NIRWatchdog: Cross-Domain Product Quality Assessment Using Miniaturized Near-Infrared Sensors

Hui Huang, Yangjie Xu, Jin Zhang, Radu State

***Abstract*—Near-infrared spectroscopy (NIRS) has been widely applied to quality assessment for various products. The recent breakthrough in the miniaturization of NIR sensors allows users to scan samples onsite and get results in seconds, making the technology suitable for mobile sensing and IoT applications. However, external factors such as temperature, humidity, and illumination can affect the sensor response and the samples, leading to distorted spectra. This causes a domain shift problem in statistical learning algorithms where the spectra collected from one environment may have a different distribution from the spectra collected from another environment. As a result, the performance of the pre-trained model can be severely degraded when the operating environment differs significantly from the training one. Existing works suggest fine-tuning the pre-trained model using the spectra of reference samples collected from the target environment. Although the number of samples required for model fine-tuning is usually much smaller than that required for model pre-training, it is still impractical to ask users to always carry many reference samples in mobile sensing scenarios. This article presents the NIRWatchdog to address the cross-domain issue of NIRS-based mobile sensing tasks. The proposed approach provides much flexibility and practicality as the transfer dataset can be automatically generated based on as few as one reference sample onsite. With only one reference sample, the AUC of the NIRWatchdog is higher than 85% even when the target environment is considerably different from the training one. In comparison, the conventional approach needs more than 15 reference samples onsite to achieve a comparable performance under the same conditions.**

***Index Terms*—cross-domain, near-infrared spectroscopy, mobile sensing, product quality assessment, anti-counterfeiting**

I. Introduction

Near-infrared spectroscopy (NIRS) is a popular spectroscopy method that measures the infrared light reflected off objects, providing detailed information about the chemical compositions and physical attributes in the form of a spectrum. Conventionally, NIR spectrometers are not portable and expensive, so they are often used in laboratory environments. The recent breakthroughs in the miniaturization of NIR sensors and their integration with wireless communication interfaces, such as Bluetooth and WiFi, make the technology suitable for a wide range of mobile sensing scenarios and Internet of Things (IoT) applications.

A promising application is onsite product quality assessment: extracting compact and descriptive features from spectra of samples with desired quality and standard (i.e., in-class spectra) to score the quality of input spectra and identify nonconforming ones. This is highly desirable for stakeholders to control product quality internally and externally in anti-counterfeiting and line quality control scenarios as the samples in question can be assessed onsite, saving labor costs and time consumption on appropriate packaging, transportation, and storage. Several recent works have shown the vast potential of miniaturized NIR sensors in such tasks. For example, the authors in [1] developed a prototype system that pairs smartphones with portable NIR spectrometers, allowing non-experts to quickly scan samples onsite. The spectra can be uploaded and analyzed in the cloud to provide almost instantaneous results. They later extended the system to identify specific pharmaceuticals [2] or commercial drinks [3], [4] using machine learning approaches.

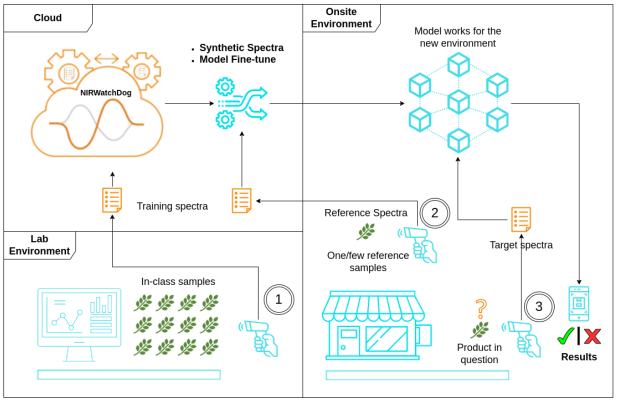


Fig. 1: A typical onsite product quality assessment scenario. Step 1: Collect spectra from in-class samples in a Lab and train the NIRWatchdog on the cloud. Step 2: scan one reference sample onsite to adapt the pre-trained model to the new environment. Step 3: use the fine-tuned model to assess the quality of the product on site.

Although device miniaturization improves the portability of NIRS, due to the adoption of low-cost hardware and lack of control over the scanning environment, external factors, including temperature, humidity, and illumination, lead to different spectral forms of absorption regions and intensity [5], [6]. Since NIRS is a non-selective sensing technology, both the sensor response and the samples could be affected

This work was supported by the Luxembourg National Research Fund (Fonds National de la Recherche - FNR) under PoC grant of NirWatchDog.

Corresponding authors: Jin Zhang.

Hui Huang, Yangjie Xu, and Radu State are with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg (e-mail: hui.huang, yangjie.xu, radu.state@uni.lu).

Jin Zhang is with National Engineering Laboratory for Big Data System and Computing Technology, College of Computer Science and Software Engineering, Shenzhen University (e-mail: jin.zhang@szu.edu.cn).

by external conditions. For example, the spectra of dry materials are susceptible to the humidity while samples with higher moisture content suffer more from the temperature fluctuations [7]. The temperature fluctuation can also cause different levels of drift in wavelength calibration depending on the quality of hardware adopted [8]. Regarding onsite applications with portable NIR equipment, the obtained spectra may be influenced by environmental illumination. This level of random variation may cause distribution shifts between spectra of varying environments, resulting in the well-known domain shift problem in most statistical learning algorithms [9]. Consequently, the pre-trained model may suffer severe performance degradation when the operating environment significantly differs from the training one [10]. The cross-domain issue is particularly problematic in anomaly detection tasks because the model could give opposite classification results: in-class samples may be incorrectly identified as anomalies, while non-conforming ones may be missed.

Intuitively, collecting more training data under varying conditions could improve the model's cross-domain ability in new environments. However, this approach increases the workload on data collection and incurs high costs in setting up a laboratory that allows flexible control over external factors. The anomaly detection model might also learn a generic representation of underlying regularities from spectra collected in different domains, which are not optimized for detecting irregularities [11]. Another solution is to apply conventional calibration transfer methods that are widely adopted in benchtop instruments, such as direct normalization (DS), piecewise direct standardization (PDS) [12], [13], and canonical correlation analysis (CCA) [14] to correct spectra from the target environments. Unfortunately, the comparative study conducted in [15] shows that these methods do not work well with miniaturized spectrometers in mobile scenarios. This is because benchtop spectrometers are usually used in a laboratory with controlled scanning environment conditions, the equipment aging, replacement of parts and batch-to-batch variation of samples are the main contributing factors to the domain shift issue. Yet miniaturized spectrometers have to work under frequently changed environments, conventional calibration transfer methods utilize simple linear transformations to characterize the impact of external factors and thus cannot represent the complex relationship among possible environmental changes.

To address this issue, the state-of-the-art [16], [17], [18], [19] suggests using model fine-tuning to adapt pre-trained models to new and unseen environments. Such methods train a base model using available spectral data and then fine-tune it with a transfer dataset collected from the target environment. The size of the transfer dataset is usually much smaller than the training dataset: the experiment results reported in [15] show that spectra of around 15 to 20 reference samples per class would be required to update the pre-trained model effectively. However, manually collecting transfer data is time-consuming for onsite tasks as the operating environment consistently changes over time. In addition, users need to carry many reference samples with them, and transportation and storage costs make this approach impractical in many mobile sensing scenarios.

This paper presents the NIRWatchdog to address the cross-domain issue of NIRS-based onsite product qualtiy assessment. Figure 1 illustrate a typical onsite product quality assessment scenario using the proposed system. Like the state-of-the-art, the NIRWatchdog fine-tunes the pre-trained model in new and unseen environments using transfer datasets. However, it provides much flexibility as the transfer dataset can be automatically generated based on as few as one reference sample onsite. Therefore, the proposed approach is more practical in mobile sensing tasks. The proposed approach is based on the observation from the spectra of two off-the-shelf commodities in different environments: although the spectra collected in different environments, even for the same batch of samples, exhibit low similarity, the distributions of their spectral deviations are very close to each other. This finding encourages us to explore the possibility of generating synthetic transfer datasets as if they were collected in the target environments to reduce the cost of onsite data collection and simplify the process of model fine-tuning.

At the highest level, the NIRWatchdog consists of two deep learning models: a base anomaly detector and a synthetic spectra generator. The base anomaly detector learns the latent pattern from in-class spectra collected in the training environment (e.g., a laboratory) to identify abnormal samples. The synthetic data generator takes as input the spectra of one/a few reference samples collected from the target environment to generate the synthetic transfer dataset. Both the anomaly detector and the data generator are trained one-off using available in-class spectra only. When the operating environment changes from the training environment, the base anomaly detector can be fine-tuned by scanning one/few reference samples onsite.

The main contributions of this paper are:

- Conduct experiments using tea bags and hazelnut cocoa spread to investigate the impact of environmental factors on spectra data. The results show that the changes in the operating environment result in varying data distributions even for the same batch of samples. Still, the distributions of their spectra deviations remain similar from one environment to another.
- Propose a novel approach to enable fast and efficient model adaptation in new environments. The proposed approach relaxes the requirement of collecting transfer datasets from many samples onsite; a synthetic dataset can be generated on demand with as few as one reference sample. The generated dataset can then fine-tune the base anomaly detector to reuse the learned knowledge in the training environment.
- Design and implement the NIRWatchdog system and conduct extensive experiments using 10 off-the-shelf commodities representing varying characteristics and compositions to demonstrate the benefits and performance of our proposed approach.

The remainder of the paper is organized as follows. Section II presents related works. Section III conducts priliminary studies to demonstrate the impacts of the operating environment on the spectra. Section IV-B presents the overall architecture of the proposed NIRWatchdog system and detailed designs of each component. Section V comprehensively evaluates the system by conducting experiments in varying environments. Finally, Section VI concludes the paper.

II. Related works

A. Near-infrared spectroscopy

Near-infrared spectroscopy (NIRS) is a popular spectroscopic method that uses near-infrared light in $780 - 2500$ nm. The working principle is that hydrogen-containing groups, such as O-H, N-H, C-H, and S-H, tend to absorb specific frequencies of the incident optical radiation in near-infrared frequency bands [20]. Therefore, the spectra of samples contain rich information on their inner configurations. As a fast, efficient, and non-invasive chemometric technique, NIRS has remarkable value in analyzing the chemical compositions and physical properties of various objects. Many machine learning-based chemometric models, including partial least squares (PLS) regression, random forest (RF), and support vector machine (SVM), have been developed to extract the chemical concentrations or unique patterns from input spectra. Recently, deep learning-based approaches have become increasingly popular as they can overcome the limitations of shallow machine learning by automatically extracting informative features from raw data without prior knowledge to improve accuracy and robustness further [21], [22], [23].

