Machine learning and application to spectral analysis on TXRF spectrometry

Makoto Doi* and Shinya Kikuta*, **

1. Introduction

Total Reflection X-ray Fluorescence (TXRF) analysis is a non-destructive and surface-sensitive analysis method using X-rays[1, 2], in which incident X-rays are irradiated on a sample at an extremely low grazing angle (about 0.1°) and the fluorescent X-rays from the sample generated by the incident X-rays are measured with extremely low background because of the total reflection characteristics of the incident X-rays. TXRF analysis does not require special sample preparation for flat samples. Because of this, TXRF analysis has been widely used for the evaluation of contamination on wafers in semiconductor manufacturing processes[3] as well as in industrial and environmental analysis. Contamination control in semiconductor manufacturing processes becomes more rigorous every year.

Contaminants are usually distributed inhomogeneously on wafers in semiconductor manufacturing processes. High concentrations of contaminants (more than 10^{10} atoms/cm^2) can exist locally on a wafer, even if the average contamination over the wafer surface is within allowance. This fact causes a decrease in production yield. The "SWEEPING-TXRF"[4, 5] method has been applied to those cases and established as one of the most effective methods. In this method, the TXRF measurements are performed over the entire surface of wafers in a short time, resulting in high-speed mapping. Figure 1 shows an example of the result of SWEEPING-TXRF, where a wafer was analyzed for the verification of contamination after the cleaning process. Sulfur and chlorine were detected over the entire surface. These elements are derived from the residues of chemical cleaning liquids, such as sulfuric and hydrochloride acid. Figure 1 also indicates that K, Ca, Ti, Cr, and Zn were detected on several areas, including the edges of the wafer, whereas Fe appeared all over the surface of the wafer. SWEEPING-TXRF can give not only the analysis results for each position on the wafer but also the average quantities over the surface. It is a useful feature that it is possible to analyze the contamination condition in detail and control the manufacturing processes from various aspects regarding contamination by combining these results.

In TXRF, an energy spectrum, which is composed of scattered incident X-rays and the fluorescent X-rays generated by the elements on the wafer surface, is obtained according to the contaminants and their amounts. There are multiple peaks in the spectrum, and the analysis is performed by the peak fitting method to obtain intensities of individual peaks using deconvolution.

It is easy to obtain good results using the peak fitting method when peak intensities are large using regular measurement times. However, trace element analysis is more complicated when peak intensities are low and measurement times are short, making it difficult to distinguish signals due to fluorescent X-rays from noise. In particular, from the viewpoint of throughput, the required time for each point for SWEEPING-TXRF measurement of the entire surface of the wafer is only 5 to 10 seconds. Therefore, the conventional peak fitting method sometimes results in false-positive or false-negative detections for such tiny peaks due to noise and statistical errors in the X-ray counts.

Recently, Artificial Intelligence (AI) technologies have developed rapidly along with progress in computer hardware, software and software libraries to deal with big data. One main benefit of AI is that it automatically extracts and analyzes unique and notable characteristics from a huge amount of data. In the field of image processing, particularly, image recognition[6]—for example, handwritten character recognition—has been actively researched and many results—such as super-resolution techniques[7] that convert low-resolution images to high-resolution ones—have been achieved. Although there are many cases where AI is used for image processing, it seems that there are few cases where AI technologies are applied to one-dimensional spectrum analysis instead of to a two-dimensional image. Therefore, in this paper, we applied the machine learning method to the data processing of TXRF analysis and introduce the results, especially on the quantification of contaminations on wafers from the spectrum obtained by short-time measurements.

2. Machine Learning Method

2.1. Training data set and training process in Machine Learning

Machine Learning (ML) is a technology to learn the correlations among a large amount of data, discover some patterns and rules from them and make various discriminations and predictions by using them. In this study, we used the ML technique classified as supervised learning. In this technique, a learning algorithm analyzes

* Rigaku Corporation.
** Co-author Mr. Shinya Kikuta had passed away suddenly before submission of this paper. Another author, Makoto Doi, would like to express our sincere gratitude for his great work to this research.

