


# Pharmaceutical Raw Material Identification Using Miniature Near-Infrared (MicroNIR) Spectroscopy and Supervised Pattern Recognition Using Support Vector Machine

Lan Sun, Chang Hsiung, Christopher G. Pederson, Peng Zou, Valton Smith, Marc von Gunten, and Nada A. O'Brien

Applied Spectroscopy  
2016, Vol. 70(5) 816–825  
© The Author(s) 2016  
Reprints and permissions:  
sagepub.co.uk/journalsPermissions.nav  
DOI: 10.1177/0003702816638281  
asp.sagepub.com  


## Abstract

Near-infrared spectroscopy as a rapid and non-destructive analytical technique offers great advantages for pharmaceutical raw material identification (RMID) to fulfill the quality and safety requirements in pharmaceutical industry. In this study, we demonstrated the use of portable miniature near-infrared (MicroNIR) spectrometers for NIR-based pharmaceutical RMID and solved two challenges in this area, model transferability and large-scale classification, with the aid of support vector machine (SVM) modeling. We used a set of 19 pharmaceutical compounds including various active pharmaceutical ingredients (APIs) and excipients and six MicroNIR spectrometers to test model transferability. For the test of large-scale classification, we used another set of 253 pharmaceutical compounds comprised of both chemically and physically different APIs and excipients. We compared SVM with conventional chemometric modeling techniques, including soft independent modeling of class analogy, partial least squares discriminant analysis, linear discriminant analysis, and quadratic discriminant analysis. Support vector machine modeling using a linear kernel, especially when combined with a hierarchical scheme, exhibited excellent performance in both model transferability and large-scale classification. Hence, ultra-compact, portable and robust MicroNIR spectrometers coupled with SVM modeling can make on-site and in situ pharmaceutical RMID for large-volume applications highly achievable.

## Keywords

Near-infrared spectroscopy, NIR, MicroNIR, support vector machine, SVM, model transferability, large-scale classification, raw material identification, RMID

Date received: 28 August 2015; accepted: 13 December 2015

## Introduction

Raw material identification (RMID) or verification of the packaging label is a common quality-control practice in the pharmaceutical industry. The increasing global footprint of the supply chain and public health concerns resulting from contaminated or mislabeled materials have driven many regulatory bodies to require inspection of every barrel in every shipment of materials used in pharmaceutical drugs. Traditionally, pharmaceutical RMID has relied on laboratory-based analytical techniques such as chromatography, wet chemistry, and titrations among others. Most of these techniques are destructive in nature, time consuming and labor intensive, and hence it is challenging to handle an enormous number of analyses.<sup>1</sup>

Vibrational spectroscopy, including near-infrared (NIR), mid-infrared (mid-IR), and Raman spectroscopy, has gained wide acceptance in the pharmaceutical industry for RMID in recent years due to its non-destructive nature, minimal sample preparation, and fast data acquisition. Especially, with substantial progress in portable NIR, mid-IR, and Raman spectrometers, on-site and in situ analysis of a large number of samples has become practical for material

---

Viavi Solutions Inc. (formerly JDSU), Santa Rosa, CA, USA

### Corresponding author:

Lan Sun, Viavi Solutions Inc. (formerly JDSU), 1402 Mariner Way, Santa Rosa, CA 95407, USA.  
Email: lan.sun@viavisolutions.com

identification, which opens up more application opportunities.<sup>2</sup>

Among the three vibrational spectroscopic techniques, NIR and IR measure absorbance, while Raman measures scattering. NIR and IR are sensitive to the change in the dipole moment of a vibrating molecule, while Raman is sensitive to the change in the polarizability of a vibrating molecule. Mid-infrared is less popular in RMID than NIR due to the strong absorption coefficient in the mid-IR spectral range, which limits the path length into the samples and sometimes requires dilution of the samples using infrared transparent materials.<sup>3</sup> In general, NIR and Raman are complementary in nature. Both techniques have found broad applications in pharmaceutical analysis,<sup>4,5</sup> but have their own advantages and disadvantages.<sup>6</sup> Raman spectroscopy has outstanding molecular selectivity, can be easily used in a non-contact fashion through common container materials, and is free of water interference from aqueous solutions. However, interference from fluorescent molecules can be a limitation, and the high energetic laser power may decompose sensitive samples. Conversely, NIR spectroscopy does not suffer from the fluorescence problem and can also measure through plastic or glass containers. The limiting factor of NIR is the complexity of the spectra, thus low molecular selectivity, resulting from vibrational overtones and combination bands, which require the use of multivariate data analysis. Over the past decade, the computing power and algorithms have improved dramatically allowing NIR to become more powerful and user friendly. In this work, we chose NIR as the analytical tool for pharmaceutical RMID.

Near-infrared techniques have generally been adopted by major pharmacopoeias. The United States Pharmacopoeia (Chapter 1119)<sup>7</sup> and the European Pharmacopoeia (Chapter 2.2.40)<sup>8</sup> have addressed the suitability of NIR instrumentation for application in pharmaceutical testing. Luypaert et al. reviewed a wide range of NIR applications for pharmaceutical material identification,<sup>4</sup> such as identifying commonly used excipients and active pharmaceutical ingredients (API),<sup>9,10</sup> distinguishing between closely related substances,<sup>11,12</sup> and classifying different polymorphic forms of the same product.<sup>13,14</sup> More recently, Grout incorporated NIR material qualification outputs with statistical process control (SPC) charts (through historical trending) to link material attributes to both product quality and process behavior, which enables rapid material qualification on receipt with better understanding of process performance.<sup>15</sup> Moreover, in the last couple of years, miniaturized NIR spectrometers became commercially available. Their performance for pharmaceutical applications has been successfully demonstrated,<sup>2,16</sup> which facilitates more practical and flexible use of NIR spectrometers for on-site material identification.

