$\frac{(F-B)\times(2T-F-B+1)}{2}$
	Minimize expected errors	$\frac{C\times(F-B)\times(2T-F-B+1)}{2}$	$\approx \frac{3C\times(F-B)^2\times(T-B)^2+(F-B)^3}{3}$
SSL	Self-training	$\frac{2\times(F-B)+(T-B)\times(T-B+1)}{2}$	$\frac{(T-B)\times(T-B+1)}{2}$
	Label-spreading	$F-B$	0
Hybrid	Uncertainty + Self-training	$\frac{2\times(F-B)+(T-B)\times(T-B+1)}{2}$	$\frac{(F-B)\times(2T-F-B+1)+(T-B)\times(T-B+1)}{2}$
	Uncertainty + Label-spreading	$F-B$	$\frac{(F-B)\times(2T-F-B+1)}{2}$

C is the number of classes for classification; T is the total number of samples, B is the initial labeled samples, and F is the number of final labeled labels.

applications of fluorescence spectroscopy in food safety. To address some of these limitations, this study evaluated the applications of Fluorescence EEM spectroscopy for food safety. EEM is a technique that allows for the complete, quantitative determination of the fluorescence profile of a given material and has been used for biomedical and material characterization applications (Ramsay et al., 2018). Application of EEM spectroscopy including their integration with ML methods can address some of the key challenges in application of fluorescence spectroscopy for food applications. In addition, our approaches of active learning, semi-supervised learning, and hybrid are general, and thus are expected to be applicable to other research domains.

5. CONCLUSION

In this paper, we target data efficiency of ML applications for spectroscopy analysis in food science. To mitigate the annotation cost by reducing the number of labeled samples, we explore different approaches of data annotation and model training: passive learning, active learning, semi-supervised learning, and the hybrid of active learning and semi-supervised learning. We evaluate these approaches in two spectroscopy datasets and find that advanced approaches can greatly improve data efficiency for ML model training. These approaches are general and thus can be applied to various ML-based food science research.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, upon request to NN.

AUTHOR CONTRIBUTIONS

NN and XL conceptualized and supervised the project. QZ supervised the project. HZ conceptualized and implemented the project. NW collected the pathogen detection dataset and conducted a preliminary machine learning model evaluation. HC collected the plasma dosage classification dataset and conducted a preliminary machine learning model evaluation. All authors contributed to the revision of the manuscript and approved the final submitted version.

FUNDING

The work was partially supported by NSF through grants (USDA-020-67021-32855 and OIA-2134901) and USDA National Institute of Food and Agriculture grant 2015-68003-23411.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). "Optuna: a next-generation hyperparameter optimization framework," in *International Conference on Knowledge Discovery and Data Mining (KDD)* (Anchorage, AK), 2623–2631. doi: 10.1145/3292500.3330701

Arthur, D., and Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding," in *ACM-SIAM Symposium on Discrete algorithms (SODA)* (New Orleans, LA), 1027–1035.

Bagcioglu, M., Fricker, M., Johler, S., and Ehling-Schulz, M. (2019). Detection and identification of *Bacillus cereus, Bacillus cytotoxicus* and *Bacillus thuringiensis* and *Bacillus mycoides* and *Bacillus weihenstephanensis* via machine learning based FTIR spectroscopy. *Front. Microbiol.* 10, 902. doi: 10.3389/fmicb.2019.00902

Ballesteros, R., Intrigliolo, D. S., Ortega, J. F., Ramírez-Cuesta, J. M., Buesa, I., and Moreno, M. A. (2020). Vineyard yield estimation by combining remote sensing, computer vision and artificial neural network techniques. *Precis. Agric.* 21, 1242–1262. doi: 10.1007/s11119-020-09717-3

Chapelle, O., Scholkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/9780262033589.001.0001

Cumberland, S., Bridgeman, J., Baker, A., Sterling, M., and Ward, D. (2012). Fluorescence spectroscopy as a tool for determining microbial quality in potable water applications. *Environ. Technol.* 33, 687–693. doi: 10.1080/09593330.2011.588401

Dasgupta, S. (2004). "Analysis of a greedy active learning strategy," in *International Conference on Neural Information Processing Systems (NIPS)* (Vancouver, BC), 1–8.

de Sousa, C. A. R., Rezende, S. O., and Batista, G. E. A. P. A. (2013). "Influence of graph construction on semi-supervised learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)* (Prague), 160–175. doi: 10.1007/978-3-642-40994-3_11

Dreiseitl, S., and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* 35, 352–359. doi: 10.1016/S1532-0464(03)00034-0

Georgios, K., Stamatis, K., Sotiris, K., and Omiros, R. (2018). Semi-supervised regression: a recent review. *J. Intell. Fuzzy Syst.* 35, 1483–1500. doi: 10.3233/JIFS-169689

