24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Application of Random Forests in ToF-SIMS Data

Yinghan Zhao[a,*], Svenja-K. Otto[c], Nico Brandt[a], Michael Selzer[a,b], Britta Nestler[a,b]

[a]*Institute for Applied Materials (IAM), Karlsruhe Institute of Technology (KIT), Karlsruhe 76131, Germany*
[b]*Institute of Digital Materials (IDM), Karlsruhe University of Applied Sciences, Karlsruhe 76133, Germany*
[c]*Institute for Physical Chemistry, Justus Liebig University Giessen (JLU), Giessen 35392, Germany*

Abstract

Surface analysis techniques are particularly important in the field of materials science, which help researchers to understand the mechanism behind complex chemical reactions and study the properties of different materials. Time-of-flight secondary ion mass spectrometry (ToF-SIMS), a highly sensitive surface analysis technique, allows the reliable determination of various materials. ToF-SIMS spectra of materials are usually enormously complex since typical raw data may include many peaks over large mass-to-charge ratio (m/z) ranges. Hence, the use of data-mining methods in processing ToF-SIMS data is becoming more popular and important. In this study we show that random forests model can be used to automatically classify several different lithium-containing materials and to extract representative peaks from ToF-SIMS spectra of these materials. Our study shows good performance in analyzing spectra of materials with similar and dissimilar compositions, which can provide researchers with the possibility of quick and automatic analysis of ToF-SIMS data.

1. Introduction

The emerging requirement of novel and specific engineering materials has increased the need for surface analysis methods that can accurately characterize and verify the actual surface chemistry. This is particularly challenging when taking the complexity of modifications to material surface into consideration. The modern trend towards using multicomponent and patterned surface chemistry further increases the difficulty and complexity during the analysis. Though no technique can solely provide a complete characterization of such complex material surface, time-of-flight secondary ion mass spectrometry (ToF-SIMS), exhibits great promising prospects in solving complex surface analysis problems due to its advantages of molecular specificity, relatively high mass resolution, high sensitivity and precision [1]. ToF-SIMS is a mass spectrometry technique which probes the chemical composition and structure of the surface

* Corresponding author. Tel.: +49-721-608-41470.
  E-mail address: yinghan.zhao@kit.edu

of the material by impacting energetic primary ions onto the sample followed by analyzing the secondary ions emitted from the surface. Commonly used primary ions in ToF-SIMS range from mono-atomic ions (Ar^+, Ga^+, Au^+, Bi^+, etc.) to cluster ions such as Au_n^+ (n= 1–3), Bi_n^+ (n= 1–5), etc. By performing this analysis the primary ion beam is carefully controlled to a sufficiently low intensity to guarantee that the surface molecules are not completely broken into individual atoms. The fragment ions emitted from the surface (secondary ions) are transferred into a "flight tube" and the mass-to-charge ratio is determined by measuring the time at which they reach the detector, that is, the so-called "time-of-flight". ToF-SIMS is considered as a technique for elemental analysis of solid materials which exploits the sputtering process as an efficient mean for ion formation from the material surface. It is practically useful for the characterization of samples of technological interest in the field of microelectronics, corrosions science and materials science such as polymeric materials and general organic samples [2]. In addition, it is widely used to perform laterally revolved analysis and the in-depth characterization of structured samples with resolution small to nanometer scale, which yields 3D rendering of the chemical compositions of materials [3]. Besides, with modification and improvement of the instrumentations, it can also be applied to characterizing samples from biological research such as cells and tissues [4, 5].

A typical ToF-SIMS spectrum can comprise hundreds of peaks, the intensity of which may vary due to composition, structure, order and orientation of the surface materials [6]. As a result, ToF-SIMS data are usually regarded as inherently multivariate since the relative intensities of different peaks within a given spectrum are related to some extent, as they often originate from the same surface species. Therein lies the central challenge, that is, the determination of which peaks are related to each other and how they are linked with the chemical differences presented on the sample surface. As the number of samples increases, the amount of information to investigate also grows rapidly and widely. Hence, data reduction/compression methods and automatic analysis approaches are desperately demanded for extracting the most significant information from the original dataset.

In this work, we focus on ToF-SIMS data sampled from 6 different lithium-containing materials, namely, Li_2CO_3, Li_2O, Li_2O_2, Li_3N, LiH and LiOH. These materials are regarded as the most possible chemical products by exposing the lithium metal in the air. These data were acquired from our research partner in the study of all-solid-state batteries (ASSBs) [7]. ASSBs have emerged as a particularly promising next generation battery owing to their potential to have high safety, high energy density and long cycle life. There is considerable interest in the usage of lithium metal as the anode in the battery by virtue of its high theoretical capacity (3860 mAh/g) and large negative potential (−3.06 vs. NHE) [8]. However, Li metal is prone to unstable side reactions, leading to solid electrolyte interphase (SEI) and dendrite formation [9]. Hence, compositional analysis of lithium anode has significant importance in understanding the chemical reaction during battery operation, which may help researchers to improve and develop all-solid-state batteries.