Near-infrared-based pharmaceutical RMID requires the aid of chemometric tools for classification due to the

complex nature of NIR spectra, which is essentially a pattern recognition problem. The pattern recognition techniques can be divided into two categories, the supervised and the unsupervised methods, with the former being the most common for pharmaceutical applications.<sup>17</sup> In a classification study, unsupervised principal component analysis (PCA) is often used first to show discriminating tendencies and then more effective supervised analyses are applied for enhanced discrimination.<sup>18</sup> Correlation based methods, distance based methods, linear discriminant analysis (LDA), soft independent modeling of class analogy (SIMCA), and partial least squares discriminant analysis (PLS-DA) are classical methods for the supervised classification.<sup>17</sup> For example, Blanco and Romero constructed a library including NIR spectra for 125 different raw materials using the correlation coefficient as the discriminating criterion.<sup>9</sup> Dreassi et al. utilized LDA to distinguish pharmaceutical compounds with different physical properties.<sup>19</sup> Krämer and Ebel demonstrated the discrimination of powdered and microcrystalline celluloses as well as cellulose and cellulose ethers by SIMCA.<sup>12</sup> Andre assessed chemical quality of 7-aminocephalosporanic acid from a large number of production lots using PLS-DA.<sup>20</sup> Some additional methods were also discussed in the review article by Roggo et al.<sup>17</sup>

Despite past successful applications of NIR for pharmaceutical RMID, there are still challenges that need to be addressed. One of the challenges is model transferability, i.e., transferring the chemometric model from one or more master instruments to multiple target instruments. Multiple instruments are often needed to fulfill the tasks of RMID on a regular basis. Built-in differences and changes induced by wear and varying environments in the instrument response function can render a model established on one instrument invalid on another. However, it is often not practical and/or economical to develop models for individual instruments. Thus, transferable robust models that require the least number of master instruments are highly desirable for RMID. Another challenge of NIR-based RMID is dealing with a large library of NIR spectra when the total number of classes reaches hundreds, which is not uncommon for RMID, since there are a wide range of APIs and excipients with different physical properties from different manufacturers and different lots involved. In this case, some conventional pattern classifiers suffer from the resolution issue that the chemometric models' discrimination power is diluted by the increased number of classes and thus not able to distinguish smaller differences among them.

In recent years, the support vector machine (SVM), originally popular in the neural networks and machine learning community, has been introduced to chemometrics and proven to be powerful in NIR spectra classification.<sup>21,22</sup> However, this classifier has been less studied in NIR-based pharmaceutical RMID. In this work, we successfully addressed the challenges of model transferability and large-scale classification in NIR-based RMID by the use of SVM.

Moreover, miniature MicroNIR spectrometers were used for all of the measurements. The MicroNIR spectrometer is designed to use a novel thin-film linear variable filter (LVF) as the dispersive element on top of an InGaAs array detector. The filter coating is physically tapered with position, resulting in a continuous change in the center wavelength of the filter with position. This ultra-compact spectrometer offers high-speed measurement, ruggedness, stability, portability, and low power consumption, and has been successfully used in several applications.<sup>16,23–26</sup> The use of MicroNIR spectrometers with excellent performance consistency and the powerful SVM algorithm makes rapid and reliable on-site RMID highly achievable.

## Experimental Details

### Materials

Two sets of materials were used in this work. The first set consisted of 19 of the most commonly used APIs and excipients, including acetaminophen, ascorbic acid, aspirin, benzocaine, caffeine, cellulose, corn starch, fructose, hydroxypropyl cellulose (HPC), (hydroxypropyl)methyl cellulose (HPMC), ibuprofen, lactose, magnesium stearate (Mg-stearate), poly(ethylene oxide) (PEO), polyvinylpyrrolidone (PVP), polysorbate 80, sodium starch glycolate (SSG), talc, and titanium dioxide (TiO<sub>2</sub>). For each compound, we purchased two to three different products, which were different in grades and/or physical properties, or were from different manufacturers. We divided the samples into lots A, B, and C. This set of samples was used to study model transferability. The second set consisted of 253 commonly used APIs and excipients that were provided by one of our collaborators in the pharmaceutical industry (the full list of materials is available upon request). For each compound, multiple samples were collected to include natural variability. This set of samples was used to investigate large-scale classification.

### Spectra Collection

Miniature MicroNIR Pro 1700 spectrometers developed and commercialized by Viavi Solutions Inc. (formerly JDSU, Santa Rosa, CA) were used for all data acquisition. All spectra were collected using MicroNIR Pro spectrometer software version 2.1 (Viavi Solutions Inc.) in the range of 908–1676 nm. A reference measurement was performed on the MicroNIR approximately 15 min after the lamps were turned on and every hour thereafter while performing scans. A 99% diffuse reflectance panel was used for the 100% reference value and the 0% reference value was taken by leaving the tungsten lamps on with an empty vial holder. This scenario was used to account for any scattered light from the sample vial holder.