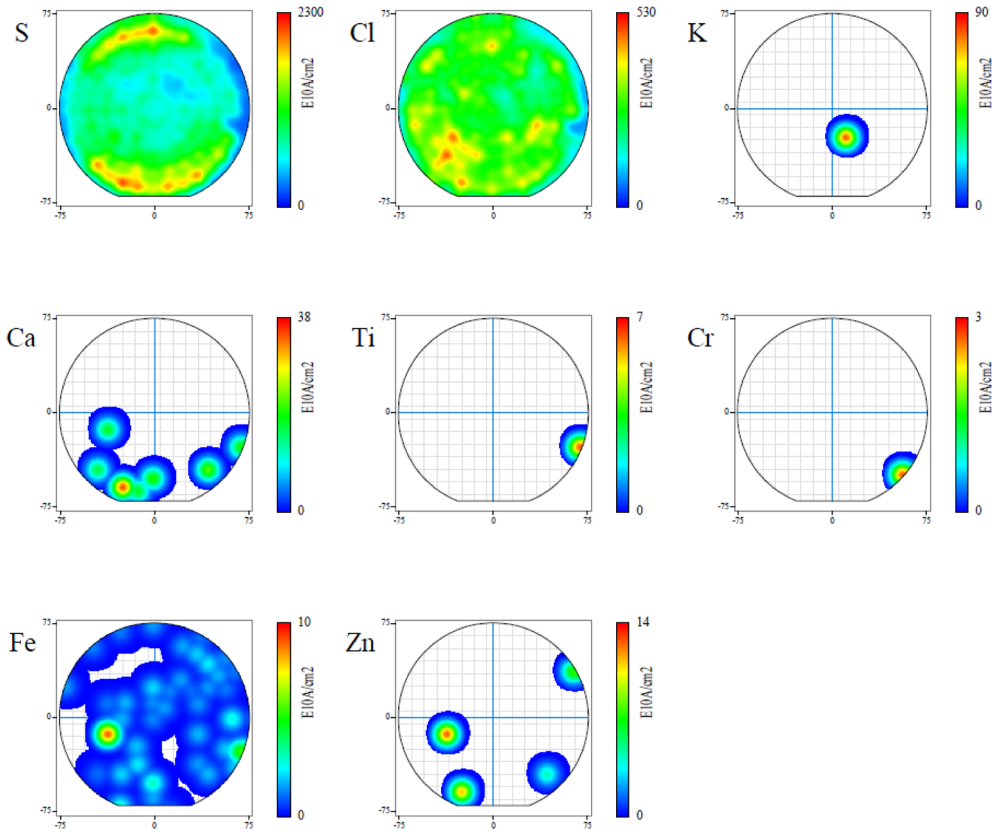

Fig. 1. Mapping results by SWEEPING-TXRF.

a training data set (supervised data set) and produces a mapping function that associates a measured spectrum as input with an elemental quantified result as output involved in each sample in the training data set. Then the trained algorithm is applied to a validation data set, which is independent from the training data, and the output results derived by the trained algorithm are checked. The known output results of the training data and the validation data were prepared by the conventional peak fitting method from a measured spectrum using a longer measurement time at the same point where the input spectrum of the training data and the validation data were obtained for a shorter measurement time.

The ML model used in this study is a convolutional neural network model constructed with one convolutional layer and six full connected layers (Fig. 2). Calculations for this ML model are performed using TensorFlow[8] as a numerical software library. The X-ray intensities corresponding to individual energies—i.e. individual channels in a multichannel analyzer—of the measurement spectrum are assigned to the input layer, and the evaluated values of the target elements are obtained from the output layer. ReLU (Rectified Linear Unit)[9], $f(x)=\max(0, x)$, is used for the activation function of each layer in the ML model except for the last output layer. The output values cannot be negative since they are X-ray intensities from elements. Therefore, the following function (1) was used as the activation function of the last layer.

$$f(x)=\log(1+\exp(x)) \tag{1}$$

This function asymptotically approaches $f(x)=x$ when the variable x is positive large, and approach $f(x)=0$ from a positive value when x becomes negative large.