The last decade has witnessed the fast development of the miniaturization of NIR spectrometers, making the technology suitable for mobile sensing scenarios. Many works have been conducted to assess the performance of portable NIR spectrometers on onsite tasks, such as material/object classifications [2], [24], liquid sensing [3], [4], [25], food calorie estimation [26], soil composition estimation [27], and medical examination [18], [28].

B. Calibration transfer

Due to noise caused by the external factors of the detection environment and system errors of spectrometers, a machine learning model trained on spectra collected from one domain could be invalid in another [10]. This is because that temperature, humidity, illumination, and many other factors could cause complex impacts on both the spectrometers and the samples themselves [13], [5], resulting in spectral distortion in the forms of baseline drift, additive, and multiplicative effects, and random noises. Therefore, even for the same set of samples, the distribution of spectral data collected from different scanning conditions may differ significantly.

Many approaches have been developed to address this issue. They are collectively referred to as calibration transfer or calibration maintenance in the literature [15]. Earlier works focus on reducing spectra variances due to different scanning conditions by employing appropriate spectral preprocessing to establish robust models. For example, taking the first and second derivatives on the spectra can remove baseline offsets and scatter effects, respectively; performing multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations reduce the impacts of scattering and enhance the correlation between spectral absorption and the target concentration [29], [30]. The main advantage of this class of approaches is simplicity. However, as they do not incorporate prior knowledge about the scanning conditions, the improvements are limited for most cases.

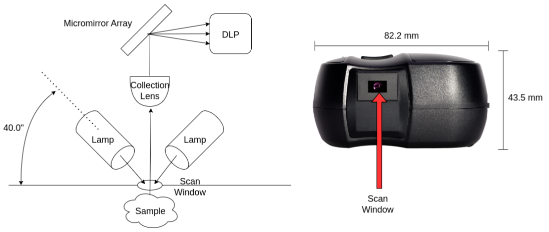


Fig. 2: The schematic diagram of the portable DLP-based NIR sensor used in this study.

To overcome the limitations of simple signal preprocessing, some works propose transforming the spectra under new conditions to resemble the spectra on the original condition so that the trained model can be directly used for prediction. Representative methods include direct normalization (DS), piecewise direct normalization (PDS) [31], and canonical correlation analysis (CCA) [14]. They have in common the need to select a set of standard samples according to specific rules [12] and then compute a transformation matrix to establish the relationship between their spectra scanned under two conditions. This way, the impacts on the spectra due to the changes in the involved conditions are explicitly incorporated into the calibration transfer, thus improving the adaptability of the trained model. However, these methods were initially proposed to transfer models between different benchtop spectrometers. In onsite tasks, the scanning environments might change significantly, and the impacts could be too complex to be described by simple linear transformations. More recently, a few research [16], [19], [18], [17] have been conducted to implement calibration transfer using transfer learning methods. Rather than mapping spectra to the original conditions, these works attempt to fine-tune the trained model using a transfer dataset collected under the new conditions to preserve the knowledge learned from the original dataset. The comparative study conducted in [15] shows that the transfer learning based method has a dominant advantage in prediction accuracy and stability when the new condition significantly differs from the original one.

III. Motivation and Preliminary Study

A. Environmental impacts

The preliminary study in this section uses tea bags and hazelnut cocoa spread to investigate the impacts of operating environments. The reason for choosing these two commodities is that the hazelnut cocoa spread is sealed in glass bottles and is less affected by the external environment, while the tea bags are dry materials wrapped in fabric bags, which are more sensitive to changes in temperature and humidity. The spectra of 40 Earl Grey tea bags and 20 jars of hazelnut cocoa spread were collected from three sites in Luxembourg using a portable NIR sensor as shown in Figure 2. The data collection at each site was conducted under different weather conditions to increase the diversity of the operating environments. Table I summarizes the environmental parameters of the experiments. Each teabag was scanned 5 times at each site, while each jar of hazelnut cocoa spread was scanned 10 times, yielding a total of 200 spectra per site per commodity. More details of the spectrometer configuration and data collection protocols are given in Section V.

The collected spectra were pre-processed following the steps described in Section IV-B to remove baseline offsets and background noise. Figure 3a compares the average of pre-processed spectra of the tea bags collected from the above three sites. [1]. As can be seen from the figure, the changes in environments cause complex variations and distortions in the absorbance of the collected spectra, especially at the three peaks around 1150, 1400, and 1500 nm, which differ in both magnitude and wavelength. The spectra collected from **Site-3** exhibit large fluctuations from 900 to 1300 nm, which are not as smooth as those of **Site-1** and **Site-2**. Figure 3b complements our observation by projecting the processed spectra onto a 2-dimensional space using t-distributed stochastic neighbor embedding (t-SNE). It is obvious that the t-SNE embeddings of spectra collected from 3 different sites are grouped into 3 distinct clusters, implying a low similarity of the data distributions over varying environments.

To gain additional insights into the impacts of the operating environments, the spectral deviations, defined as the differences from the average spectra of the dataset, are computed and analyzed using the same method as the above. Since spectral deviations exhibit more substantial randomness, they are projected into 3-dimensional space using t-SNE. Figure 3c visualizes the t-SNE embeddings on one of the 2-D planes of projected space; the other two planes lead to the same conclusion, so they are not presented. Interestingly, even though the projections of spectral deviations from the same environment tend to stay closer, there does not exist a clear separation boundary: spectral deviations from different sites mix together, just like they were collected from the same environment.

To quantify the difference, the Earth Mover's Distance (EMD), also known as the Wasserstein distance, is employed to measure the similarity of the spectra and spectral deviations between different sites. Given the limited data, it is difficult to calculate the EMD between two datasets using their empirical multivariate distributions. Instead, the spectral measurements at each wavelength are treated as a feature and compute the average EMD between all features of the two spectral datasets. Specifically, each feature is first normalized between 0 to 1 using the minimum and maximum values observed from the three datasets. The frequency distribution of each feature is then computed by grouping them into 10 bins ranging from 0 to 1. Let $\mathbf{p^k}$ and $\mathbf{q^k}$ denote the non-negative vector representing the frequency distribution of the kth spectra feature from two datasets. By definition, $||\mathbf{p^k}||_1 = 1$ in the context of this work, the EMD between them can be calculated as [32]:

$$EMD(\mathbf{p^k}, \mathbf{q^k}) = \sum_{i=1}^{10} |\sum_{j=1}^{i} (p_j^k - q_j^k)| \qquad (1)$$

The average EMD of two datasets is then defined as $\frac{1}{N}\sum_{k=1}^{N} EMD(\mathbf{p^k}, \mathbf{q^k})$, where N is the number of features in the spectra vector. By definition, the greater the difference between the distributions of the two datasets, the larger the average EMD is. Since the frequencies are binned into 10 buckets, the maximum value of the distance is 9.

Table II presents the average EMD of the spectra and spectral deviations of tea bags and hazelnut cocoa spread between different sites. Overall, the EMD of the spectral deviation of samples under different environments is an order of magnitude smaller than the EMD of their spectra, with the average value of tea bags ranging from 0.2 to 0.4 and the average value of hazelnuts ranging from 0.1 to 0.3. The results are consistent with the t-SNE analysis: the spectral deviations under different operating environments have closer distributions.

B. Lessons learned

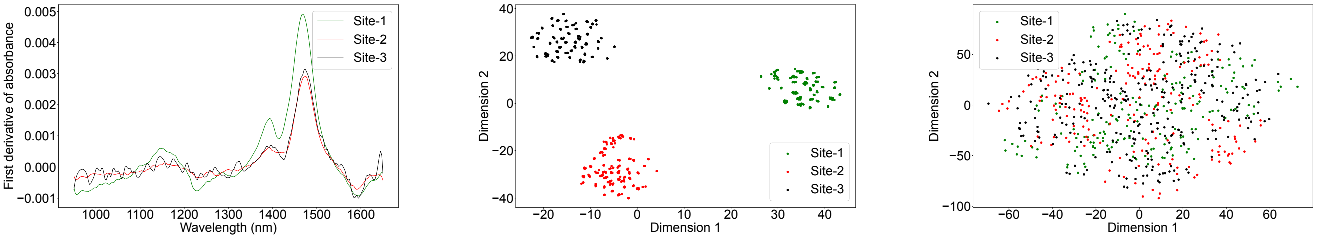
The above experiments demonstrate how alterations in environmental conditions affect the spectra of a particular sample batch. These changes impact the amplitude, smoothness, and slope of the spectra. The visualization results from t-SNE and the quantitative analysis using average EMD reveal significant distribution shifts of spectra between different environments. As most machine learning-based models assume that the training and testing data are independent and identically distributed (i.i.d.), applying the trained model on spectra collected from new and unseen environments would violate this assumption and lead to significant performance degradation. In Section V, we conduct experiments to investigate this issue in more details. However, it is interesting to see that the distributions of the spectra deviations of different datasets remain similar. In other words, a generative model learned from the spectral deviations of the training dataset could be used to approximate the data generation process of spectral deviations in new and unseen environments. The generated spectral deviations, combined with a few onsite collected spectra, can be used as the transfer dataset to fine-tune the model to adapt to the target environment.

It is noteworthy that the EMDs of the spectral deviation of **Site-1&2** and **Site-1&3** are higher than that of **Site-2&3**. This is because the data collection at **Site-1** was conducted in early spring rather than mid-summer, with more differences in temperature and humidity than the other two sites. Therefore, this distance may continue to increase when environmental changes intensify further. In this sense, the assumption of sufficiently similar distributions of spectral deviations could be compromised in the case of extreme changes in external conditions. However, we argue that the operating environments in this study cover considerable variations, with a maximum temperature difference of approximately 16 degrees, a 58%

[1] The paper only shows the figures of spectra of tea bags as we are not allowed to mention the brand of the hazelnut cocoa spread nor disclosure the spectra information of their products according to the nondisclosure agreement signed between the company and us.

TABLE I: Environment parameters of the data collections over three sites

ID	Location	Temperature	Humidity	Illumination	Date
Site-1	SnT, Outdoor	16.7 C	73 %	109 Lux	22/03/2022
Site-2	SnT, Indoor	24.6 C	29 %	323 Lux	13/07/2022
Site-3	Kayl, Outdoor	32.1 C	15 %	597 Lux	19/07/2022



(a) (b) (c)

Fig. 3: (a) The average spectra of the same tea bag batch collected from three sites, (b) the t-SNE embeddings of the three spectra datasets in a 2-D plane, and (c) the t-SNE embeddings of the **spectral deviation** of the same tea bag batch collected from three sites. The t-SNE embeddings of spectra collected from 3 different sites are grouped into 3 distinct clusters, implying a low similarity of the data distributions over varying environments. However, the embeddings of the **spectral deviation** of different sites are projected into proximity region; there does not exist a clear separation boundary.