Goujot, D., Meyer, X., and Courtois, F. (2012). Identification of a rice drying model with an improved sequential optimal design of experiments. *J. Process Control* 22, 95–107. doi: 10.1016/j.jprocont.2011.10.003

Hassoun, A., Sahar, A., Lakhal, L., and Ait-Kaddour, A. (2019). Fluorescence spectroscopy as a rapid and non-destructive method for monitoring quality and authenticity of fish and meat products: impact of different preservation conditions. *LWT Food Sci. Technol.* 103, 279–292. doi: 10.1016/j.lwt.2019.01.021

Hennessy, S., Downey, G., and ODonnell, C. P. (2009). confirmation of food origin claims by Fourier transform infrared spectroscopy and chemometrics: extra virgin olive from Liguria. *J. Agric. Food Chem.* 57, 1735–1741. doi: 10.1021/jf803714g

Hong, X., Wang, J., and Qi, G. (2015). E-nose combined with chemometrics to trace tomato-juice quality. *J. Food Eng.* 149, 38–43. doi: 10.1016/j.jfoodeng.2014.10.003

Hsu, W.-N., and Lin, H.-T. (2015). "Active learning by learning," in *Association for the Advancement of Artificial Intelligence (AAAI)* (Austin, TX), 2659–2665.

Jiang, Y., Chen, S., Bian, B., Li, Y., Sun, Y., and Wang, X. (2021). Discrimination of tomato maturity using hyperspectral imaging combined with graph-based semi-supervised method considering class probability information. *Food Anal. Methods* 14, 968–983. doi: 10.1007/s12161-020-01955-5

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *International Conference on Neural Information Processing Systems (NIPS)* (Long Beach, CA), 3149–3157.

Khullar, S., and Singh, N. (2021). Machine learning techniques in river water quality modelling: a research travelogue. *Water Supply* 21, 1–13. doi: 10.2166/ws.2020.277

Konyushkova, K., Raphael, S., and Fua, P. (2017). "Learning active learning from data," in *Conference on Neural Information Processing Systems (NIPS)* (Long Beach, CA), 1–11.

Krumperman, P. H. (1983). Multiple antibiotic resistance indexing of *Escherichia coli* to identify high-risk sources of fecal contamination of food. *Appl. Environ. Microbiol.* 46, 165–170. doi: 10.1128/aem.46.1.165-170.1983

Leca, J. M., Pereira, A. C., Vieira, A. C., Reis, M. S., and Marques, J. C. (2015). Optimal design of experiments applied to headspace solid phase microextraction for the quantification of vicinal diketones in beer through gas chromatography-mass spectrometric detection. *Anal. Chim. Acta* 887, 101–110. doi: 10.1016/j.aca.2015.06.044

Li, L., Wang, Y., Zhang, W., Yu, S., Wang, X., and Gao, N. (2020). New advances in fluorescence excitation-emission matrix spectroscopy for the characterization of dissolved organic matter in drinking water treatment: a review. *Chem. Eng. J.* 381, 1–12. doi: 10.1016/j.cej.2019.122676

Li, Y.-F., and Zhou, Z.-H. (2011). "Towards making unlabeled data never hurt," in *International Conference on Machine Learning (ICML)* (Bellevue, WA), 1081–1088.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: a review. *Sensors* 18, 1–29. doi: 10.3390/s18082674

Liang, W., Zou, J., and Yu, Z. (2020). "ALICE: active learning with contrastive natural language explanations," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4380–4391. doi: 10.18653/v1/2020.emnlp-main.355

Liao, X., Liu, D., Xiang, Q., Ahn, J., Chen, S., Ye, X., et al. (2017). Inactivation mechanisms of non-thermal plasma on microbes: a review. *Food Control* 75, 83–91. doi: 10.1016/j.foodcont.2016.12.021

Liu, N., Chen, C.-B., and Kumara, S. (2020). Semi-supervised learning algorithm for identifying high-priority drug-drug interactions through adverse event reports. *IEEE J. Biomed. Health Inform.* 24, 57–68. doi: 10.1109/JBHI.2019.2932740

Liu, W., Wang, J., and Chang, S.-F. (2012). Robust and scalable graph-based semisupervised learning. *Proc. IEEE* 100, 2624–2638. doi: 10.1109/JPROC.2012.2197809

Long, B., Bian, J., Chapelle, O., Ya Zhang, Y. I., and Chang, Y. (2015). Active learning for ranking through expected loss optimization. *IEEE Trans. Knowledge Data Eng.* 27, 1180–1191. doi: 10.1109/TKDE.2014.2365785

Lookman, T., Balachandran, P. V., Xue, D., and Yuan, R. (2019). Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *Comput. Mater.* 5, 1–17. doi: 10.1038/s41524-019-0153-8