The spectra of the pharmaceutical materials were collected in ambient conditions. As shown in Figure 1,



**Figure 1.** MicroNIR spectrometer equipped with a vial holder and tethered to a rugged 7" Windows 8.1 tablet for pharmaceutical raw material identification.

all samples were presented to the spectrometer housed in 14 mm diameter borosilicate glass vials with measurements performed through the bottom of the vials. A vial holder that slips over the end of the spectrometer was used by the operator to easily introduce and remove samples while maintaining an optimal 3 mm distance between the material and the sapphire window of the spectrometer. The vial was rotated approximately 10–15° after every scan. Each scan had an integration time of 10 ms with spectrum averaged over 50 collections.

For the set of 19 compounds, three vials were prepared for every sample in lots A, B, and C, respectively. Data acquisition was performed at three different dates and different times of the day using different vials of the same sample to take into account varying ambient temperature. A minimum of five spectra were collected for each vial. To study the model transferability, six spectrometers were used to collect data from the set of 19 pharmaceutical compounds. For the set of 253 compounds, a minimum of 20 spectra were collected from multiple samples for each compound. Figure 1 also illustrates the portability of the device for on-site and in situ pharmaceutical RMID with spectra acquisition by the ultra-compact MicroNIR spectrometer tethered to a rugged seven-inch Windows 8.1 tablet.

### Spectral Pretreatment and Chemometric Analysis

All of the steps of spectral processing and chemometric analysis were performed using Matlab (The MathWorks, Inc.). All of the spectra collected were pretreated using





spectrometer for training to achieve perfect prediction, and the SVM-rbf model needed two. In contrast, five or six spectrometers were needed to achieve perfect prediction for the other models. These results indicate that SVM algorithms outperformed all the other algorithms in terms of model transferability.

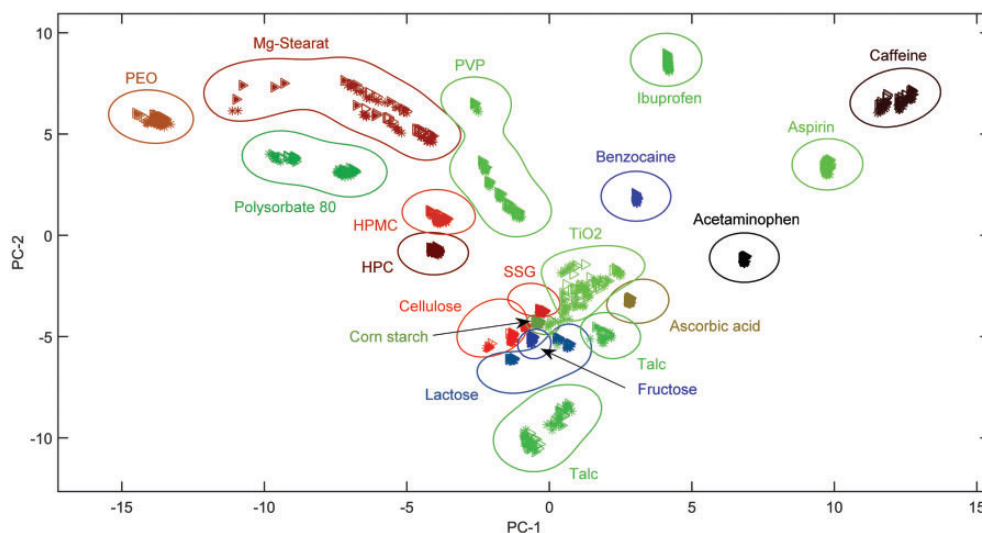
As an example, the training set acquired by one spectrometer was used to predict the test set acquired by another spectrometer implementing the SVM model. The clear partitions of classes and successful predictions are visualized by the PCA-SVM plot in Figure 2, where the data collected were plotted against PC1 and PC2 and the class boundaries were generated by SVM calculations. The filled triangles in each color-coded class represent the support vectors that constrain the width of margin between different classes and the open triangles represent the rest of the training data. The color-coded stars represent the predicted data with the color denoting the true identity of the test sample. The prediction success rate was 100% for the data shown in Figure 2. The few TiO<sub>2</sub> data not perfectly enclosed by the boundary need a higher dimensional space to visualize the classification, since the data structure of the spectra has 121 dimensions and the PCA-SVM plot only shows two.

It should be noted that for the library of 19 pharmaceutical compounds, with PLS-DA and SIMCA, prediction accuracy of >98.8% was achieved with only two instruments for training. This excellent instrument-to-instrument method transferability can be attributed to the consistent performance of MicroNIR spectrometers spectrally, optically and physically, which was enabled by the wavelength calibration reproducibility and photometric stability of the MicroNIR spectrometers.<sup>23</sup>

To further demonstrate the robustness of the SVM model and its superb transferability, we conducted more

stringent model validation. The materials in the library of 19 pharmaceutical compounds were from three different lots (lots A, B, and C) with three products different in grades and/or physical properties, or from different manufacturers for each compound. Thus, we were able to use data from one lot (lot A) as the training set and data from the other two lots (lots B and C) as the external testing set. Three validation tests were designed to systematically test the impact of material and instrument variation and robustness of the calibration model. For the baseline case, the same-unit-same-lot test (SUSL), the combined datasets from all three lots were used with half of the data for training and the other half for testing. The six spectrometers were tested respectively. For the second case, the same-unit-cross-lot test (SUXL), the lot A data were used as the training set, and the lot B and lot C data were used as the testing set. The six spectrometers were tested respectively. For the third case, the cross-unit-cross-lot test (XUXL), the lot A data from one unit were used as the training set, and the lot B and lot C data from a different unit were used as the testing set. Pairwise cross-unit validation was performed between the six spectrometers with 30 assessments in total. The XUXL case simulated the real world situation, where the calibration model needs to be transferred to different instruments, applied to materials from different sources and/or lots, and used at different test locations. Six models were compared for each case. The hier-SVM-linear model was left out of the test here, because theoretically the results from hier-SVM-linear model are expected to be the same as from the SVM-linear model when the number of classes is relatively small.