In order to deal with the complexity of ToF-SIMS data and to perform fast and effective analysis, various data analysis approaches are adopted, such as PCA (principal component analysis) and MCR (multivariate curve resolution) in the field of multivariate analyses (MVA) [10]. Besides, data-mining methods are gaining more and more popularity in recent years with the rising of machine learning. Support vector machines, decision trees, hierarchical clustering analysis, artificial neural network and other methods have been successfully applied in the field of mass spectrometry [11, 12]. Machine learning models possess the ability to explore latent relationships behind the complex data. Among these algorithms, random forests model, a decision tree based ensemble method which owns many favorable properties such as robustness to over-fitting, relatively high prediction accuracy and the capability of dealing with a great number of input variables, has successfully been applied to various kinds of spectral data, including ToF-SIMS [13].

In this study we adopt random forests to classify the 6 different lithium-containing materials according to their ToF-SIMS spectra. Besides, we also explore the usage of this algorithm to determine the characteristic peaks of these samples we tested, which can provide researchers with a quick and automatic method to analyze ToF-SIMS data in order to distinguish materials with similar and dissimilar compositions.

2. Data and Methodology

2.1. Data

For ToF-SIMS analysis, 6 different lithium-containing materials, namely, Li_2CO_3, Li_2O, Li_2O_2, Li_3N, LiH and LiOH, were pressed into pellets (3 t, 1.5 min) under Ar-atmosphere. The subsequent transfer to the instrument was also done under Ar-atmosphere. The ToF-SIMS measurements were performed using a TOF.SIMS 5-100 instrument (ION-TOF GmbH, Muenster, Germany) equipped with a 25 kV Bi-cluster primary ion gun for analysis and with a 20 kV gas cluster ion beam (GCIB) for sputtering. The 6 samples were examined under the same experimental condition. Before the analysis, corresponding surface regions were sputter-cleaned with Ar_{1500}^+ (10 kV, 10 nA) using a dose density of 4×10^{15} ions/cm^2. For analysis, areas of $100{\times}100\,\mu m^2$ were measured in spectrometry mode, using Bi_3^+ (20 kV, 1 pA) as primary ion with an ion dose density of 10^{12} ions/cm^2. Measurements were done in negative ion mode, with a cycle time of 100 μs. A flood gun was used for charge compensation during sputtering and analysis. For each material, 12 spots were examined, hence the dataset consists of 72 spectra in total. The mass range was truncated to m/z = 120 for the consideration that large mass regions have weak ability of identifying materials in our research. For each spectrum 143 possible important mass-to-charge ratio candidates were manually selected based on researchers' domain knowledge.

2.2. Pre-processing

Data pre-processing is one of the important phases in data mining process and results produced by various analysis methods depend largely on pre-processing of the raw data [14]. It involves many factors, including: accuracy, completeness, consistency, credibility, interpretability and so on. The quality of data plays a very important role in determining the prediction and generalization ability of the model. In the real world, data may contain a lot of missing values, may mix with a lot of noise, or there may be abnormal points due to manual input errors, which are not conducive to the training of algorithm models. The result of data preprocessing is to process various "dirty" raw data in a corresponding manner to obtain standard, clean and continuous data, which is then provided for data statistics and data mining. Pre-processing methods usually include cleaning, normalization, scaling, transformation, feature extraction and selection, etc. They may affect the way how the results of the final data processing can be interpreted and hence they are usually carefully considered in the multivariate processing of data. One of the important objectives of data pre-processing such as scaling is to avoid features with greater value ranges dominating those have smaller numeric ranges, in other words, to bring data into similar ranges and to minimize the effect of variance so that all of the variables can make a comparable contribution to the result of the model.

By performing the measurements, variations can be introduced by instrumentations, due to topography or other factors, which cause differences in the absolute values of intensities of the spectrum. As ToF-SIMS is a highly sensitive surface analysis technique, a large proportion of its spectral data may be noise. Thus pre-processing of the raw data must be carefully considered. Some physical considerations are worth paying attention to, such as Poisson scaling for the reason that ToF-SIMS data often follow Poisson distributions [15]. In this study, for simplification, each intensity value is first subtracted from the minimum value in the corresponding channel, and then divided by the range (difference between the original maximum and original minimum in that channel). Hence the value of intensity in each mass channel is scaled to a value between 0 and 1. An example of this scaling method is depicted in figure 1. It shows one of the raw spectra of Li_2CO_3 and its corresponding scaled spectrum, with intensity at each mass-to-charge ratio scaled to a value between 0 and 1.