TABLE II: The average EMD between the spectra and spectra deviations collected from the three sites

Sites	EMD of Spectra	EMD of Spectra deviations
1&2	3.3168	0.3652
1&3	**3.5126**	**0.4085**
2&3	1.4361	0.2762

(a) Teabag

Sites	EMD of Spectra	EMD of Spectra deviations
1&2	**2.4642**	0.2314
1&3	2.1570	**0.2916**
2&3	1.0826	0.1232

(b) Hazelnut cocoa spread

difference in humidity, and a 488 lux difference in illumination. The results presented in this study is still of practical value in the applications of on-site anomaly detection.

IV. System Model

A. Problem statement

The output of the NIR spectrometer is a one-dimensional vector $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{n})$. Depending on the operation mode (e.g., transmission, reflection, transflection, or interactance), each value x_i represents the measured intensity of the incident light at a certain wavelength in the near-infrared region of the electromagnetic spectrum (i.e., 780 nm to 2500 nm). The sensor response varies with the composition and inner configuration of the scanned sample, therefore, contains rich structural information at the molecular level. This paper investigates the use of miniaturized NIR spectrometers for onsite anomaly detection. Unlike existing works that focus on building regression models to predict specific compositions [33], [4], [34], this work aims at learning a binary classifier to determine whether query spectra come from abnormal samples. In practice, abnormal samples are usually unknown beforehand, and their inner configurations might differ completely. Only in-class samples are assumed to be available during training. Therefore the problem can be well formulated as a one-class classification task.

B. System overview

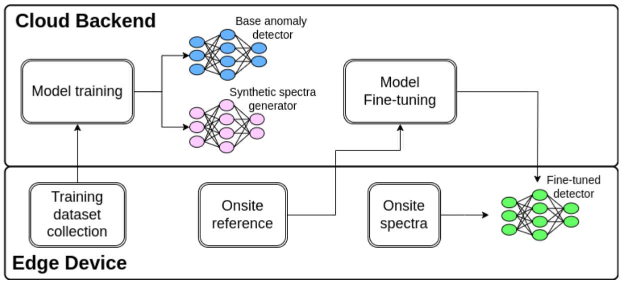
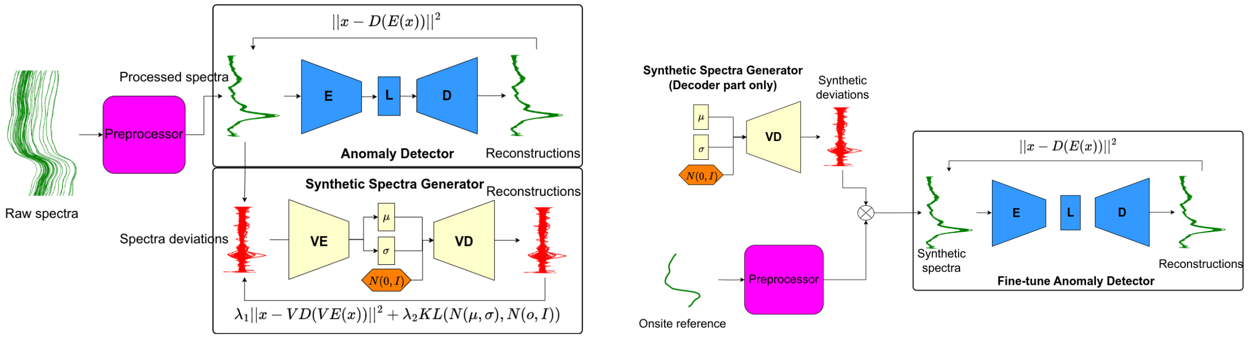


Fig. 4: Overview of the NIRWatchdog system.

The overall architecture of the proposed NIRWatchdog is shown in Figure 4. The system consists of two conceptual subsystem: the cloud backend and the edge device (the NIR sensor). The cloud backend is responsible for training the base anomaly detector and the synthetic spectra generator using the training dataset collected from the edge devices, as well as fine-tuning the base detector to onsite environments. The fine-tuned model is then off-loaded to the edge to predict the quality of the samples in question.

As shown in Figure 5, the major components of the system include the pre-processor, the anomaly detector, and the synthetic spectra generator. The pre-processor aims at improving spectra quality by reducing background noise and the impacts of physical factors. The base anomaly detector learns compact and descriptive feature representations from in-class spectra to identify non-conforming samples. Based on the assumption that the distribution of spectra deviations is largely environmentally independent, the synthetic spectra generator produces transfer datasets as if they were collected from the target environments using the knowledge learned



(a) (b)

Fig. 5: The detailed design of the NIRWatchdog system (a) The training phase. (b) The onsite fine-tuning phase

from the spectra deviations of the training dataset. In the fine-tuning phase, the pre-trained anomaly detector can be quickly adapted using the generated spectra.

C. Spectra pre-processing

Pre-processing is an important step on NIR spectra analysis. In this work, the pre-processor performs re-sampling, spectra derivatives, and signal smoothing to eliminate noise from various sources. The re-sampling is necessary as the wavelength sampling of spectra (i.e., x axis) collected by different spectrometers or even the same spectrometers under different environments might not be the same. This is because the mapping between measurements and wavelength is estimated by a quadratic equation, and the wavelength step is not uniform. Re-sampling help improves the consistency by aligning the wavelength to the same ranges. The cubic spine interpolation is applied to re-sample the spectra. The raw data contains 228 measurements. After re-sampling, a spectra has 400 data points, evenly distributed from 900 to 1700 nm. Two sections of the spectra, 900 – 950 nm and 1650 – 1700 nm, exhibit random variations as the signals are susceptible to the spectrometer's temperature and thus do not contain informative features. Therefore, only the measurements from 950 to 1650 nm are preserved for further analysis.

Next, the pre-processor takes numerical $1st$ derivatives on the re-sampled spectra to remove additive and multiplicative effects. This simple approach can reduce baseline offset, compensates for hardware drift, and enhances minor spectra variations. However, derivatives tend to increase noise. To solve this problem, the spectra are finally smoothed using the Savitzky-Golay (SG) filter to improve the signal-to-noise ratio (SNR) without greatly distorting the spectra. The SG filter applies convolution on a sequence of data points. The least-squares method with polynominal of order n is used to fit adjoining data points:

$$x_j* = \frac{1}{N}\sum_{h=-k}^{k} c_h x_{j+h} \tag{2}$$

Where x_j is the jth point after smoothing, N is the normalising coefficient, k is the window size defining the number of adjoining data points and c_h is a coefficient depends on the selected nominal order. The best combination of the window size and the polynomial order depends on the scanned samples. In this work, the window size is chosen as 21, and the polynomial order is set as 3.

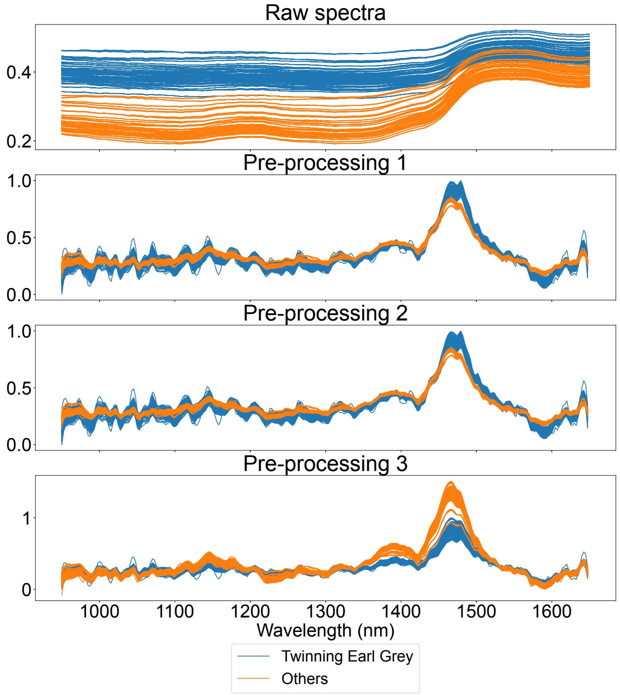


Fig. 6: Comparison between the effects of three different pre-processing procedures on Twining earl tea bags and tea bags of other types. The MSC and SNV tend to reduce the dissimilarity between spectra collected from samples of different types.

Note that unlike conventional procedures reported in the literature, the pre-processor of the NIRWatchdog neither applies multiplicative scatter correction (MSC) nor standard normal variate (SNV). This is because recent works show that deep learning models can learn similar transformation behavior compared with MSC and other commonly used spectra pre-processing algorithms [35], [36], [37]. Excessive pre-processing could compromise the information contained in the training dataset. To see this, Figure 6 compares the spectra after pre-processing using the above-described pipeline with the following two alternative pipelines:

- Pre-processing 1: spectra re-sampling, SNV, $1st$ derivatives and SG smoothing.

- Pre-processing 2: spectra re-sampling, MSC, $1st$ derivatives and SG smoothing.

The spectra in Figure 6 are collected from 40 Twining earl grey tea bags and 20 tea bags of other types. For **Pre-processing 2**, the MSC uses the mean of the earl grey spectra as the reference spectrum. As expected, MSC and SNV achieves similar effects on the raw spectra: the pre-processed results are more condensed and the noise due to scattering has been filtered away. However, pre-processing pipelines involving MSC or SNV tend to reduce the dissimilarity between spectra collected from earl grey and other tea bags, while our method maintains the spectral differences between types of samples while reducing noise. In Section V, we conduct experiments to show that additional MSC or SNV does not improve overall anomaly detection performance and instead leads to slightly worse performance.

D. Anomaly detector

It is assumed that only in-class samples are available during model training, as collecting many out-class samples in practice is difficult. Therefore, the proposed system employs an autoencoder network as the anomaly detector to only learn the data regularities from in-class spectra. The anomaly detector comprises an encoding network (E) and a decoding network (D), parameterized by θ_e and θ_d, respectively. The encoder E maps the in-class spectra onto a latent feature space, and the decoder D tries to reconstruct the input data from there. The usual practice is to have a bottleneck network architecture, where the dimension of the latent space is much smaller than the input data, to prevent the autoencoder from learning the identity function. Similar to the work in [38], the $tanh$ activation function is used in the output layer of the E to have bounded support $(-1, 1)^d$, where d is the dimension of the latent space. The anomaly detector is trained with the objective of minimizing the reconstruction error defined as the $L2$ norm between the input spectra and reconstructed spectra:

$$\theta_e^*, \theta_d^* = \underset{\theta_e, \theta_d}{\arg\min} \sum_{\mathbf{x} \in \chi} ||\mathbf{x} - D(E(\mathbf{x}; \theta_e); \theta_d)||^2 \tag{3}$$

Where $\chi = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_m\}$ are the training dataset containing m in-class spectra. Minimizing the reconstruction error enforces the encoder to learn latent features that retain the most relevant information to reconstruct the data instance. Consequently, the mapping from out-class spectra to the low-dimensional space could produce irregular features that lead to poor reconstructions. The reconstruction error of a input spectra $s_{\mathbf{x}} = ||\mathbf{x} - D(E(\mathbf{x}; \theta_e); \theta_d)||^2$ can therefore be directly used as the anomaly score: the greater the reconstruction error, the more likely the tested sample is an outclass sample.