The validation results are presented in Figure 3 and Table 2. In Figure 3, individual prediction success rates are shown for each case: (a) SUSL; (b) SUXL; and (c) XUXL. In Table 2, average prediction success rates



**Figure 2.** PCA-SVM plot for 19 pharmaceutical compounds based on the SVM model.

and the corresponding standard deviation across different spectrometers or different combinations of spectrometers are reported. As shown in Figure 3a and Table 2, for the baseline case (SUSL), all of the models performed well with most of the prediction success rates close to 100%. The lowest prediction success rate was 92.44% for the sixth spectrometer for both LDA and QDA models. For the SUXL case, the prediction success rate was still above 90% for all of the models using all of the spectrometers. However, it can be clearly seen from Figure 3b and Table 2 that the two SVM models outperformed the others with most of the prediction success rates greater than 97% except for the second spectrometer, where the success rate was 94.23% using the SVM-rbf model. Among the other models, SIMCA provided the lowest prediction success rate. For the XUXL case, as shown in Figure 3c and Table 2, the prediction success rate significantly reduced when using LDA, QDA, and SIMCA. For the LDA and QDA models, the lowest success rate was below 62%. For the SIMCA model, the lowest success rate was below 60%. The SVM models outperformed the others in terms of not only the average success rate, but also consistency in the high success rate (>94%) for all the data points in Figure 3c. Between the two SVM models, SVM-linear performed better with 100% prediction success rate for 29 out of the 30 pairs, except when using the second spectrometer for training and the first spectrometer for testing, where the success rate was 96.43%, which was still higher than the success rates obtained by all the

other models using the same two spectrometers. PLS-DA performed better than SIMCA, LDA, and QDA with prediction success rate ranging from 88.26% to 97.08%. These validation results clearly demonstrate that the SVM-linear model is superior in terms of both generalization capability and model transferability.

### Large-Scale Classification

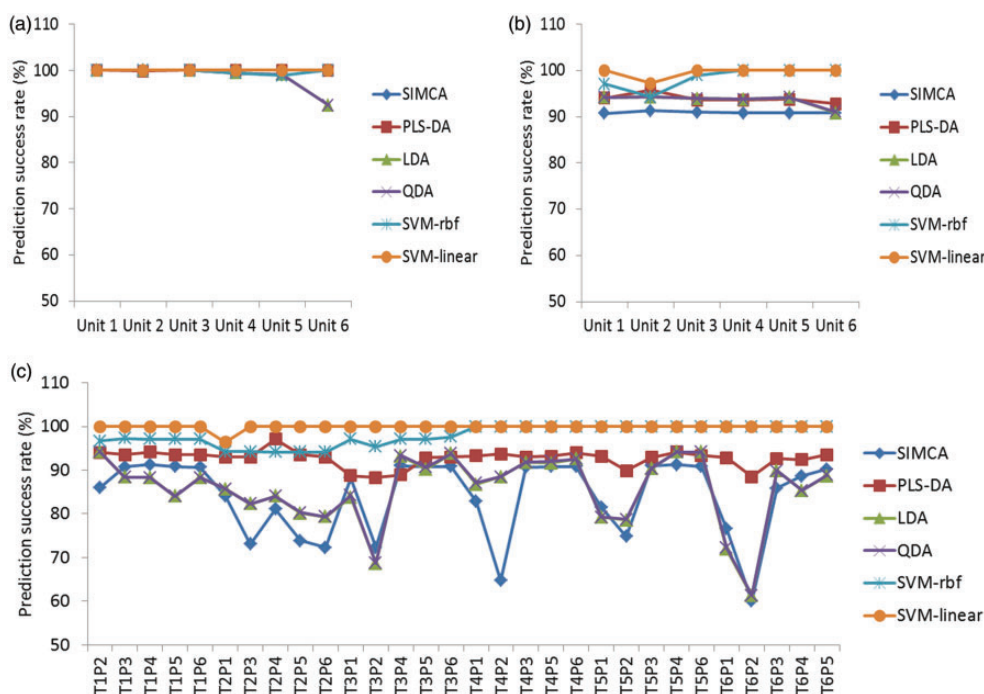
The library of 253 pharmaceutical compounds was used to investigate the capability of the chemometric models to solve the large-scale classification problem. Several types of differences were contained in this library. The main

**Table 2.** Model validation using different lots of materials and different MicroNIR spectrometers.

Classifier	SUSL		SUXL		XUXL	
	AVG <sup>a</sup>	STD <sup>b</sup>	AVG	STD	AVG	STD
SIMCA	100	0	90.93	0.22	83.92	8.92
PLS-DA	99.98	0.06	93.94	1.03	92.69	1.94
LDA	98.49	2.99	93.52	1.32	85.64	7.93
QDA	98.49	2.99	93.55	1.25	85.73	7.91
SVM-rbf	99.72	0.47	98.38	2.33	98.01	2.26
SVM-linear	100	0	99.54	1.14	99.88	0.65