In this study, considering the limited number of samples (72), we adopt k-fold cross-validation [16] to take full advantage of the data and assess the performance of the model in a fair manner. In k-fold cross-validation the original samples are partitioned into k equal sized disjoint subsamples randomly. Among the k subsamples, one single subsample is kept as the validation data for testing the performance of the model, while the remaining k − 1 subsamples are retained as training data. This cross-validation process is then repeated k times, with each of the k subsamples used exactly only once as the validation data. The k results can finally be averaged to produce the overall estimation. 3, 5, 10 are commonly used values for k, and in this work before training the machine learning model, the original dataset is shuffled and then is divided into training part and validation (test) part following the 5-fold cross-validation scheme. In each cross-validation, scaling is firstly conducted on the training data, and the test part is also scaled with
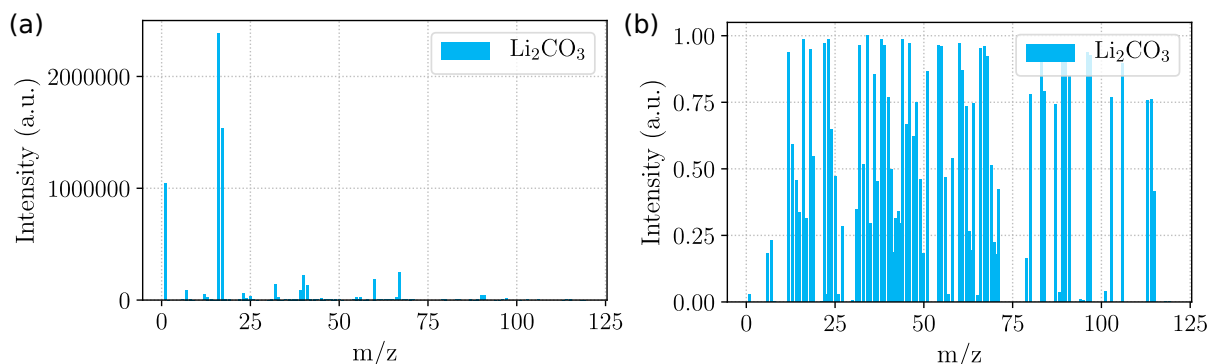
Figure 1: Example of ToF-SIMS spectrum of Li_2CO_3: (a) raw spectrum without pre-processing; (b) its scaled spectrum after the pre-processing. Intensity (a.u.) at each mass-to-charge ratio (m/z) is scaled to a value between 0 and 1.

the same values as those of the training data. In this way, the test data are kept "unforeseen" for the model to ensure the fairness by judging the performance of the model.

2.3. Performance Evaluation

To evaluate the result of our model, the standard performance measures of pattern classification including precision, recall and F_1 score are calculated. In information retrieval, precision is a measure of result relevancy while recall is a measure of how many truly relevant results are returned. F_1 score is the harmonic mean of precision and recall. These three evaluation metrics are defined based on the following four basic test statistics in classification task:

- True positive (TP): the cases in which the model predicts positive and the actual label is positive
- True negative (TN): the cases in which the model predicts negative and the actual label is negative
- False positive (FP): the cases in which the model predicts positive but the actual label is negative
- False negative (FN): the cases in which the model predicts negative but the actual label is positive

So they can be calculated as:

- Precision (P) = TP/(TP+FP)
- Recall (R) = TP/(TP+FN)
- F_1 score = (2PR)/(P+R)

In this work, these three measures are calculated and the confusion matrix is given to evaluate the performance of random forests models by the classification of the six lithium-containing materials.

2.4. Machine Learning Model

Machine learning algorithms aim to optimize the performance of certain tasks based on existing samples. The problem of materials determination discussed in this paper falls into the supervised learning category. Supervised learning has the intention of finding the unknown function which connects inputs to labeled outputs. In the field of machine learning, these inputs are usually denoted as "features" of the data and the outputs are usually called "labels" (in classification tasks). In this study signal intensities at different mass-to-charge ratios are used as inputs and species of materials are considered as outputs. Random forests [17, 18], a kind of statistical learning method, is one of the most popular and well-known models in the field of machine learning and data-mining. Considering its many advantages, such as relatively good resistance to over-fitting and the ability to evaluate the importance of variables, we adopt this approach in this study to classify different materials and to determine characteristic peaks of these materials at the same time.