Training is performed by comparing the quantitative values of the target elements obtained by analyzing the long-time measurement data—which we call the "true values" here—with the calculated results from the ML model. Optimization of the parameters that make up the neural network proceeds so as to minimize the loss function. The loss function is defined as the sum of squares of the differences from the true values through equation (2).

$$\text{Loss}=\sum w_i\left(I_i-I_i^0\right)^2 \tag{2}$$

where I_i^0 and I_i represent the true value and the calculated value from the ML model respectively. When the deviations from the true values follow Gaussian distribution, minimizing this loss function drives the maximum likelihood estimation for the distribution of I_is (Maximum Likelihood method). The actual fluctuations of the detected X-rays are, however, not in Gaussian distribution; when the counted amounts of the X-rays are small, they are in Poisson distribution, and when the X-ray intensities are so high that a count-loss correction (dead time correction) is needed, they are neither in Poisson distribution nor Gaussian distribution[10), (11]. Therefore the minimization of the loss function expressed in the eq.

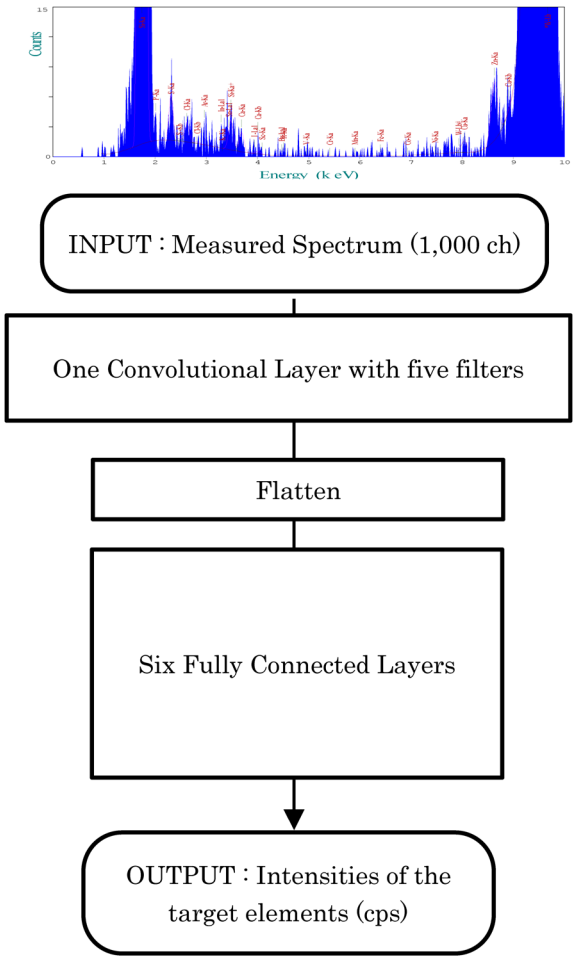

Fig. 2. Convolutional Neural Network Model used in this study.

(2) does not give the maximum likelihood estimation, but we used it here since it is easy to analyze by computing.

Next, we discuss the weights, w_is, in the summation of eq. (2). X-ray generation is a random event and the number of counted X-ray photons fluctuates around the mean value. The standard deviation of the counted photons is proportional to the square root of the X-ray intensity.

$$\sigma_i \propto \sqrt{I_i^0} \tag{3}$$

The square of a residual, $(I_i - I_i^0)^2$, is evaluated with the uncertainty of the square of the standard deviation, σ_i^2. Therefore, it is adequate to define the weights as the inverse of the square of the standard deviation, $w_i = 1/\sigma_i^2$. Then the loss function is as follows:

$$\begin{aligned} \text{Loss} &= \sum w_i \left(I_i - I_i^0\right)^2 \\ &= \sum \frac{\left(I_i - I_i^0\right)^2}{\sigma_i^2} \propto \sum \frac{\left(I_i - I_i^0\right)^2}{I_i^0} \end{aligned} \tag{4}$$

The AdaMax method[12], which is a stochastic optimization algorithm, was employed to search the optimized parameters to minimize the loss function defined above.

The training data set and the validation data set were obtained from the results of measuring various Si wafers using Rigaku TXRF spectrometers TXRF3760 and TXRF-V310[13].