<sup>a</sup>AVG: Average

<sup>b</sup>STD: Standard deviation



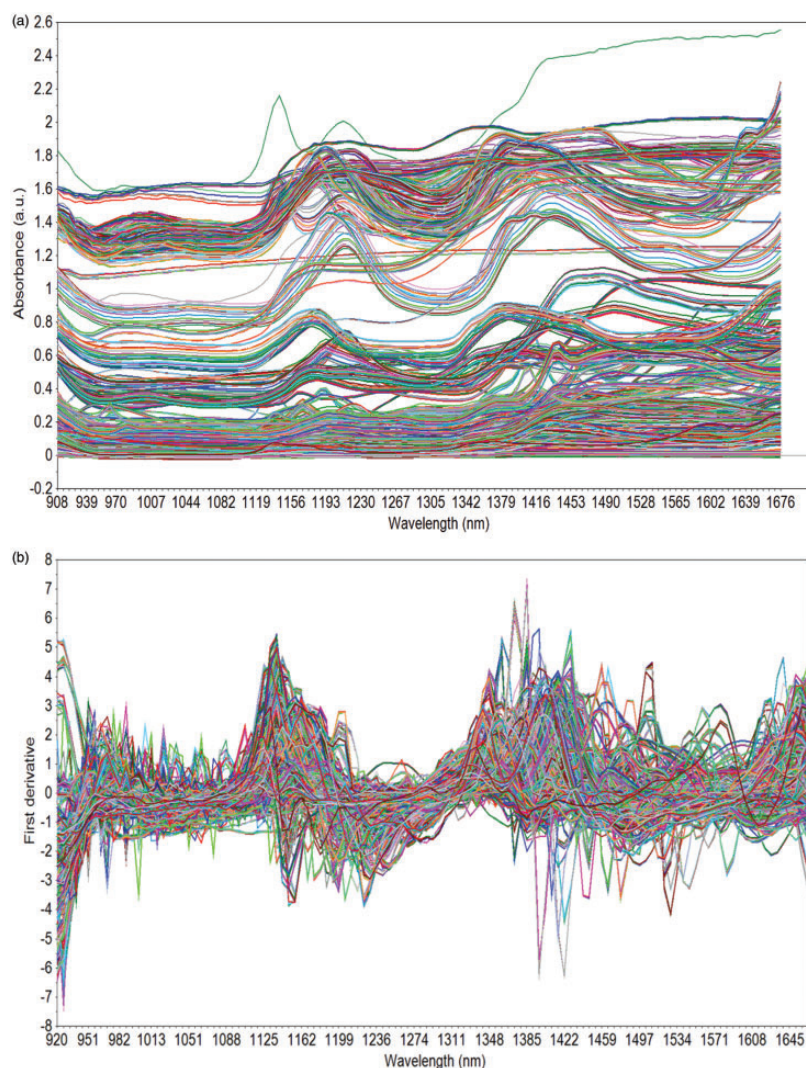
**Figure 3.** Model validation using different lots of material and different MicroNIR units. (a) Same-unit-same-lot (SUSL) validation; (b) same-unit-cross-lot (SUXL) validation; (c) cross-unit-cross-lot (XUXL) validation. Unit# denotes the spectrometer number. T#P# denotes the spectrometer number for training (T#) and the spectrometer number for testing or prediction (P#).

difference was chemical structure, but for some compounds, there were also differences in particle size, potency and formulation (coated versus uncoated) of the same materials. The goal was to test if all of these 253 compounds with both chemical and physical differences can be discriminated simultaneously.

As shown in Figure 4a, the raw spectra of different compounds exhibited broad bands and large variability in shape, intensity, and baseline over the whole spectral range. Spectral pretreatment is necessary to eliminate or minimize variability unrelated to the key spectral features for classification. The unwanted variation is usually caused by uncontrollable physical effects such as non-homogeneous distribution of the particles, changes in refractive index, sample packing/density variability, sample morphology, etc. However, it should be noted that some spectral features related to physical properties are important here since we also wanted to differentiate compounds with different physical properties in addition to chemically different

compounds. Therefore, moderate spectral pretreatment using Savitzky–Golay first derivative followed by SNV was selected. As shown in Figure 4b, more resolvable peaks were observed and the baseline drift was minimized in the pretreated spectra.

All of the seven models were then applied to the pretreated spectra of these 253 compounds. Half of the spectra were used as the training set and the other half was used as the test set. The results are summarized in Table 3 and the prediction accuracy is compared. Among all of these models, PLS-DA performed most poorly. A necessary condition for PLS-DA to work reliably is that each class is tight and occupies a small and separate volume in X-space consisting of the multivariate data, which was not satisfied by our dataset. Experience shows that PLS-DA is more appropriate with a small number of classes,<sup>30</sup> and thus not suitable for large-scale classification. Moreover, it took more than 20 hours to build this PLS-DA model in this work. Linear discriminant analysis (LDA) and QDA



**Figure 4.** NIR spectra of the training set from 253 pharmaceutical compounds. (a) Raw spectra; (b) pretreated spectra.

showed excellent performance for this set of data, which can probably be attributed to good settings of this dataset satisfying the preferred conditions for LDA and QDA, such as multivariate normal distribution of the class population and equal co-variances in the classes.<sup>31</sup> It only took seconds to build these models. However, it should be noted that LDA and QDA did not perform well compared with other models in terms of model transferability (Table 1, Table 2, and Figure 3). Also, when an additional compound with much less number of spectra was included (imbalanced dataset), the performance of LDA and QDA significantly worsened, while there were only slight changes in prediction success rate for the other models (data not shown). Soft independent modeling of class analogy (SIMCA) also performed well for this dataset. The good performance could be explained by the special feature of SIMCA that PCA is applied to each group separately and the number of PCs is selected individually and not jointly for all groups, which allows for an optimal dimension reduction in each group in order to reliably classify new objects.<sup>32</sup> However, it took minutes to build the SIMCA model, much slower than the LDA/QDA models, and model transferability of SIMCA was among the worst (Table 1, Table 2, and Figure 3).