Random forests model is a kind of decision tree [19, 20] based ensemble method which takes the advantage of the bootstrap aggregation. It is suitable for both regression and classification analyses. In contrast to the single decision tree model, a randomized tree ensemble is robust to noisy data and can reduce the probability of over-fitting. In the typical training setup, a few hundred to thousand single decision trees are constructed and they build up the "forest" as a whole. Given the original dataset with n samples: $X = x_1, x_2, ..., x_n$ and its corresponding target set $Y = y_1, y_2, ..., y_n$, the algorithm first randomly selects n samples with replacement (the so-called "bootstrapping") from the original dataset to form a sub-training set. The extracted sub-training set is then used for building the decision tree model (classification or regression tree depending on tasks), whereas the rest of the samples from the original dataset (out-of-bag samples, OOB samples) is used for estimating the out-of-bag error (OOB error). This procedure is repeated m times to obtain m different decision tree models. After the training procedure, the prediction of the new sample can be achieved by averaging the predictions of all the m regression trees on that sample in regression task, or by using the majority voting rule in classification task. However, a problem with this traditional decision trees training procedure is that the obtained decision trees still have a lot of structural similarities which in turn have high correlation in their predictions. This is caused by the greedy algorithm adopted during the training process, as it tries to minimize the error when choosing variables to split decision nodes. Combining predictions from multiple models in ensembles works better if the predictions from the sub-models are uncorrelated or at most weakly correlated. Hence the random forests model employs an improved learning algorithm to select a random subset of features in each candidate splitting during the learning process, which is sometimes called "feature bagging". In this way sub-trees are learned in the way that the resulting predictions from all of the sub-trees have low correlation with each other, which can lead to better generalization of the model.

The importance of the different features (in this study they are the mass channels of spectra) for the classification of different materials can be estimated with the "mean decrease accuracy" using out-of-bag (OOB) samples. The first step in measuring the feature importance of a dataset is to train a random forests model based on the training set. During the training process, the OOB error of each data point is recorded, and then averaged across the whole forests. In order to measure the importance of the i-th feature, values of the i-th feature are scrambled in the training data, and the OOB error of the scrambled data is recalculated. Then the importance score of the i-th feature can be obtained by calculating the average of the OOB errors before and after the disruption of this feature. Afterwards this score is normalized by calculating the standard deviation of these differences. Features that produce larger scores are more important than features that have smaller scores.

In our study, we use random forests algorithm implemented by the Python library scikit-learn [21], which is a popular and open source library for machine learning. Models are trained on a desktop computer (CPU: Intel Core i5-6700K, RAM: 16GB, no discrete GPU) and can be finished in less than 1 minute.

3. Results and Discussion

The ToF-SIMS dataset in this study consists of 72 spectra from 6 different species of materials (12 spectra from each material). Samples we tested are lithium-containing materials, namely, Li_2CO_3, Li_2O, Li_2O_2, Li_3N, LiH and LiOH. The random forests model is adopted to classify these six kinds of materials. Since we also need to determine characteristic mass-to-charge ratios (m/z) which can best separate each material from others, we break this problem down to 6 binary classification tasks. For each material we train a random forests model, and for each model, the selected material is marked as positive sample while others are marked as negative samples. As we adopt 5-fold cross-validation scheme, in each training procedure, 80% of the data is used as the training set and the rest 20% is left as the test set. Each random forests model consists of 50 decision trees, the number of which is determined by trial-and-error to ensure its generalization performance. The final characteristic mass-to-charge ratios (m/z) are selected according to the average feature importances calculated by the models during 5-fold cross-validation.

Here we focus on the analysis of the tested Li_2CO_3 samples as an example. During the training procedure we set Li_2CO_3 as the positive class and treat others as negative ones to perform the classification task. After the training we calculate the importances of different features (the mass-to-charge ratios, m/z). The top 10 most important features selected by the model when separating Li_2CO_3 from other materials are summarized in figure 2. These features correspond to characteristic peaks (m/z) of the Li_2CO_3 samples we tested. The importances of peaks are arranged in descending order from top to bottom with darker color representing higher degree of importance. It can be clearly