Ma, W., Cheng, F., Xu, Y., Wen, Q., and Liu, Y. (2019). Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy. *Adv. Mater.* 31, 1–9. doi: 10.1002/adma.201901111

Munson-McGee, S. H. (2014a). D- and G-optimal experimental designs for the partition coefficient in freeze concentration. *J. Food Eng.* 121, 80–86. doi: 10.1016/j.jfoodeng.2013.08.018

Munson-McGee, S. H. (2014b). D-optimal experimental designs for uniaxial expression. *J. Food Process Eng.* 37, 248–256. doi: 10.1111/jfpe.12080

Naik, A. W., Kangas, J. D., Langmead, C. J., and Murphy, R. F. (2013). Efficient modeling and active learning discovery of biological responses. *PLoS ONE* 8, e83996. doi: 10.1371/journal.pone.0083996

Nakar, A., Schmilovitch, Z., Vaizel-Ohayon, D., Kroupitski, Y., Borisover, M., and Saldinger, S. S. (2020). Quantification of bacteria in water using PLS analysis of emission spectra of fluorescence and excitation-emission matrices. *Water Res.* 169, 1–9. doi: 10.1016/j.watres.2019.115197

Ramsay, H., Simon, D. E., Steele, Hebert, A., Oleschuk, R. D., and Stamplecoskie, K. G. (2018). The power of fluorescence excitation-emission matrix (EEM) spectroscopy in the identification and characterization of complex mixtures of fluorescent silver clusters. *RSC Adv.* 8, 42080–42086. doi: 10.1039/C8RA08751B

Reker, D., Schneider, P., and Schneider, G. (2016). Multi-objective active machine learning rapidly improves structure-activity models and reveals new protein-protein interaction inhibitors. *Chem. Sci.* 7, 3919–3927. doi: 10.1039/C5SC04272K

Reutlinger, M., Rodrigues, T., Schneider, P., and Schneider, G. (2014). Multi-objective molecular *de novo* design by adaptive fragment prioritization. *Angew. Int. Ed. Chem.* 53, 4244–4248. doi: 10.1002/anie.201310864

Settles, B. (2012). *Active Learning*. San Rafael, CA: Morgan & Claypool. doi: 10.2200/S00429ED1V01Y201207AIM018

Sharma, M., and Bilgic, M. (2017). Evidence-based uncertainty sampling for active learning. *Data Mining Knowledge Discov.* 31, 164–202. doi: 10.1007/s10618-016-0460-3

Tamposis, I. A., Tsirigos, K. D., Theodoropoulou, M. C., Kontou, P. I., and Bagos, P. G. (2019). Semi-supervised learning of hidden markov models for biological sequence analysis. *Bioinformatics* 35, 2208–2215. doi: 10.1093/bioinformatics/bty910

Triguero, I., Garcia, S., and Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge Inform. Syst.* 42, 245–284. doi: 10.1007/s10115-013-0706-y

Tsakanikas, P., Karnavas, A., Panagou, E. Z., and Nychas, G. J. (2020). A machine learning workflow for raw food spectroscopic classification in a future industry. *Nat. Sci. Rep.* 10, 1–11. doi: 10.1038/s41598-020-68156-2

van de Voort, F. R. (1992). Fourier transform infrared spectroscopy applied to food analysis. *Food Res. Int.* 25, 397–403. doi: 10.1016/0963-9969(92)90115-L

van Engelen, J. E., and Hoos, H. H. (2020). A survey on semi-supervised learning. *Mach. Learn*. 109, 373–440. doi: 10.1007/s10994-019-05855-6

Velusamy, V., Arshak, K., Korostynska, O., Oliwa, K., and Adley, C. (2010). An overview of foodborne pathogen detection: in the perspective of biosensors. *Biotechnol. Adv*. 28, 232–254. doi: 10.1016/j.biotechadv.2009.12.004

Wang, X., Rai, N., Pereira, B. M. P., Eetemadi, A., and Tagkopoulos, I. (2020). Accelerated knowledge discovery from omics data by optimal experimental design. *Nat. Commun*. 11, 1–9. doi: 10.1038/s41467-020-18785-y

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometr. Intell. Lab. Syst*. 2, 37–52. doi: 10.1016/0169-7439(87)80084-9

Wu, D., Lin, C.-T., and Huang, J. (2019). Active learning for regression using greedy sampling. *Inform. Sci*. 474, 90–105. doi: 10.1016/j.ins.2018.09.060

Yang, X., Wisuthiphaet, N., Young, G. M., and Nitin, N. (2020). Rapid detection of *Escherichia coli* using bacteriophage-induced lysis and image analysis. *PLoS ONE* 15, e0233853. doi: 10.1371/journal.pone.0233853

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Scholkopf, B. (2003). "Learning with local and global consistency," in *International Conference on Neural Information Processing Systems (NIPS)* (Vancouver, BC), 321–328.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.