One of the major features of SVM models is that they can operate in a kernel-induced feature space allowing both linear and nonlinear modeling.<sup>21,22</sup> In our work, the linear kernel (SVM-linear) performed well with a prediction success rate slightly lower than SIMCA, while the nonlinear kernel (SVM-rbf) did not perform well. It should be noted that parameter optimization is often conducted to build a SVM model, one parameter ( $C$ ) for the linear kernel and two parameters ( $C$  and  $\gamma$ ) for the RBF kernel, which usually involves an exhaustive search algorithm and takes a long computation time. From the practical perspective, in order to develop a robust method with minimal model building time and efforts, we intentionally skipped the parameter optimization procedure to test the SVM models using the default settings ( $C = 1.0$  and  $\gamma = 1/\text{number of variables}$ ). The prediction success rate of 96.57% of the SVM-linear model was excellent considering that no optimization

was performed, and it only took seconds to build the model. This good performance of the linear kernel could be related to the linear relationship between the absorbance and the concentration of the material according to Beer's law. In fact if optimization was performed, the performance of SVM-rbf was also excellent whereby the prediction success rate improved from 85.78% to 98.52%.

To further improve the performance of SVM, a hierarchical scheme was applied using the linear kernel (hier-SVM-linear), again without parameter optimization. With this approach, the prediction accuracy reached 100%. In pattern classification, the difference between the probability of assigning the sample to the winner class (the maximum probability) and the probability of assigning the sample to the second place class (the second maximum probability) is the key for prediction accuracy. In our work, the multi-class classification was realized by pairwise comparisons, known as all-pairs approach. The probability was calculated based on a multi-class probability estimate by combining the pairwise class probabilities using the algorithm proposed by Wu et al.<sup>33</sup> The maximum probability and the second maximum probability of classifying all the samples in the test set are shown in Figure 5. It can be clearly seen that for all the spectra, the differences between these two probabilities were significant, more than 0.33 for all the data except one of 0.08, indicating the strong discrimination power of hier-SVM-linear. It also only took seconds to build the hier-SVM-linear model, slightly longer than the SVM-linear model though.

External validation was performed to further substantiate the generalization capability of the SVM models. Spectra from 14 compounds out of the library of 19 compounds, which are also included in the library of 253 compounds, were used as the testing set. These 14 compounds are polysorbate 80, acetaminophen, corn starch, HPC, PVP, SSG, talc, ascorbic acid, benzocaine, caffeine, HPMC, lactose, Mg-stearate, and  $\text{TiO}_2$ . To be comparable with the testing sets, for these 14 chemically different materials in the library of 253 compounds, spectra from multiple compounds of the same chemical structure were consolidated to remove physical differences. There were 231 classes in the training set after the consolidation. All of the seven models were compared in this test. This validation test represents perhaps the most challenging situation, and often a real world situation, where a significantly larger number of classes, including chemically similar compounds to the compounds in the testing set, were used in the training set than in the testing set. It should be noted that the spectra in the testing sets were collected in our lab using six spectrometers, while the training set was collected from different samples using a different spectrometer (the seventh spectrometer) in a different environment by one of our external collaborators in the pharmaceutical industry. Therefore, this external validation test represented real world application. It can demonstrate not

**Table 3.** Comparison of different models for classification of the 253 pharmaceutical compounds.

Classifier	No. of spectra	Prediction success rate (%)	No. of missed predictions
SIMCA	2566	97.54	63
PLS-DA	2566	85.23	379
LDA	2566	99.61	10
QDA	2566	99.73	7
SVM-rbf	2566	85.78	365
SVM-linear	2566	96.57	88
hier-SVM-linear	2566	100	0



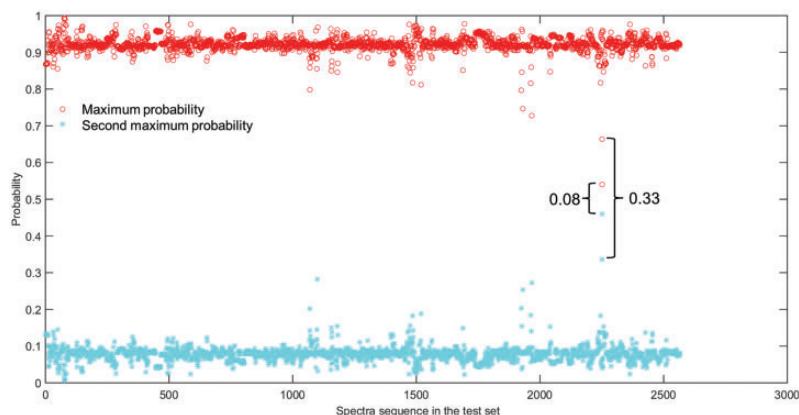


Figure 5. Estimation probabilities of class membership.

Table 4. Model validation using the large-scale classification model to predict samples from different sources with different MicroNIR units.