seen that the mass-to-charge ratio (m/z) channels 59.9862 and 51.0072 are regarded by the random forests model as the most important peaks (features) of this material. It is reasonable for our model to select m/z = 59.9862 as this mass channel corresponds to mass of the ion CO_3^{2-} which comes from Li_2CO_3. Channel m/z = 67.0031 ranks the second place followed by channels m/z = 79.9589 and 86.9765 which together rank the third place among all the features. Besides, several different channels (m/z = 39.0072, 14.0317, 54.0011, 113.027 and 91.0279) are ranked the same order by the model. Signals in different mass channels marked as equally important are treated in the same manner by the identification of that material. From the figure we can draw the conclusion that some peaks found by the model fit well with the chemical information (e.g. CO_3^{2-}: m/z = 59.9862 and $LiCO_3^-$: m/z = 67.0031), while others may come from contaminations or products generated by the complex gas phase reactions during the measurement.
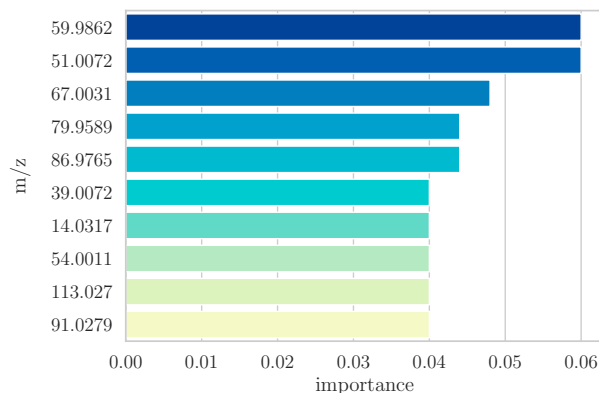

Figure 2: Top 10 important peaks of the tested Li_2CO_3 samples which are selected by random forests model. The larger value the peak has (marked with darker color), the more important it is to represent this sample.

In order to investigate why our model treats these mass channels as the most important ones by identifying Li_2CO_3 samples, we plot the intensities of the 10 selected mass channels of all the 6 different materials in figure 3. It illustrates the scaled average intensities of these corresponding mass-to-charge ratios. The intensities (a.u.) of the Li_2CO_3 samples are marked as red while other samples are colored in blue for better distinction. On the one hand, as can be clearly seen from the diagram, except for the channel m/z = 14.0317, in all other 9 channels, Li_2CO_3 exhibits the strongest intensity among the 6 materials. This means these selected mass channels are indeed the characteristic peaks of the Li_2CO_3 samples (though some peaks may come from the impurities attached to the samples we tested). On the other hand, it is noticeable that in the seventh mass channel where m/z = 14.0317, Li_2CO_3 exhibits the lowest intensity among the 6 materials. As the task of the machine learning model here is to try to find the most significant attributes that can best discriminate the target material from others despite the absolute value of the peak, it is reasonable that the model also regards peak with relatively low intensity as one of the most important features. It indicates that sometimes the machine learning model may interpret data in such a way that may be nonintuitive to human beings. In practical use mass channels with larger intensities are preferred by researchers as they can be better recognized. Because the model has chosen several characteristic peaks, we can further adopt some filter rules by the selection of them, such as picking up those peaks with intensities above the average value in the whole spectra or selecting those match their elemental or compositional representations. Overall, random forests model has detected specific peaks that can be used to uniquely determine one certain material. Situations are similar for other materials, but sometimes a deeper inspection into materials with similar compositions (such as to distinguish between Li_2O and Li_2O_2) is required. A complete comparison and analysis of other materials are not described in details in this paper.

For a better visualization, we represent these 6 different materials in a low dimensional space with two of the most important mass channels of the Li_2CO_3 samples (m/z = 59.9862 and m/z = 67.0031 which fit well with the chemical information) in figure 4(a). Materials with corresponding ions excited from the surface in these two mass channels should have greater values and hence should appear in upper right side of the diagram. Different materials are marked with different colors with Li_2CO_3 in blue. It is apparent from the figure that Li_2CO_3 data points are located at upper right corner, representing relatively strong intensities in these two channels. It can be noted that even for the same
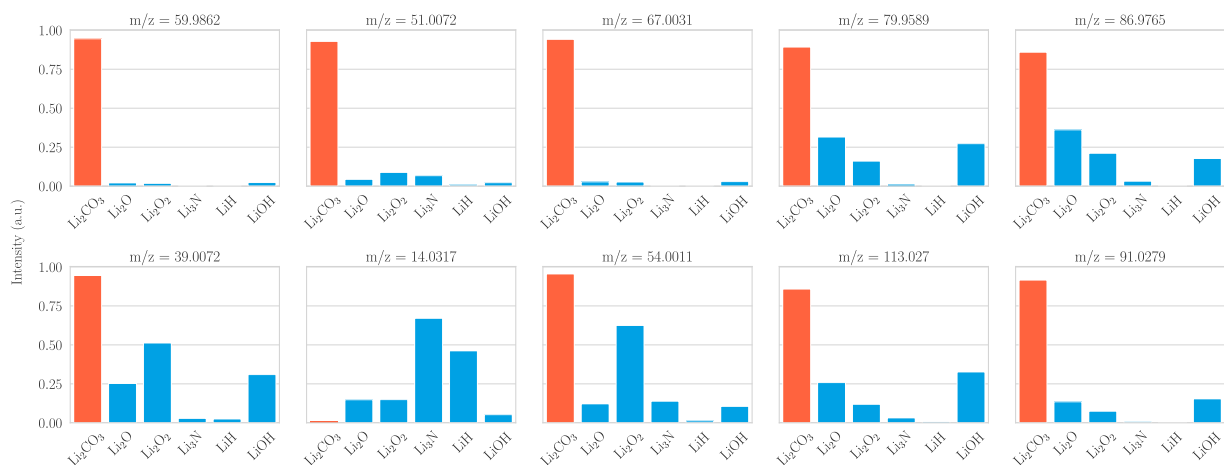
Figure 3: Scaled average intensities (a.u.) in the top 10 most important mass-to-charge ratios of the Li_2CO_3 samples selected by random forests model corresponding to peaks shown in figure 2. Li_2CO_3 is marked red while other materials are colored in blue.