When outputting classification results is required, the anomaly detector can utilize a threshold of reconstruction errors to distinguish between in-class samples and non-conforming ones. The threshold could be determined through either a supervised or unsupervised method. In the supervised method, the value can be tuned as a hyper-parameter based on the performance of detecting labeled anomalies. However, this approach is not suitable for our case since only in-class samples are assumed to be available during the model training. Additionally, even if labeled out-class samples are available, the threshold found in this way might be too aggressive if these samples differ significantly from normal ones. This work adopts the unsupervised scheme that sets the threshold by statistical measures on the reconstruction errors obtained from the in-class samples. Commonly used indicators include n-percentile, median plus a specified interquartile range, and z-scores. Due to its simplicity, the anomaly detector uses the 95 percentile of the reconstruction errors as the threshold, which is the most commonly used in many works.

E. Synthetic spectra generator

The objective of the synthetic spectra generator is to generate a transfer dataset using spectra collected from a few, or even just one, randomly chosen reference samples (i.e., in-class samples) onsite as if they were collected from the target environment. Conventionally, the transfer dataset has to be collected manually from many reference samples, which is impractical in mobile sensing tasks. According to the observations in Section III, the distribution of spectral deviations remains similar from one environment to another. Therefore, the NIRWatchdog first generates synthetic spectra deviations using a VAE trained on the training dataset and then adds the generated spectra deviations to the average of spectra collected from one or a few reference samples in the target environment to produce the transfer dataset. The assumption here is that the average of the collected spectra is close to the mean of the in-class spectra distribution. Indeed, the quality of the synthetic spectra generated using one/a few reference samples strongly depends on the variance of the in-class spectra distribution. If the variance is large, more reference samples are needed to approximate the dataset mean. However, in real-world applications, high-quality commodities that meet specific standards and requirements tend to have similar chemical and physical characteristics, so the distribution of their spectra can be assumed to have a small variance. Section V shows that a randomly chosen reference sample can still perform satisfactorily using our approach.

Specifically, the core component of the synthetic spectra generator is a variational autoencoder (VAE) trained using the spectra deviations of the training dataset. The VAE is a form of autoencoder designed to learn the data distribution using principles of variational inference. Similar to the standard anomaly detector, it also consists of an encoding network (VE) and a decoding network (VD). Yet, the encoder VE maps the input data into a normal distribution $N(\mu, \sigma)$ over the latent space instead of a single point. The decoder VD tries to reconstruct the input by sampling that distribution. In order to ensure the learned latent space is continued and complete, which is essential to enable the generative process, the training of the VAE is regularised as follows:

$$\begin{aligned}\theta_{ve}^*, \theta_{vd}^* = & \underset{\theta_{ve}, \theta_{vd}}{\arg\min} \sum_{\mathbf{x} \in \chi} [\lambda_1 ||\mathbf{x} - VD(VE(\mathbf{x}; \theta_{ve}); \theta_{vd})||^2 \\ & + \lambda_2 KL(N(\mu, \sigma), N(0, I))]\end{aligned} \tag{4}$$

Where θ_{ve} and θ_{ve} are parameters of VE and VD respectively. As can be seen from the equation, the loss function is composed of the reconstruction term that minimizes the $L2$ norm between the input and reconstructed spectra deviation and a regularization term expressed as the Kulback-Leibler divergence between $N(\mu, \sigma)$ and a standard Gaussian. λ_1 and λ_2 are tunable hyper-parameters representing the importance of these two loss terms. After the training process has been completed, the VD can be used as a data generator to produce synthetic spectra deviations with a similar distribution to the training dataset's spectra deviations.

F. Model adaptation

Since the early layers of a deep neural network usually focus on abstracting the input data and are largely independent of the optimization goal [39], similar to previous works [19], [18], [17], the NIRWatchdog applies transfer learning techniques to adapt the base anomaly detector to new and unseen environments. The knowledge learned from the training environment can then be reused to reduce the cost of model re-calibration.

Specifically, the proposed approach fine-tunes the pre-trained model to the target environment at the parametric level. This can be done by copying the learned parameters of the base anomaly detector to initialize a new model and then training it with the generated synthetic transfer dataset. During fine-tuning, the model's architecture remains the same as the base anomaly detector. All layers are frozen except the output layer of the encoder, and the output layer of the decoder are allowed to update to learn new variability in the transfer dataset. The base anomaly detector is fine-tuned with a variable and smaller learning rate than the pre-training: it linearly increases from 0 to 0.0001 for the first few epochs to warm up and then linearly decreases to 0 afterward.

V. Evaluations

A. Samples and data collection protocol

A portable NIR spectrometer, InnoSpectra NIR-S-G1, as shown in Figure 7, was used to collect spectra from prepared samples. NIR-S-G1 is a diffuse reflective spectrometer that uses two built-in 0.7 W tungsten filament lamps as light sources. The optical engine adopts a post-dispersive architecture consisting of slits, collimating lenses, bandpass filters, digital micromirror devices, and single-point InGaAs detectors. The working spectral region of the device covers a wide part of the near-infrared band, ranging from 900 – 1700 nm.

Before scanning samples, the spectrometer was calibrated using a Spectralon 99% diffuse reflectance standard. According to the recommendation of the NIR-S-G1 user manual, the device needs around 5 minutes of warm-up, during which the reflectance standard was scanned repeatedly until the resultant spectra stabilize. For each scan, the spectrometer was configured to produce the average of 6 consecutive spectra. Exposure time was set to 0.635 ms, resulting in a spectra bandwidth of 7.03 nm and 228 data points ranging from 900 nm and 1700 nm. The scanning type was set to Hadamard to provide better SNR. Note that when scanning the reference Spectronlon, the device automatically determines a maximum PGA gain allowed in the current environment (affected mainly by the environment's illumination). This parameter was set to the same when scanning samples. Table III summarizes the configuration used in our experiments.

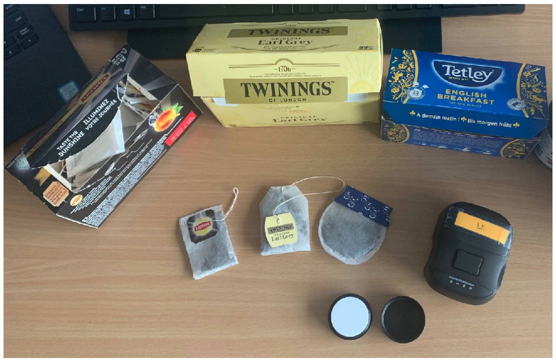


Fig. 7: The teabags and NIR spectrometer used in our experiments.

TABLE III: Spectrometer configurations

Scan type	Start wavelength
Hadamard	900 nm
End wavelength	**Wavelength width**
1700 nm	7.03
Exposure time	**Num. of scans to average**
0.635 ms	6
PGA gain	**Number of data points**
Auto	228

Two off-the-shelf commodities, tea bags, and hazelnut cocoa spread, were used as examples to validate the proposed NIRWatchdog system. In the tea bag experiments, 60 packs of Twinings Earl Grey were used as in-class samples. The out-class samples include 10 packs of Lipton Earl Grey and 10 packs of Tetley English Breakfast tea. Note that there is no assumption that Lipton's Earl Grey tea is of lower quality than Twinings'. Ideally, the out-class samples should be counterfeit or low-quality Earl Grey tea bags but they are difficult to obtain. The main objective of the experiments is to evaluate the ability of the proposed NIRWatchdog to differentiate between the manufacturing variations of different brands. In the hazelnut cocoa spread experiments, 20 jars of qualified products from a well-known brand were used as in-class samples, and 4 jars of counterfeits discovered on the market as out-class samples [2]. Since both the tea bags and the hazelnut cocoa spread are packaged to allow light to pass through, they were scanned without the packaging being removed. The tea bag was placed on top of the spectrometer lens, each scanned 5 times. Between two consecutive scans, the tea bag is moved slightly around the lens to increase the diversity of collected spectra. As a result, the tea bag dataset per site includes 300 spectra of Twinings earl grey, 50 spectra of Lipton Earl Grey, and 50 spectra of Tetley English Breakfast tea. When scanning the hazelnut cocoa spread, the jar was turned upside down to scan 10 random locations on

[2] We are not allowed to mention the name of the brand according to the nondisclosure agreement signed between the company and us

TABLE IV: Network architecture of the anomaly detector and the synthetic spectra generator

Anomaly Detector				Synthetic Spectra Generator			
Network	**Layer**	**Activation**	**Output**	**Network**	**Layer**	**Activation**	**Output**
Encoder				Encoder	Input	-	(279,1)
	Input	-	(279,1)		Conv1D(8,5)	ELU	(138,8)
	Conv1D(16,16)	ELU	(279,16)		BN-DP	-	(138,8)
	BN-MP	-	(139,16)		Conv1D(16,5)	ELU	(67,16)
	Conv1D(4,4)	ELU	(139,4)		BP-DP	-	(67,16)
	BN-MP	-	(69,4)		Conv1D(32,8)	ELU	(30,32)
					BN-DP	-	(30,32)
Latent	Flatten	-	(276,1)	Latent	Flatten	-	(960,1)
	Dense	TanH	(30,1)		z_mean	Linear	(10,1)
					z_log_var	Linear	(10,1)
Decoder				Decoder	Conv1D(32,8)	ELU	(7,32)
	Conv1D(4,4)	ELU	30,8		BN-DP-US	-	(14,32)
	BN-US	-	(60,8)		Conv1D(16,5)	ELU	(5,16)
	Conv1D(16,16)	ELU	(60,16)		BN-DP-US	-	(10,16)
	BN-US	-	(120,16)		Conv1D(8,5)	ELU	(5,8)
	Conv1D(1,1)	ReLU	(120,1)		BN-DP	-	(5,8)
	Dense	Linear	(279,1)		Flatten	-	(40,1)
					Dense	Sigmoid	(279,1)

TABLE V: Parameters for model training and fine-tuning

	Learning rate	**Batch size**	**Maximum epoch**
Anomaly Detector	0.001	16	200
Spectra Generator	0.001	16	200
Fine-tuning	0.0001	4	50

the bottom. The collected dataset per site includes 200 spectra of qualified products and 40 spectra of counterfeits. Although the data collected in this way result in a mixture of spectra from the products and their packaging, it is sufficient to study the performance of the proposed system. The teabag dataset has been uploaded to the IEEE DataPort with the DoI of 10.21227/t935-wm91.