Classifier	T <sup>a</sup> -Unit7 P <sup>b</sup> -Unit1	T-Unit7 P-Unit2	T-Unit7 P-Unit3	T-Unit7 P-Unit4	T-Unit7 P-Unit5	T-Unit7 P-Unit6	AVG <sup>c</sup>	STD <sup>d</sup>
SIMCA	84.84	76.10	84.75	84.92	85.40	83.90	83.32	3.57
PLS-DA	85.71	85.89	85.00	85.71	85.71	85.69	85.62	0.31
LDA	33.02	31.97	57.75	45.87	67.78	70.57	51.16	16.86
QDA	33.10	32.13	58.33	47.78	68.41	70.73	51.75	16.91
SVM-rbf	91.67	92.08	93.42	90.32	90.79	86.67	90.83	2.30
SVM-linear	92.30	92.95	95.33	92.54	92.38	87.80	92.22	2.44
hier-SVM-linear	94.92	93.03	97.92	95.71	95.56	92.85	95.00	1.89

<sup>a</sup>T: Training

<sup>b</sup>P: Prediction

<sup>c</sup>AVG: Average

<sup>d</sup>STD: Standard deviation

only the predictive power of the model, but also the model transferability.

The individual prediction success rate using different spectrometers for the testing sets and the average value based on all of the six spectrometers for each model are summarized in Table 4. Support vector machine (SVM) models have clearly shown outstanding performance in both predictive power and model transferability. The average prediction success rates obtained by the three SVM models (>90% in most cases) were significantly higher than the other models. Moreover, the good performance was consistent across different spectrometers. The hier-SVM-linear model has shown the best performance, with prediction success rates ranging from 92.85% to 97.92% across the six spectrometers, which agrees with the results presented previously. By closely examining the misclassification results for the SVM models, we observed that a majority of the misclassified spectra were predicted as the classes of chemically similar materials. Among the other models, PLS-DA performed better. As shown in Table 4 and Figure 3, PLS-DA exhibited pretty good model transferability. However, it took very long time

(>20 h) to build the model and the prediction accuracy was not sufficient for large-scale classification. LDA and QDA performed very poorly in this validation test.

Successful on-site and in situ pharmaceutical RMID by NIR requires a robust chemometric model with high prediction success rate, reliable model transfer from instrument to instrument, the capability to handle a large number of materials with both chemical and physical differences, as well as short model building time and simple model settings for easy model building and rebuilding when necessary. Support vector machine (SVM) modeling, especially the hier-SVM-linear algorithm, meets all of these requirements and can be potentially used as a powerful classification tool in pharmaceutical industry.

## Conclusions

In this study, we demonstrated the use of MicroNIR spectrometers for NIR-based pharmaceutical RMID and solved two challenges in this area, model transferability and large-scale classification. The successful application can be attributed to the consistent instrument–instrument

performance of MicroNIR spectrometers as well as robust SVM modeling with accurate classification, excellent method transferability, fast model building and simple implementation. With capability of rapid, reliable and non-destructive analysis, it is highly promising to use the ultra-compact and portable MicroNIR spectrometers for on-site and in situ pharmaceutical RMID in order to inspect every barrel in every shipment of materials used in the manufacture of pharmaceutical drugs for the fulfillment of quality and safety standards in pharmaceutical industry.

### Conflict of Interest

The authors report there are no conflicts of interest.

### Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

### References

- M.R. Siddiqui, Z.A. AlOthman, N. Rahman. "Analytical Techniques in Pharmaceutical Analysis: A Review". *Arab. J. Chem.* (2013).
- D. Sorak, L. Herberholz, S. Iwaszek, S. Altinpinar, F. Pfeifer, H.W. Siesler. "New Developments and Applications of Handheld Raman, Mid-Infrared, and Near-Infrared Spectrometers". *Appl. Spectrosc. Rev.* 2011. 47(2): 83–115.
- M.A. Druy. "Molecular Spectroscopy Workbench Applications for Mid-IR Spectroscopy in the Pharmaceutical Process Environment". *Spectroscopy*. 2004. 19(2): 60–63.
- J. Luybaert, D.L. Massart, Y. Vander Heyden. "Near-Infrared Spectroscopy Applications in Pharmaceutical Analysis". *Talanta*. 2007. 72(3): 865–883.
- T. Vankeirsbilck, A. Vercauteren, W. Baeyens, G. Van der Weken, F. Verpoort, G. Vergote, J.P. Remon. "Applications of Raman Spectroscopy in Pharmaceutical Analysis". *Trends Anal. Chem.* 2002. 21(12): 869–877.
- B. Swarbrick. "Review: Advances in Instrumental Technology, Industry Guidance and Data Management Systems Enabling the Widespread Use of Near Infrared Spectroscopy in the Pharmaceutical/biopharmaceutical Sector". *J. Near Infrared Spectrosc.* 2014. 22(3): 157–168.
- "Near-Infrared Spectrophotometry". United States Pharmacopoeia USP 38 NF33. 2015.
- "Near-Infrared Spectroscopy". European Pharmacopoeia. 8th ed. 2013.
- M. Blanco, M.A. Romero. "Near-Infrared Libraries in the Pharmaceutical Industry: A Solution for Identity Confirmation". *Analyst*. 2001. 126(12): 2212–2217.
- W.B. Mroczek, K.M. Michalski. "Application of Modern Computer Methods for Recognition of Chemical Compounds in NIRS". *Comput. Chem.* 1998. 22(1): 119–122.
- K. Kreft, B. Kozamernik, U. Urleb. "Qualitative Determination of Polyvinylpyrrolidone Type by near-Infrared Spectrometry". *Int. J. Pharm.* 1999. 177(1): 1–6.
- K. Krämer, S. Ebel. "Application of NIR Reflectance Spectroscopy for the Identification of Pharmaceutical Excipients". *Anal. Chim. Acta*. 2000. 420(2): 155–161.
- M. Blanco, J. Coello, H. Iturriaga, S. Maspocho, C. Pérez-Maseda. "Determination of Polymorphic Purity by near Infrared Spectrometry". *Anal. Chim. Acta*. 2000. 407(1–2): 247–254.
- M. Blanco, A. Villar. "Development and Validation of a Method for the Polymorphic Analysis of Pharmaceutical Preparations Using near Infrared Spectroscopy". *J. Pharm. Sci.* 2003. 92(4): 823–830.
- B. Grout. "Application of Near Infrared Conformance Trending for Material Quality, Container Consistency and Minimisation of Process Risk". *J. Near Infrared Spectrosc.* 2014. 22(3): 169–178.
- M. Alcalá, M. Blanco, D. Moyano, N. Broad, N. O'Brien, D. Friedrich, F. Pfeifer, H.W. Siesler. "Qualitative and Quantitative Pharmaceutical Analysis with a Novel Handheld Miniature near-Infrared Spectrometer". *J. Near Infrared Spectrosc.* 2013. 21(6): 445–457.
- Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent. "A Review of Near Infrared Spectroscopy and Chemometrics in Pharmaceutical Technologies". *J. Pharm. Biomed. Anal.* 2007. 44(3): 683–700.
- I. Storme-Paris, H. Rebiere, M. Matoga, C. Civade, P.A. Bonnet, M.H. Tissier, P. Chaminade. "Challenging Near Infrared Spectroscopy Discriminating Ability for Counterfeit Pharmaceuticals Detection". *Anal. Chim. Acta*. 2010. 658(2): 163–174.
- E. Dreassi, G. Ceramelli, P. Corti, S. Lonardi, P.L. Perruccio. "Near-Infrared Reflectance Spectrometry in the Determination of the Physical State of Primary Materials in Pharmaceutical Production". *Analyst*. 1995. 120(4): 1005–1008.
- M. Andre. "Multivariate Analysis and Classification of the Chemical Quality of 7-Aminocephalosporanic Acid Using Near-Infrared Reflectance Spectroscopy". *Anal. Chem.* 2003. 75(14): 3460–3467.
- J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, J.A.K. Suykens. "A Tutorial on Support Vector Machine-Based Methods for Classification Problems in Chemometrics". *Anal. Chim. Acta*. 2010. 665(2): 129–145.
- O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, J.-P. Huvenne. "Support Vector Machines (SVM) in Near-Infrared (NIR) Spectroscopy: Focus on Parameters Optimization and Model Interpretation". *Chemom. Intell. Lab. Syst.* 2009. 96(1): 27–33.
- D.M. Friedrich, C.A. Hulse, M. von Gunten, E.P. Williamson, C.G. Pederson, N.A. O'Brien. "Miniature Near-Infrared Spectrometer for Point-of-Use Chemical Analysis". *Proc. SPIE*. 2014. 8992: 899203.
- C.G. Pederson, D.M. Friedrich, C. Hsiung, M. von Gunten, N.A. O'Brien, H.-J. Ramaker, E. van Sprang, M. Dreischor. "Pocket-Size Near-Infrared Spectrometer for Narcotic Materials Identification". *Proc. SPIE*. 2014. 9101: 91010O.
- N.A. O'Brien, C.A. Hulse, D.M. Friedrich, F.J. Van Milligen, M.K. von Gunten, F. Pfeifer, H.W. Siesler. "Miniature Near-Infrared (NIR) Spectrometer Engine for Handheld Applications". *Proc. SPIE*. 2012. 8374: 837404.
- J.J.R. Rohwedder, C. Pasquini, P.R. Fortes, I.M. Raimundo, A. Wilk, B. Mizaikoff. "iHWG- $\mu$ NIR: A Miniaturised Near-Infrared Gas Sensor Based on Substrate-Integrated Hollow Waveguides Coupled to a Micro-NIR-Spectrophotometer". *Analyst*. 2014. 139(14): 3572–3576.
- S. Liu, H. Yi, L. Chia, D. Rajan. "Adaptive Hierarchical Multi-Class SVM Classifier for Texture-Based Image Classification". In: *Multimedia and Expo (ICME), IEEE International Conference on*. Amsterdam, the Netherlands: 6–8 July 2005. doi: 10.1109/ICME.2005.1521640.
- C.N. Silla, A.A. Freitas. "A Survey of Hierarchical Classification across Different Application Domains". *Data Min. Knowl. Discov.* 2011. 22(1–2): 31–72.
- R.W. Kennard, L.A. Stone. "Computer Aided Design of Experiments". *Technometrics*. 1969. 11(1): 137–148.
- L. Eriksson, T. Byrne, J. Johansson, J. Trygg, C. Vikström. *Multi- and Megavariable Data Analysis: Basic Principles and Applications*. Umetrics Academy, 2013.
- W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni. "Comparison of Regularized Discriminant Analysis Linear Discriminant Analysis and Quadratic Discriminant Analysis Applied to NIR Data". *Anal. Chim. Acta*. 1996. 329(3): 257–265.
- K. Varmuza, P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton, FL: CRC Press, 2009.
- T.-F. Wu, C.-J. Lin, R.C. Weng. "Probability Estimates for Multi-Class Classification by Pairwise Coupling". *J. Mach. Learn. Res.* 2004. 5: 975–1005.