material, values of the signal intensity in the same channel may fluctuate slightly. As can also be drawn from the same figure, other materials with relatively low intensities in these two channels all gather at the bottom left corner of the diagram. In other words, the Li_2CO_3 samples have generated specific ions in these two mass channels during the characterization process, while others have not. This once again proves that these mass channels selected by our model are the characteristic peaks of the Li_2CO_3 samples. For comparison, we choose another two channels (m/z = 49.0305 and m/z = 1.0084) that are not considered important for Li_2CO_3 by the model. How these 6 materials are distributed in these two channels is illustrated in figure 4(b). This diagram reveals that in these two channels, even the same material can exhibit relatively large fluctuations in its intensities, as its data points are scattered loosely over relatively large range. This phenomenon could be caused by the complicated chemical or physical gas phase reactions during the characterization process, or the samples themselves may have slight differences in their chemical compositions over different testing spots. It is well-known that ToF-SIMS analysis is a highly sensitive technique and is usually performed under strict experimental conditions. Therefore, any tiny changes of chemical compositions on the surface of the examined sample may result in a relatively large fluctuation in the spectrum. Furthermore, matrix effects caused by components in the sample other than the analyte, further add significant interference and hence would affect the accuracy of analysis results. As can be seen from the figure, data points of Li_2CO_3 (blue points) are mixed with data points of Li_3N (red points) and parts of of Li_2O_2 (green points). As these materials exhibit similar intensities in these two channels, it is difficult to distinguish Li_2CO_3 from other materials using information from these two mass-to-charge ratios, which may serve to explain why our machine learning model gives them low importance percentages. Besides, it is interesting to find that at m/z = 1.0084 where anion hydride (H^-, negatively charged) contributes the most intensities, most of the examined samples in this study show some signals in this channel. We infer these signals might come from the hydrocarbons attached to the sample surface or were introduced during the process of instrument operations. Unsurprisingly, materials LiOH and LiH reveal relatively large intensities at m/z = 1.0084 as they themselves contain element H in their compositions. As some contaminations or complex intermediate products generated during the characterization process will raise obstacles to an accurate analysis of materials, a better design of the experiment may be required to minimize these side effects. In summary, we introduce machine learning model to determine the characteristic peaks of the 6 materials we tested, rather than basing solely on their physical representations at corresponding mass-to-charge ratios.

In this work, as the raw dataset is limited and not complex, the model of each material can achieve good performance with all the mass-to-charge ratios (original features) for the classification. As we need to know the characteristic peaks (important features) which can best separate each material from others, binary classification models of each material are trained individually to classify material pairs and to obtain corresponding important mass-to-charge ratios of different materials. Confusion matrices related to the results of models are plotted in figure 5. Examining the confu-
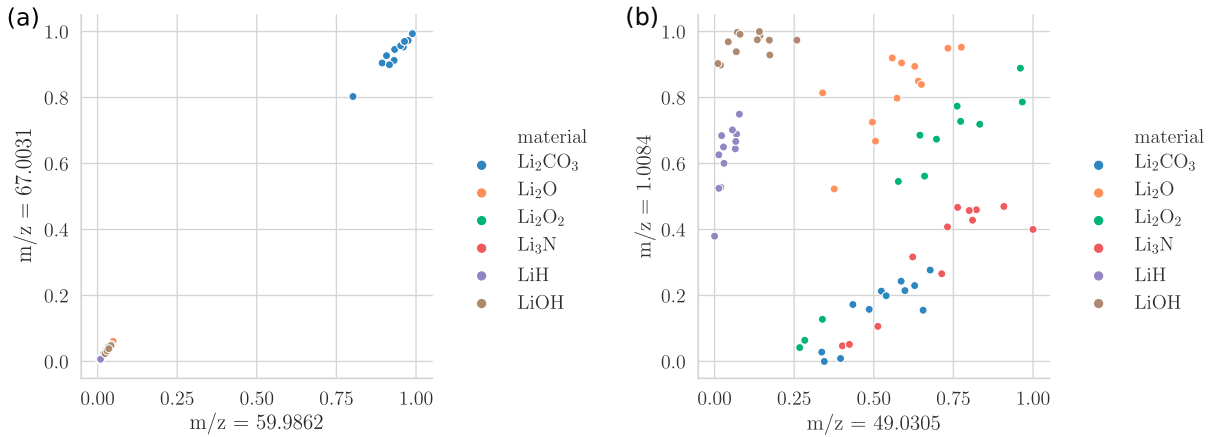
Figure 4: Intensities (a.u.) at selected mass-to-charge ratios (m/z) of six different tested materials: (a) m/z = 59.9862 and m/z = 67.0031, random forests model predicts high importance in these two channels by separating Li_2CO_3 from others; (b) m/z = 49.0305 and m/z = 1.0084, random forests model predicts low importance for Li_2CO_3. Different materials are marked with various colors shown in the legend.