B. Experiments setup

The proposed NIRWatchdog was implemented in Python 3.8.3 using Keras 2.8.0 with Tensorflow backend and run on a laptop equipped with an Intel Core i7 3.5 GHz and 16 GB RAM. Determining the optimal model hyper-parameters and network architecture is a highly iterative process. This paper only presents the model with the best performance obtained from our experiments. The encoder E of the anomaly detector employs 2-layer convolutions neural networks (CNN) to map the pre-processed spectra into the latent space. Batch normalization (BN) and max pooling (MP) layers are appended after each 1-D convolution layer (Conv1D) to reduce the risk of over-fitting and stabilize the training process. The decoder D of the anomaly detector incorporates a structure like the inverted encoder to reconstruct the input spectra from the latent space.

Similarly, the VAE used for the synthetic spectra generator is also a CNN, except that it has a deeper depth to capture the complex constitutions of the spectra deviations. Both the encoder VE and the decoder VD have three Conv1D layers, interconnected by BN and dropout layers. The latent space is reparameterized by two fully connected dense layers representing the mean and standard deviation vectors of the learned multivariate normal distribution. Table IV summarizes the network architectures of the anomaly detector and the synthetic spectra generator. All models and the fine-tuning of the base anomaly detector are trained by the Adam optimizer with early stop mechanism, whose training parameters are summarized in Table V.

For each type of the chosen commodities, three sets of experiments were conducted to evaluate the performance of the proposed system. Each set of experiments trained and validated the base anomaly detector and the synthetic spectra generator using in-class spectra from one site. The trained anomaly detector is then fine-tuned using the transfer dataset generated by the synthetic spectra generator to adapt it to the other two sites. The performance of the NIRWatchdog is compared with the baseline approach applied in [16], [17], [18], [19]. Since their works focus on the regression problem (e.g., predicting polymer length or blood glucose) rather than anomaly detection. We slightly modified their algorithm to fit our experiments where the anomaly detector is fine-tuned using the spectra collected from 5, 10, or 15 in-class samples on the target sites. The learning rate, batch size and maximum number of epoch remain the same as the NIRWatchdog. The rest of the paper uses the term **Onsite-N** to refer to the results of the baseline approach, where N represents the number of reference samples used onsite to prepare the fine-tuning dataset.

The main metric to evaluate the anomaly detector is the **Area Under the Curve (AUC) of Receiver Operating Characteristics (ROC) curve**, recall, and precision, most commonly used for one-class problems [40], [41]. To calculated these metrics, the inclass samples are regarded as positive. The performance of the synthetic spectra generator is evaluated by visually inspecting the results of the t-SNE projections together with quantitative analysis using the **average EMD** as described in section III. Each experiment was repeated 10 times with a randomly split training and testing set, and the results were averaged to obtain the final result. Note that since each sample has to be scanned multiple times to collect enough spectra, the training and testing set for each experiment do not contain spectra from the same sample. This can be done by randomly selecting samples first and then putting their spectra into the training set, the spectra of the remaining samples then belong to the testing set.

C. Anomaly detection in the same environment

To evaluate the performance of anomaly detectors in the same environment, 80% of the in-class spectra collected from one site were used to train the model, and the rest of the in-class spectra, together with the out-of-class spectra collected from the same site, were used as the testing set. Table VI shows the average AUC, recall, and precision obtained from the tea bags and hazelnut cocoa spread experiments conducted at the three sites, and Figure 8 plots histograms of reconstruction errors for in-class and out-class testing spectra, respectively. As mentioned above, the anomaly detector uses reconstruction errors as an indicator to determine anomalies. The vertical black line in each figure represents the decision threshold defined as the $95th$ percentile of the testing set reconstruction errors. As expected, out-class spectra are poorly reconstructed by the anomaly detector compared to in-class spectra: reconstruction errors for out-class spectra mostly fall to the right side of the threshold, while those for in-class spectra are much smaller. The results show that the hidden feature representations extracted by the encoder part are compact and descriptive enough to identify anomalies.

TABLE VI: Performance of the anomaly detector trained and tested on the same site.

	Average AUC	Average Recall	Average Precision
Teabag site 1	0.9612	0.9583	0.9490
Teabag site 2	0.9780	0.9580	0.9924
Teabag site 3	0.9710	0.942	1.0
Hazelnut site 1	0.9758	0.9516	1.0
Hazelnut site 2	0.9800	0.9600	1.0
Hazelnut site 3	0.9580	0.9660	0.9613

TABLE VII: Performance of the anomaly detector with different pre-processing pipelines

	Pre-processing pipeline	Average AUC	Average Recall	Average Precision
Teabag (site 1)	NIRWatchdog	**0.9612**	0.9583	**1.0**
	Pre-processing 1	0.9590	**0.9630**	0.9623
	Pre-processing 2	0.9611	0.9597	0.9871
Hazelnut (site 1)	NIRWatchdog	**0.9758**	0.9516	**1.0**
	Pre-processing 1	0.9480	0.9577	0.9133
	Pre-processing 2	0.9534	**0.9600**	0.9139

We then investigate the impacts of pre-processing pipelines on the performance of the anomaly detector. Table VII presents the average AUC, recall and precision obtained by applying the two alternative pre-processing pipelines as described in Section IV-B and the pipeline of the NIRWatchdog, in the tea bag and hazelnut cocoa spread experiments. In general, the performance of the NIRWatchdog pipeline is close to that of the other two pipelines in terms of AUC and recall, while the precision is relatively improved. This is particularly evident in hazelnut cocoa spread experiments: the precision of applying NIRWatchdog pipeline increases by around 7% compared with **Pre-processing 1** and **Pre-processing 2**, as the MSC and SNV tend to reduce the dissimilarity between types of samples, the anomaly detector is therefore more likely to classify outclass samples to in-class.

D. Synthetic spectra generation

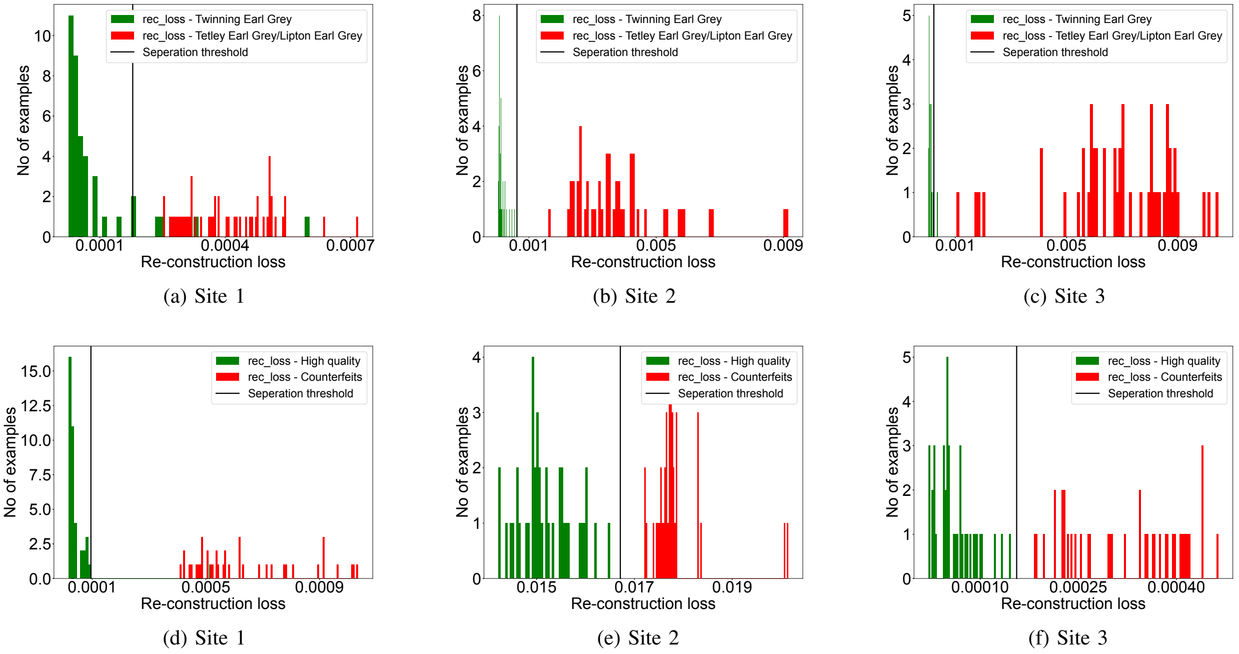
This section first evaluates the generation of spectral deviations using the synthetic spectra generator and then compares the similarity of synthetic spectra with the in-class spectra collected from the target environments.

The VAE is trained by minimizing a loss function defined as the sum of the reconstruction loss and regularization loss, weighted by two importance parameters λ_1 and λ_2. To determine them, λ_1 was empirically determined as 1 and 5, and experiments were conducted to test the performance under varying λ_2. By definition, the spectral deviations produced by the decoder of the VAE should exhibit a similar distribution to the deviations of the training dataset. The average EMD between them, as described in Section III, is applied to quantify the impact of the two hyper-parameters.

Table VIII summarizes the average EMD between the deviations of the testing dataset and the generated spectral deviations obtained from tea bags and hazelnut cocoa spread experiments in three sites. The results show that the best combination of the two parameters can be found at $\lambda_1 = 5$ and $\lambda_2 = 0.001$, in which average EMDs between the synthetic and real spectral deviations are around 0.3 to 0.4. This setting consistently helps the VAE achieve the best performance for most scenarios. In contrast, the extremely low or high settings of λ_1 and λ_2 tend to produce spectral deviations with low similarity. In the experiments below, the default values of λ_1 and λ_2 are 5 and 0.001 if not explicitly mentioned.

To investigate the effect of the number of reference samples on the quality of the generated synthetic in-class spectra, one site was defined as the training environment, and the rest were treated as the target environments. 1, 3, and 5 randomly selected in-class samples were used as reference samples to generate synthetic spectra of the target environment, respectively. The average EMD between the synthetic dataset and the in-class dataset of the target environment is expected to be small enough.