2.2. Target elements and Samples

Rigaku TXRF3760 and TXRF-V310 spectrometers are capable of analyzing from Na to U using three X-ray excitation lines. In this study, only W-Lβ X-rays from three excitation lines were considered and the target elements are Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Sn and Ba, which (except for Si, which originates from the Si wafer) are important elements for contamination analysis. The W-Lβ X-ray line, the excitation X-ray line, was also studied. It is necessary to prepare training data on samples that include various amounts of the target elements in order to train the ML model. The intensities of the X-rays of the Si line from the Si substrate and the W-Lβ X-rays scatter line from the excitation X-rays are much higher than the fluorescent X-ray intensities of the target elements. If Si and W-Lβ X-rays are considered in the same way as the target elements during training, the parameters in the ML model would be fixed by optimizing only these large intensity values; therefore, these intensities were reduced by a factor of 1000 when the ML model was trained.

3. Results and Discussion

3.1. Training and Results on Rigaku TXRF 3760

Using Rigaku's TXRF3760, we acquired spectral data consisting of 15,277 sample points with a measurement time of 5 seconds each, and also acquired the same data points with a time of 60 seconds. Each spectrum was analyzed and quantified by the conventional peak fitting method. From the 15,277 sample data, we randomly extracted 13,749 sample data as the training data. The remaining 1,528 sample data were used as the validation data.

First, the results for the validation data measured in 5 seconds, analyzed by the conventional method, are compared with those measured in 60 seconds in the correlation graphs of Fig. 3(a). The horizontal and vertical axes in Fig. 3(a) are for the results for 60 seconds and 5 seconds, respectively. For many elements, the results for 5-second measurements are close to the results for the 60-second measurement, but the correlations become poorer for the lower X-ray intensity cases. This means that the data from the 5-second measurement are not analyzed accurately for the low-intensity cases because of statistical errors and noise in the small peaks. Figure 3(b) shows correlation graphs between the results for 5-second measured data calculated by the trained ML model and the results of the conventional method for 60-second measured data. As with the results shown in Fig. 3(a), correlation is good for the high X-ray intensities. Furthermore, the

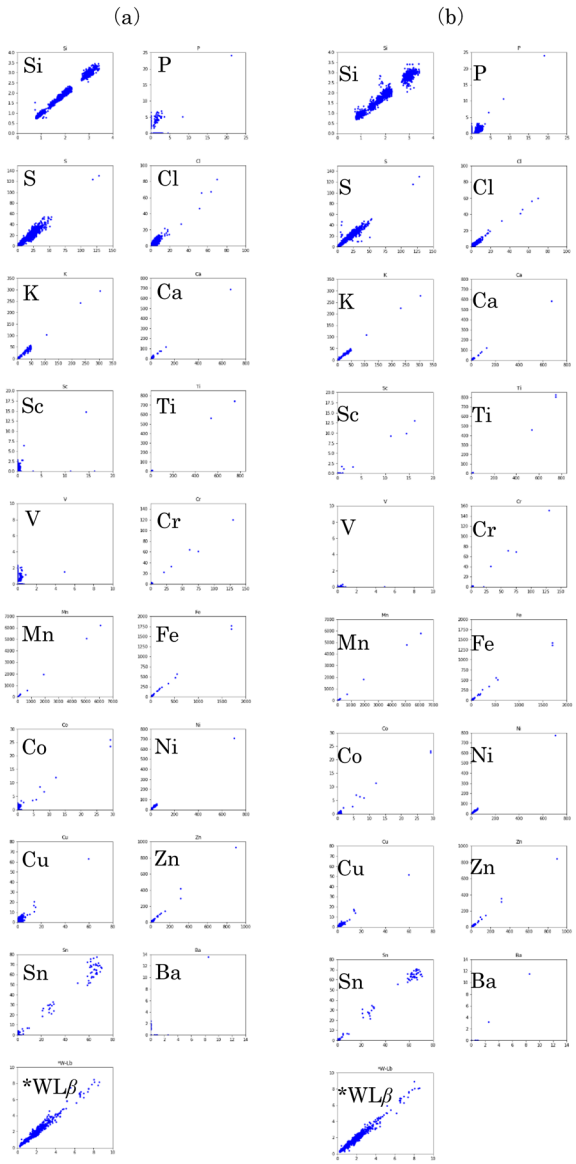

Fig. 3. Correlation plots of the analyzed results on 5-second measurements vs. 60-second measurements by TXRF 3760. (a) Conventional analysis, (b) ML analysis.

correlations seem to be good even when the X-ray intensities are low. The correlation coefficients from those graphs are summarized in Table 1, and the graphs plotted with those results are shown in Fig. 4. The horizontal axis in Fig. 4 is the correlation coefficients obtained by the conventional method (Fig. 3(a)) and the vertical axis are those obtained by the trained ML model (Fig. 3(b)). The correlation coefficients were improved by the ML analysis for most elements.