sion matrices for each binary classification indicates the ability of models to successfully differentiate between various materials. An ideal model would result in a confusion matrix with all diagonal values being the number of samples belonging to that kind of material (in each binary classification 12 positive and 60 negative samples) and all off-diagonal values being zero. It can be seen that the model makes two misjudgments by mistaking Li_2O and LiH, one possible explanation could be the similarity of their chemical compositions (Li_2O/Li_2O_2 and LiH/LiOH). Further analysis of the cause of misjudgments and how to improve the performance of model are in progress along with improving the experimental procedure to reduce signal noise and reduce contaminants on the samples.
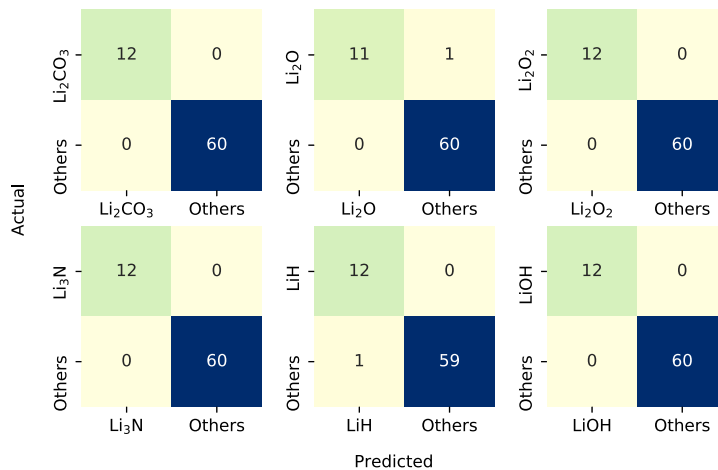

Figure 5: Confusion matrices of random forests models for each binary classification. The vertical labels indicate the actual kinds of material and the horizontal labels are those predicted by the model. It illustrates the number of corresponding prediction results for each material with the colors scaled to the number of occurrences.

To evaluate the results of our model, the precision, recall and F_1 score based on the confusion matrices of each binary classification are calculated and they are given in table 1. It can also be seen that we have obtained a model with good performance (average precision: 0.987, average recall: 0.986 and average F_1 score: 0.986) with all the mass-to-charge ratios (original features). This indicates our model is highly accurate when differentiating between various material pairs. The study of how many features should be kept to achieve a comparably high classification

performance for each material is left for the future work. Besides, considering the size of dataset is strictly limited (here we have only 72 spectra), collecting more data points from more material species will be important to establish a robust model and will be meaningful for wider application in this area.

Table 1: Precision, recall and F_1 score of random forests models based on confusion matrices in figure 5.

Materials	Precision	Recall	F_1 score
Li_2CO_3	1.000	1.000	1.000
Li_2O	1.000	0.917	0.957
Li_2O_2	1.000	1.000	1.000
Li_3N	1.000	1.000	1.000
LiH	0.923	1.000	0.960
LiOH	1.000	1.000	1.000
average	0.987	0.986	0.986

4. Conclusion and Future Work

In this study, random forests model is adopted in the analysis of ToF-SIMS data in the field of materials science. It can provide a quick and unbiased approach for detecting characteristic peaks of different materials, which can assist researchers to analyze their complex experimental data. Random forests algorithm can be used to classify different materials and calculate feature importances at the same time, allowing selecting characteristic features and further reducing the amount of required data. Our experiment demonstrates that this method can be successfully used to classify different materials according to their ToF-SIMS spectra, and to select characteristic peaks (mass-to-charge ratio, m/z) of each material.