Figure 9 presents results from tea bags and hazelnut cocoa spread experiments. The legend **Site a to Site b** in the figure denotes the scenario where the VAE was trained on **Site-a** and the synthetic transfer dataset was generated for **Site-b**. Intuitively, increasing the number of reference samples reduces the average EMD, implying that synthetic spectra are generated from a similar distribution to those collected from the target environment. This is because the average spectra obtained from more reference samples are more likely to be closer to the dataset mean. However, even with only one reference sample, the average EMDs are still less than 0.85 for all cases. The results also show that the average EMD in the case of **Site-2 to Site-3** and **Site-3 to Site-2** are smaller than the others. This is because the data collection in **Site-2** and **Site-3** were conducted in the same season. Therefore, the variation of environmental conditions is less compared with the case of **Site-1**.



(a) Site 1 (b) Site 2 (c) Site 3

(d) Site 1 (e) Site 2 (f) Site 3

Fig. 8: Reconstruction errors of teabag experiments (a to c) and hazelnut cocoa spread experiments (d to f) when the training and testing datasets come from the same site.

TABLE VIII: Impacts of λ_1 and λ_2 on the quality of generated spectra deviations

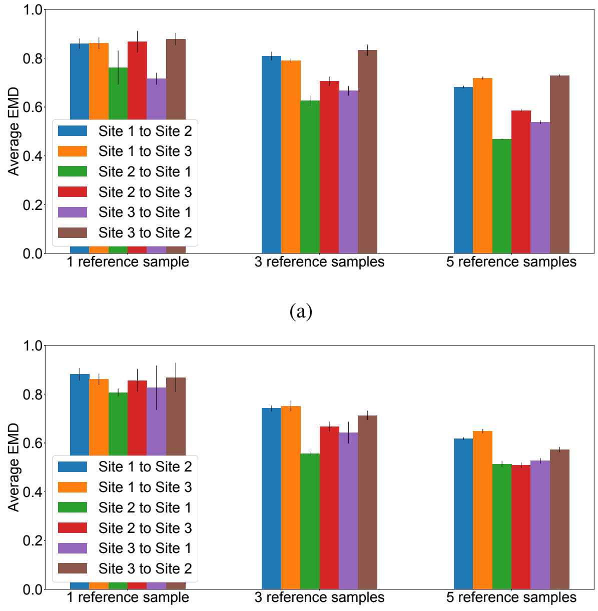
Tea Bags					Hazelnut cocoa spread				
λ_1	λ_2	**Site 1**	**Site 2**	**Site 3**	λ_1	λ_2	**Site 1**	**Site 2**	**Site 3**
5	1	0.9975	1.2914	1.3436	5	1	1.0052	1.3157	1.3841
	0.1	0.9990	0.7535	1.3439		0.1	0.9534	0.7891	0.8172
	0.01	0.5607	0.4774	0.4859		0.01	0.5281	0.5191	0.4696
	0.001	**0.2942**	**0.3441**	**0.3085**		0.001	0.3317	**0.4128**	0.3601
10	1	0.9948	1.3056	1.3555	10	1	1.0180	1.3333	1.3698
	0.1	1.0148	0.6847	0.7714		0.1	0.6010	0.7120	0.7943
	0.01	0.4233	0.4830	0.4982		0.01	0.4413	0.4108	0.4684
	0.001	0.3657	0.8112	0.3432		0.001	**0.2943**	0.4288	**0.3525**

Figure 10 further demonstrates the effectiveness of the proposed approach by visualizing the t-SNE embeddings of the in-class spectra collected from **Site-1** and **Site-3**, and the synthetic spectral dataset generated for the two sites, respectively. In this experiment, the synthetic spectra generator was trained on the spectra of tea bags collected from **Site-2**. Since the environmental changes between **Site-1** and **Site-2** are more considerable, it can be seen from Figure 10a that some of the embeddings of in-class spectra are grouped away from the embeddings of synthetic ones, yet, they remain close to each other. In the case of **Site-3** (i.e., Figure 10b), it is evident that the embeddings of the two datasets are mixed together as if they came from the same data distribution.

E. Adapt to new environments

This section demonstrates the cross-domain adaptability of the NIRWatchdog system. In this set of experiments, the anomaly detector trained from one site was fine-tuned using the generated synthetic transfer dataset and then tested by the real in-class and out-class spectra collected from the other two sites. Note that the synthetic spectra generator can generate any number of in-class spectra as if they were collected from the target environment. The experimental results are obtained using 50 synthetic spectra for model fine-tuning.

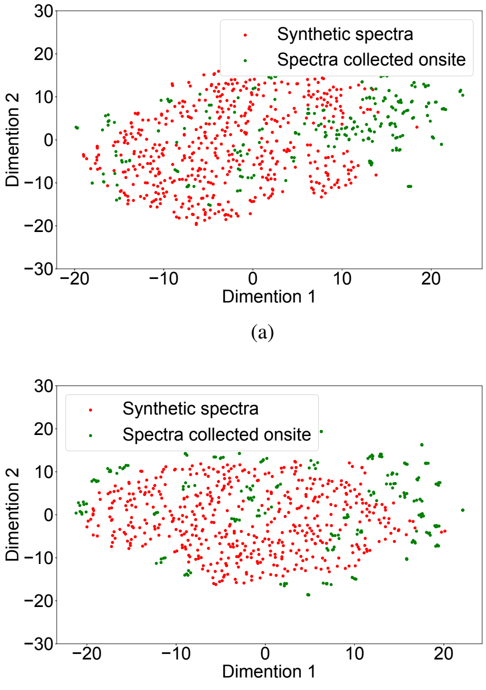
Figure 11 shows the reconstruction errors of in-class and out-class spectra obtained from a tea bag experiment before and after model fine-tuning, where the anomaly detector and the synthetic spectra generator were trained using the in-class spectra of **Site-1**. As can be seen from 11a and 11b, changes in the operating environment cause complete failures of the pre-trained anomaly detectors: the reconstruction errors fall to the right side of the threshold obtained from the training environment no matter they are in-class or out-class spectra. As explained in Section III, the reason behind this is the onsite collected spectra are no longer independent and identically distributed as the training spectra. Hence, all samples are considered as anomalies indiscriminately. It is worth noting that the performance of the pre-trained models cannot be improved by simply adjusting their thresholds: the reconstruction errors of the in-class spectra are larger than that of the out-class ones for the experiment at **Site-2** and **Site-3**. After fine-tuning the model using the synthetic transfer dataset, the thresholds obtained through pre-training can correctly separate most of the input spectra again. It is evident from



(a)

(b)

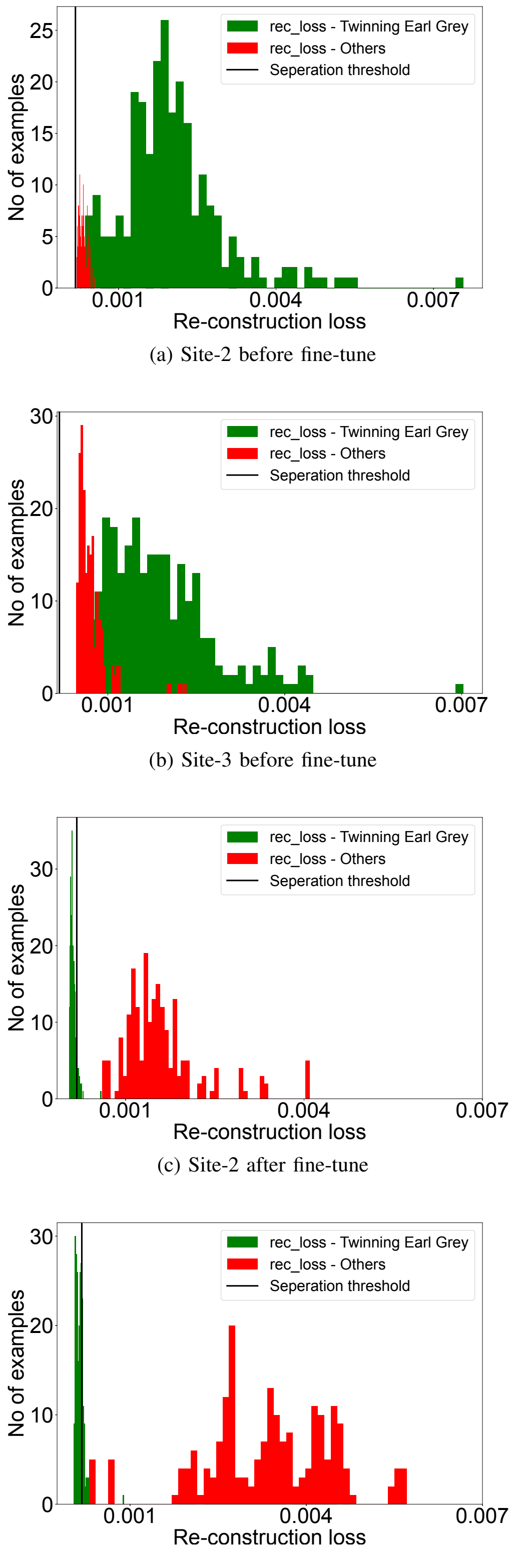
Fig. 9: The effect of the number of reference samples on the quality of generated synthetic transfer dataset. (a) Tea bags and (b) Hazelnut cocoa spread. Increasing the number of reference samples reduces the average EMD, implying that synthetic spectra are generated from a similar distribution to those collected from the target environment.



(a)

(b)

Fig. 10: The t-SNE embeddings of synthetic spectra and the spectra collected from (a) **Site-1** and (b) **Site-3**. The spectra generator is trained on data collected from **Site-2**. The synthetic spectra and onsite spectra are projected in proximity region as if they came from the same data distribution.



(a) Site-2 before fine-tune

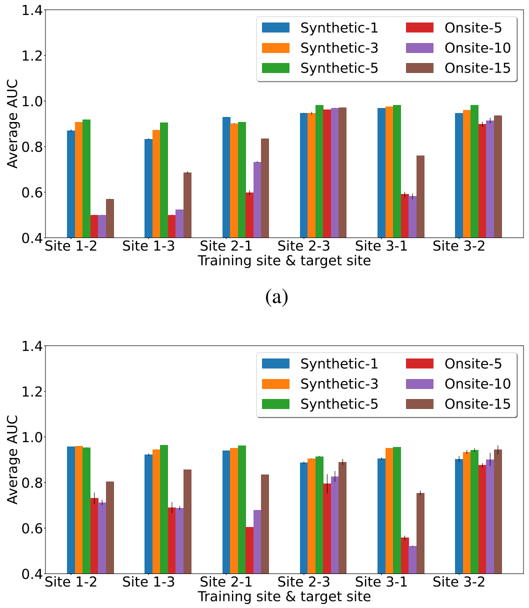
(b) Site-3 before fine-tune

(c) Site-2 after fine-tune

(d) Site-3 after fine-tune

Fig. 11: Reconstruction losses of teabags before (a and b) and after (c and d) model fine-tuning. The base anomaly detector and synthetic spectra generator were trained on **Site-1**.