In TXRF analysis, it is important to obtain accurate results for high X-ray peak intensities, but it is more important to not produce false detections when the X-ray intensities are low. In Table 2, we summarized the total numbers of cases where the results of analyzing the 5-second measurement spectra were 1 cps or less even though the 60-second measurement results were more than 1 cps (false-negative), and cases where the 5-second measurement results were more than 1 cps but

Table 1. Correlation coefficients of the plots in Fig. 3.

Element	Conventional	ML
Si	0.993	0.974
P	0.442	0.795
S	0.954	0.960
Cl	0.915	0.975
K	0.994	0.997
Ca	0.997	0.999
Sc	0.342	0.973
Ti	1.000	1.000
V	0.088	0.085
Cr	0.991	0.987
Mn	1.000	1.000
Fe	0.999	0.997
Co	0.886	0.986
Ni	0.998	0.999
Cu	0.867	0.967
Zn	0.995	0.998
Sn	0.993	0.998
Ba	0.855	0.978
*W-Lβ	0.986	0.985

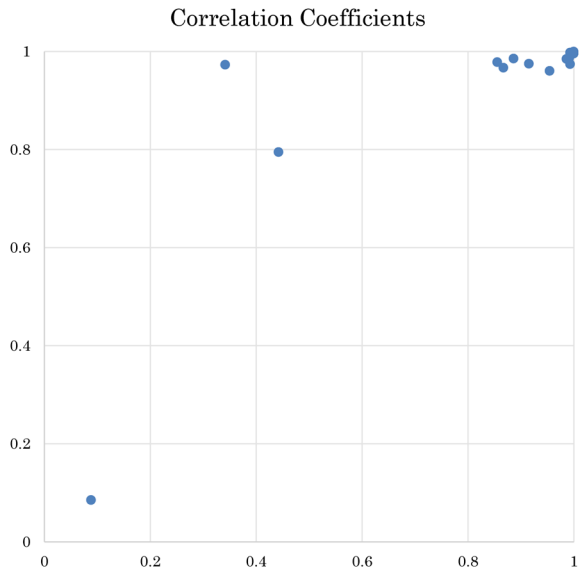

Fig. 4. Correlation coefficient of each element; x-axis: conventional result, y-axis: ML results.

the 60-second measurement results were 1 cps or less (false-positive). The number of false detection cases in the analyses using the ML model is significantly reduced compared to the conventional analysis method.

3.2. Results and discussion on Rigaku TXRF-V310

In subsection 3.1, the training data were acquired using the TXRF3760 and the validation data were acquired by the same spectrometer as well. In this subsection, we studied the application of the ML model trained with data from the TXRF3760 to data acquired by the TXRF-V310. When applying the trained ML model to others, in general, there are three methods; (1) applying it as it is, (2) re-training all the parameters of

Table 2. The number of false detection cases from the validation samples from 1,528 points.

Element	Conventional	ML
P	775	450
S	68	18
Cl	306	114
K	241	23
Ca	256	38
Sc	153	2
Ti	110	3
V	140	1
Cr	103	7
Mn	120	16
Fe	354	111
Co	165	6
Ni	99	27
Cu	431	93
Zn	282	94
Sn	182	19
Ba	15	0

the ML model using a small number of training data to be applied (Fine Tuning) and (3) re-training only the parameters involved in some layers of the model using a small number of training data to be applied (Transfer Learning). In the latter case, parameters other than those to be re-trained remain fixed. Methods (1) and (2) are discussed below.

We prepared a data set consisting of 1,634 sample points measured by the TXRF-V310 and used them as validation data. Figure 5(a) shows the correlation between 5-second measurement results and 60-second results, both analyzed by the conventional method. The "Conventional" columns in Table 3 and Table 4 show the correlation coefficients of those and the number of false detection cases in the 5-second measurement analysis, respectively, where the number of the false detection cases is defined in 3.1.