It is worth noting that data-mining algorithms do not render understanding of fundamental methodology of ToF-SIMS analysis unnecessary, but rather enable a quick and efficient way to help researchers to explore and understand the data they obtained. Combination of domain knowledge in the research area with data-mining methods will greatly benefit the analysis procedure. Furthermore, to achieve a further automation in analyzing ToF-SIMS data, various data pre- and post-processing methods and different machine learning algorithms should be studied to compare their influences on the results.

Acknowledgements

References

[1] Benninghoven A. (1994) "Chemical analysis of inorganic and organic surfaces and thin films by static time-of-flight secondary ion mass spectrometry (ToF-SIMS)." *Angewandte Chemie International Edition in English* **33**: 1023–1043.

[2] Mahoney, C. M. (2010) "Cluster secondary ion mass spectrometry of polymers and related materials." *Mass spectrometry reviews* **29** (2): 247–293.

[3] Bodzon Kulakowska, A., and Suder, P. (2016) "Imaging mass spectrometry: instrumentation, applications, and combination with other visualization techniques." *Mass spectrometry reviews* **35** (1): 147–169.

[4] Yang, J., and I. Gilmore. (2015) "Application of secondary ion mass spectrometry to biomaterials, proteins and cells: a concise review." *Materials Science and Technology* **31** (2): 131–136.

[5] Henss, A., Otto, S. K., Schaepe, K., Pauksch, L., Lips, K. S., and Rohnke, M. (2018). "High resolution imaging and 3D analysis of Ag nanoparticles in cells with ToF-SIMS and delayed extraction." *Biointerphases* **13** (3): 03B410.

[6] Michel, Roger, and David G. Castner. (2006) "Advances in time-of-flight secondary ion mass spectrometry analysis of protein films." *Surface and Interface Analysis: An International Journal devoted to the development and application of techniques for the analysis of surfaces, interfaces and thin films* **38** (11): 1386–1392.

[7] Randau, Simon, et al. (2020) "Benchmarking the performance of all-solid-state lithium batteries." *Nature Energy* : 1–12

[8] Wood, K. N., Noked, M., and Dasgupta, N. P. (2017) "Lithium metal anodes: toward an improved understanding of coupled morphological, electrochemical, and mechanical behavior." *ACS Energy Letters* **2** (3): 664–672.

[9] Zhang Z, Shao Y, Lotsch B, et al. (2018) "New horizons for inorganic solid state ion conductors." *Energy and Environmental Science* **11** (8): 1945–1976.

[10] Graham, D. J., and Castner, D. G. (2012) "Multivariate analysis of ToF-SIMS data from multicomponent systems: the why, when, and how." *Biointerphases* **7** (1): 49.

[11] Kulkarni, P., Dost, M., Bulut, O. D., Welle, A., Boeker, S., Boland, W., and Svatoš, A. (2018) "Secondary ion mass spectrometry imaging and multivariate data analysis reveal co-aggregation patterns of Populus trichocarpa leaf surface compounds on a micrometer scale." *The Plant Journal* **93** (1): 193–206.

[12] Madiona, R. M., Winkler, D. A., Muir, B. W., and Pigram, P. J. (2019) "Optimal machine learning models for robust materials classification using ToF-SIMS data." *Applied Surface Science* **487**: 773–783.

[13] Hanselmann M, Koethe U, Kirchner M, et al. (2009) "Toward digital staining using imaging mass spectrometry and random forests." *Journal of proteome research* **8** (7): 3558–3567.

[14] Alasadi, Suad A., and Wesam S. Bhaya. (2017) "Review of data preprocessing techniques in data mining." *Journal of Engineering and Applied Sciences* **12** (16): 4102-4107.

[15] Keenan, Michael R., and Paul G. Kotula. (2004) "Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images." *Surface and Interface Analysis: An International Journal devoted to the development and application of techniques for the analysis of surfaces, interfaces and thin films* **36** (3): 203–212.

[16] Kohavi R. (1995) "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Ijcai* **14** (2): 1137-1145.

[17] Ho T K. (1995) "Random decision forests." *Proceedings of 3rd international conference on document analysis and recognition, IEEE* **1**: 278-282.

[18] Breiman L. (2001) "Random Forests." *Machine Learning* **45** (1): 5–32

[19] Quinlan, J. R. (1986) "Induction of decision trees." *Machine learning* **1** (1): 81-106.

[20] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984) "Classification and regression trees." *CRC press*.

[21] Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) "Scikit-learn: Machine learning in Python." *Journal of machine learning research* **12**: 2825–2830.