Figure 11c and 11d that the reconstruction errors of in-class spectra are pushed back to the left side while those of the out-class spectra stay in the same region. The results demonstrate that our proposed approach can effectively address the cross-environmental adaptability issue of miniaturized near-infrared spectrometers.



(a)

(b)

Fig. 12: Comparison between model fine-tuning using synthetic transfer dataset and onsite collected transfer dataset. (a) Tea bags, and (b) Hazelnut cocoa spread

To further demonstrate the practicality and effectiveness of the proposed approach, another set of experiments was conducted to evaluate the performance of the conventional approach, in which the pre-trained model was fine-tuned using real in-class spectra collected from the target environment. Those experiments used the same pre-trained anomaly detector and identical hyper-parameters for model fine-tuning. The spectra of 5, 10, and 15 in-class samples collected from the target sites was used as the transfer dataset, respectively. Figure 12 compares the average AUCs obtained from the NIRWatchdog and the conventional approach, in which the legend **Synthetic x** denotes the case that the pre-trained anomaly detector is fine-tuned by using a synthetic dataset generated based on x reference samples and **Onsite x** indicates that the model is fine-tuned using spectra collected from x reference samples onsite.

According to the experiment results, with only one reference sample, the average AUCs of NIRWatchdog are higher than 85% in all cases. When the number of reference samples increases to 3, the AUCs can be further improved to above 90%, close to 1 in some scenarios. However, without using the synthetic spectra generator, the models must be fine-tuned using onsite spectra collected from many more reference samples. In the scenarios of **Site-2 to Site-3** and **Site-3 to Site-2**, the conventional approach requires 10 to 15 reference samples to achieve a comparable performance of NIRWatchdog with 3 reference samples. In other scenarios, the average AUCs of 15 reference samples are much lower than the performance of the NIRWatchdog system. This is because the data collection for **Site-2** and **Site-3** is carried out in the same season, and the distribution of their in-class spectra is closer than that for **Site-1**. In other words, the number of reference samples required by the conventional approach depends on the difference between the target environment and the training environment: the greater the difference between the two environments, the more the number of reference samples required for model fine-tuning. On the other hand, our proposed approach can significantly reduce the number of reference samples required for model fine-tuning and achieves a stable performance over different scenarios. The synthetic spectra generator can augment the transfer dataset collected from as few as one reference sample according to a similar distribution as the real in-class spectra on the target environment, thereby reducing the cost of reference sample storage and transportation and making model fine-tuning faster and more efficient.

F. Experiments on additional commodities

To thoroughly validate the proposed NIRWatchdog, this section presents the results obtained from the cross-domain experiments on eight additional types of commodities. As shown in Figure 13, the commodities were selected based on their varying characteristics and compositions, which included coffee pods, two types of effervescent tablets, and five types of out-of-counter pharmaceuticals. Samples were gathered from local shops and pharmacies with 40 units per commodity. We scanned each sample using the NIR-S-G1 spectrometer under two different environments as described in Table IX, with the same spectrometer parameters as the tea bag and hazelnut cocoa spread experiments. Similarly to the above setup, each sample was placed on top of the lens and scanned 10 times. For each commodity, the anomaly detector and the synthetic spectra generator were trained using 80% of the in-class spectra collected from **Site-A**. The network architectures and the training parameters remain the same as in the tea bag and hazelnut cocoa spread experiments. Using the synthetic transfer dataset, the trained anomaly detector is then fine-tuned to adapt **Site-B**.

TABLE IX: Environment parameters of the data collections

	Temperature	**Humidity**	**Illumination**
Coffee Pods	A: 16.3C B: 29.2C	A: 83% B: 52%	A: 426 Lux B: 537 Lux
Magnesium ET	A: 24.3C B: 12.6C	A: 42% B: 88%	A: 368 Lux B: 382 Lux
Vitamin C ET	A: 24.3C B: 12.6C	A: 42% B: 88%	A: 368 Lux B: 382 Lux
Paracetamol	A: 12.7C B: 23.8C	A: 62% B: 37%	A: 402 Lux B: 372 Lux
Panadol	A: 12.7C B: 23.8C	A: 62% B: 37%	A: 402 Lux B: 372 Lux
Ibuprofen EG	A: 12.7C B: 23.8C	A: 62% B: 37%	A: 402 Lux B: 372 Lux
Aspirin	A: 11.3C B: 23.8C	A: 65% B: 37%	A: 322 Lux B: 372 Lux
Magnesium B6	A: 23.8C B: 11.3C	A: 37% B: 65%	A: 372 Lux B: 322 Lux

Unlike the tea bag experiments that evaluated products from another brand, this set of experiments regarded contaminated

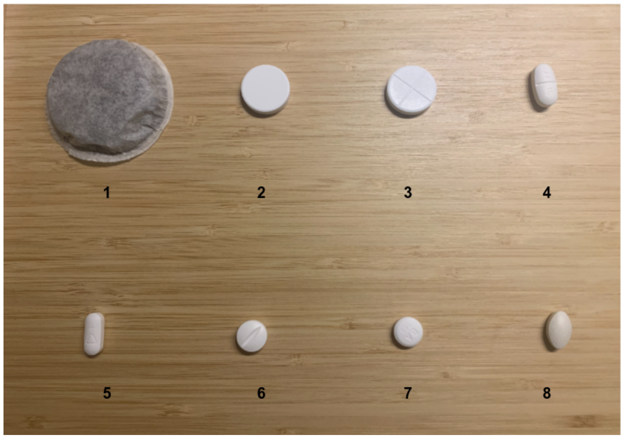


Fig. 13: The selected commodities for the experiments: 1) Coffee pods, 2) Magnesium effervescent tablets, 3) Vitamin C effervescent tablets, 4) Paracetamol, 5) Panadol, 6) IbuprofenEG, 7) Aspirin, 8) Magnesium B6.

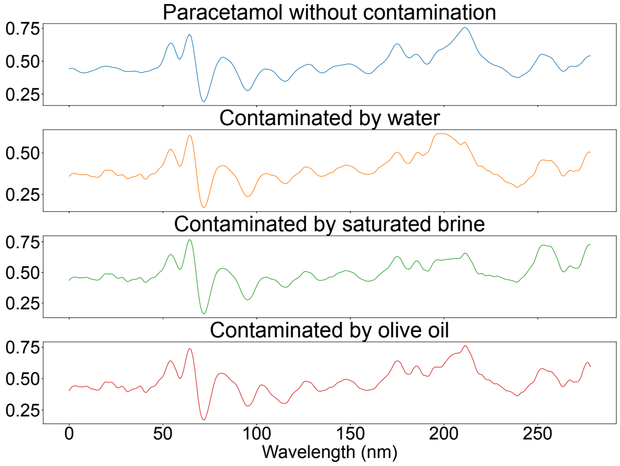


Fig. 14: The average spectra of paracetamol tablets with and without contamination.

samples as out-class. The contamination was introduced by using an eye dropper to add one drop of the solutions (water, saturated brine, or olive oil) to the normal samples. The amount of solution for each drop is around $0.04ml$ to $0.05ml$. For each commodity, 9 samples were contaminated: 3 with water, 3 with saturated brine, and 3 with olive oil. Figure 14 compares the average spectra of paracetamol tablets with and without contamination. It can be seen that dropping a small amount of solution compromises the quality of the tablets, leading to distortion in different wavelength regions of the spectra. The obtained dataset, including spectra of normal samples and contaminated samples, has been uploaded to the IEEE DataPort with the DoI of 10.21227/9c2g-dj30.

TABLE X: The EDM of spectra and spectra deviations between samples scanned on two sites

Coffee Pods	**Magnesium ET**	**Vitamin C ET**	**Paracetamol**
2.742/0.952	2.2311/0.5613	3.5008/0.5273	1.9718/0.2708
Panadol	**IbuprofenEG**	**Aspirin**	**Magnesium B6**
1.8665/0.4247	2.4449/0.3585	4.3649/0.6410	1.6186/0.2250

TABLE XI: The performance of anomaly detectors trained and tested on the data collected from the same site.

Commodity	AUC	Recall	Precision
Coffee Pods	0.970	0.934	1.0
Magnesium ET	0.984	0.968	1.0
Vitamin C ET	0.965	0.931	1.0
Paracetamol	0.971	0.942	1.0
Panadol	0.978	0.956	1.0
IbuprofenEG	0.975	0.945	1.0
Aspirin	0.971	0,944	1.0
Magnesium B6	0.990	0.981	1.0

TABLE XII: The cross-domain performance of the anomaly detector. The models are trained on in-class spectra collected from site A and tested on the spectra collected from site B.

Commodity	Metric	Before FT	Syn-1	Syn-3	Syn-5
Coffee Pods	AUC	0.5	0.919	915	963
	Recall	-	0.839	830	0.935
	Precision	-	2	3	4
Magnesium ET	AUC	0.5	0.945	0.986	0.991
	Recall	-	0.889	0.972	0.983
	Precision	-	1.0	1.0	1.0
Vitamin C ET	AUC	0.5	0.902	0.912	0.948
	Recall	-	0.804	0.825	0.896
	Precision	-	1.0	1.0	1.0
Paraceta-mol	AUC	0.625	0.856	0.869	0.912
	Recall	0.253	0.883	0.929	0.948
	Precision	1.0	0.896	0.885	0.901
Panadol	AUC	0.5	0.907	0.923	0.964
	Recall	-	0.814	0.845	0.928
	Precision	-	0.998	1.0	1.0
Ibuprofen EG	AUC	0.5	0.869	0.915	0.934
	Recall	-	0.731	0.831	0.869
	Precision	-	1.0	1.0	1.0
Aspirin	AUC	0.5	0.922	0.936	0.948
	Recall	-	0.845	0.872	0.896
	Precision	-	1.0	1.0	1.0
Magnesium B6	AUC	0.631	0.922	0.953	0.963
	Recall	0.261	0.845	0.907	0.927
	Precision	1.0	1.0	1.0	1.0

The average EMD of the spectra and spectra deviations of the 8 commodities between two sites are presented in Table X. As expected, the changes in scanning environments cause spectral distortion. Magnesium B6, panadol, and paracetamol tablets are less prone to the impact of environments, while Vitamin C and aspirin exhibit higher inter-site variation. Similar to the observations obtained from tea bags and hazelnut cocoa spread experiments, the distribution shift of the spectra deviations between the two sites is small, as confirmed by the obtained EMD of spectra deviations.

Table XI presents the average AUC, recall, and precision obtained from the same-site experiments (e.g., the anomaly detectors are trained and tested on the data collected from Site-A) of the 8 commodities. Although a very small number of in-class samples are classified as out-of-class, the autoencoder-based detector can effectively capture the quality changes caused by contamination and separate them from the in-class samples. Table XII summarizes the results obtained from the cross-site experiments. Due to the distribution shift caused by the changes in scanning environments, the pre-trained anomaly detectors suffer varying levels of performance drop when applied on different sites. Except for paracetamol and magnesium b6, the models trained for the remaining 6 commodities indiscriminately classify all samples as out-class.

Based on the results, it is clear that the proposed NIRWatchdog is effective in enhancing cross-site performance, even when there is only one in-class sample available onsite. In the case of magnesium effervescent tablets, aspirin, and magnesium b6, with three in-class samples onsite, the performance of the anomaly detector can be recovered to match the level of performance that would have been achieved in the same-site experiments.

VI. Conclusion

This paper focuses on the application of onsite product quality assessment using miniaturized NIR sensors and proposes NIRWatchdog, a system to enable fast model adaptation to new and unseen operating environments. Similar to the state-of-the-art, NIRWatchdog fine-tunes the pre-trained model at the parametric level using a transfer dataset. However, our proposed approach only needs as few as one reference sample onsite, effectively reducing the cost and human involvement of data collection and improving the feasibility of the technology on onsite tasks. The preliminary study shows that even for the same batch of samples, changes in the external factors of the scanning environment may significantly alter the distribution of their spectral data, however, the distribution of the spectral deviations remains sufficiently similar. Based on this observation, the NIRWatchdog employs a variational autoencoder to learn the data generation process of the spectral deviations from the training dataset. The generated spectral deviations, added with the average of a few onsite reference spectra, can be used to fine-tune the pre-trained model in new and unseen environments. Extensive experiments using tea bags and hazelnut cocoa spread over three distinct operating environments were conducted to demonstrate the effectiveness and efficiency of NIRWatchdog. The results show that the generated synthetic transfer dataset exhibits similar data distribution as the in-class spectra collected from the target environment. They can effectively boost the pre-trained models' performance in new environments. With only one reference sample, the average AUCs of NIRWatchdog are higher than 85% in all cases. In comparison, the baseline approach requires 10 to 15 reference samples to achieve a comparable performance of NIRWatchdog with 3 reference samples. The experiments conducted on additional 8 types of commodities also draw the same conclusion showing that the proposed approach is scalable for other types of samples.

References

[1] S. Klakegg, C. Luo, J. Goncalves, S. Hosio, and V. Kostakos, "Instrumenting smartphones with portable nirs," in *Proceedings of the 2016 International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 618–623.

[2] S. Klakegg, J. Goncalves, C. Luo, A. Visuri, A. Popov, N. van Berkel, Z. Sarsenbayeva, V. Kostakos, S. Hosio, S. Savage, *et al.*, "Assisted medication management in elderly care using miniaturised near-infrared spectroscopy," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, pp. 1–24, 2018.

[3] W. Jiang, G. Marini, N. van Berkel, Z. Sarsenbayeva, C. Luo, X. He, T. Dingler, Y. Kawahara, and V. Kostakos, "A mobile scanner for probing liquid samples in everyday settings," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 1172–1177.

[4] W. Jiang, G. Marini, N. van Berkel, Z. Sarsenbayeva, Z. Tan, C. Luo, X. He, T. Dingler, J. Goncalves, Y. Kawahara, *et al.*, "Probing sucrose contents in everyday drinks using miniaturized near-infrared spectroscopy scanners," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 4, pp. 1–25, 2019.

[5] X. Xu, L. Xie, and Y. Ying, "Factors influencing near infrared spectroscopy analysis of agro-products: a review," *Frontiers of Agricultural Science and Engineering*, vol. 6, no. 2, pp. 105–115, 2019.

[6] W. Wang, M. D. Keller, T. Baughman, and B. K. Wilson, "Evaluating low-cost optical spectrometers for the detection of simulated substandard and falsified medicines," *Applied Spectroscopy*, vol. 74, no. 3, pp. 323–333, 2020.

[7] Y. Yao, H. Chen, L. Xie, and X. Rao, "Assessing the temperature influence on the soluble solids content of watermelon juice as measured by visible and near-infrared spectroscopy and chemometrics," *Journal of food engineering*, vol. 119, no. 1, pp. 22–27, 2013.

[8] C. J. Hayes, C. V. Greensill, and K. B. Walsh, "Temporal and environmental sensitivity of a photodiode array spectrophometric system," *Journal of Near Infrared Spectroscopy*, vol. 22, no. 4, pp. 297–304, 2014.

[9] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[10] J. J. Workman Jr, "A review of calibration transfer practices and instrument differences in spectroscopy," *Applied spectroscopy*, vol. 72, no. 3, pp. 340–365, 2018.

[11] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[12] R. N. Feudale, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, and J. Ferré, "Transfer of multivariate calibration models: a review," *Chemometrics and intelligent laboratory systems*, vol. 64, no. 2, pp. 181–192, 2002.

[13] D.-W. Sun, *Infrared spectroscopy for food quality analysis and control*. Academic press, 2009.

[14] W. Fan, Y. Liang, D. Yuan, and J. Wang, "Calibration model transfer for near-infrared spectra based on canonical correlation analysis," *Analytica chimica acta*, vol. 623, no. 1, pp. 22–29, 2008.

[15] X. Li, Z. Li, X. Yang, and Y. He, "Boosting the generalization ability of vis-nir-spectroscopy-based regression models through dimension reduction and transfer learning," *Computers and Electronics in Agriculture*, vol. 186, p. 106157, 2021.

[16] L. Liu, M. Ji, and M. Buchroithner, "Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery," *Sensors*, vol. 18, no. 9, p. 3169, 2018.

[17] P. Mishra and D. Passos, "Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios," *Chemometrics and Intelligent Laboratory Systems*, vol. 212, p. 104283, 2021.

[18] Y. Yu, J. Huang, J. Zhu, and S. Liang, "An accurate noninvasive blood glucose measurement system using portable near-infrared spectrometer and transfer learning framework," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3506–3519, 2020.

[19] J. Padarian, B. Minasny, and A. McBratney, "Transfer learning to localise a continental soil vis-nir calibration model," *Geoderma*, vol. 340, pp. 279–288, 2019.

[20] D. A. Burns and E. W. Ciurczak, *Handbook of near-infrared analysis*. CRC press, 2007.

[21] L. Feng, S. Zhu, L. Zhou, Y. Zhao, Y. Bao, C. Zhang, and Y. He, "Detection of subtle bruises on winter jujube using hyperspectral imaging with pixel-wise deep learning method," *IEEE access*, vol. 7, pp. 64 494–64 505, 2019.

[22] G. Mu, T. Liu, C. Xue, and J. Chen, "Semi-supervised learning-based calibration model building of nir spectroscopy for in situ measurement of biochemical processes under insufficiently and inaccurately labeled samples," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2020.

[23] X. Zhang, J. Yang, T. Lin, and Y. Ying, "Food and agro-product quality evaluation based on spectroscopy and deep learning: A review," *Trends in Food Science & Technology*, vol. 112, pp. 431–441, 2021.

[24] L. Zou, X. Yu, M. Li, M. Lei, and H. Yu, "Nondestructive identification of coal and gangue via near-infrared spectroscopy based on improved broad learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 8043–8052, 2020.

[25] L. da Silva Dias, J. C. da Silva Junior, A. L. d. S. M. Felício, and J. A. de França, "A nir photometer prototype with integrating sphere

for the detection of added water in raw milk," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 12, pp. 2812–2819, 2018.

[26] H. Hu, Q. Zhang, and Y. Chen, "Nirscam: A mobile near-infrared sensing system for food calorie estimation," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 18 934–18 945, 2022.

[27] S. Nawar and A. Mouazen, "On-line vis-nir spectroscopy prediction of soil organic carbon using machine learning," *Soil and Tillage Research*, vol. 190, pp. 120–127, 2019.

[28] L. Wang, M. Izzetoglu, J. Du, and H. Ayaz, "Phantom and model-based near infrared spectroscopy measurements of intracranial hematoma from infants to adults," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–9, 2022.

[29] T. Isaksson and B. Kowalski, "Piece-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products," *Applied spectroscopy*, vol. 47, no. 6, pp. 702–709, 1993.

[30] P. De Groot, H. Swierenga, G. Postma, W. Melssen, and L. Buydens, "Effect on the partial least-squares prediction of yarn properties combining raman and infrared measurements and applying wavelength selection," *Applied spectroscopy*, vol. 57, no. 6, pp. 642–648, 2003.

[31] E. Bouveresse and D. Massart, "Improvement of the piecewise direct standardisation procedure for the transfer of nir spectra for multivariate calibration," *Chemometrics and intelligent laboratory systems*, vol. 32, no. 2, pp. 201–213, 1996.

[32] M. Martinez, M. Tapaswi, and R. Stiefelhagen, "A closed-form gradient for the 1d earth mover's distance for spectral deep learning on biological data," in *ICML 2016 Workshop on Computational Biology (CompBio@ ICML16)*, 2016.

[33] G. Ibáñez, J. Cebolla-Cornejo, R. Martí, S. Roselló, and M. Valcárcel, "Non-destructive determination of taste-related compounds in tomato using nir spectra," *Journal of Food Engineering*, vol. 263, pp. 237–242, 2019.

[34] Y. Huang, K. Chen, L. Wang, Y. Dong, Q. Huang, and K. Wu, "Lili: liquor quality monitoring based on light signals," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 256–268.

[35] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. Buydens, and E. Marchiori, "Convolutional neural networks for vibrational spectroscopic data analysis," *Analytica chimica acta*, vol. 954, pp. 22–31, 2017.

[36] C. Cui and T. Fearn, "Modern practical convolutional neural networks for multivariate regression: Applications to nir calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 182, pp. 9–20, 2018.

[37] C. Zhang, W. Wu, L. Zhou, H. Cheng, X. Ye, and Y. He, "Developing deep learning based regression approaches for determination of chemical compositions in dry black goji berries (lycium ruthenicum murr.) using near-infrared hyperspectral imaging," *Food chemistry*, vol. 319, p. 126536, 2020.

[38] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2898–2906.

[39] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[40] P. Oza and V. M. Patel, "One-class convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 277–281, 2018.

[41] P. Perera and V. M. Patel, "Learning deep features for one-class classification," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5450–5463, 2019.