We prepared three types of trained ML models: ML1, ML2 and ML3. ML1 is the trained ML model described in 3.1 as is. ML2 is the fine-tuned ML model using 1,000 sample data points acquired by the TXRF-V310

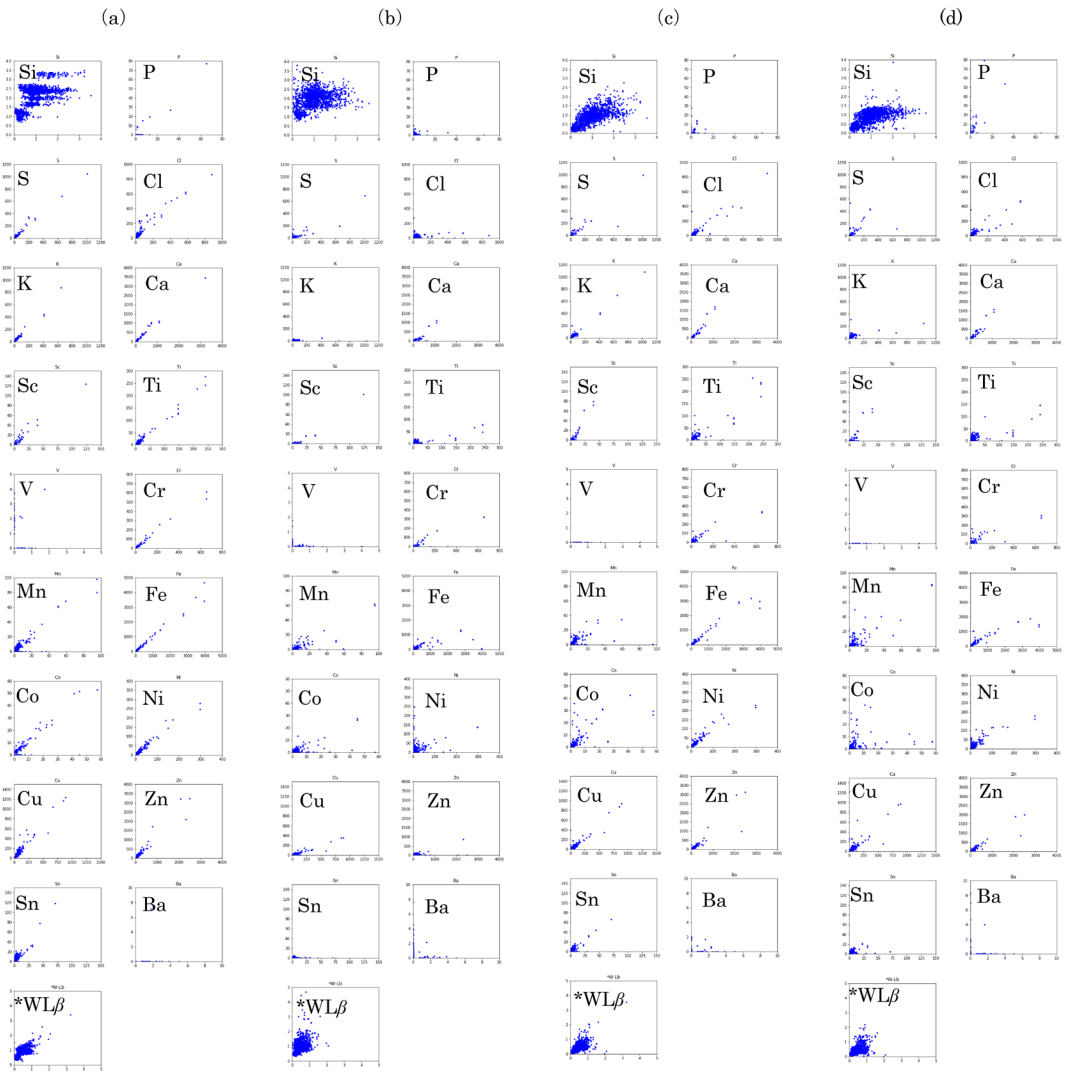

Fig. 5. Correlation plots of analyzed results measured by TXRF-V310. X-axes: 60-second measurement results, Y-axes: 5-second measurement results by (a) Conventional method, (b) ML1, (c) ML2 and (d) ML3. ML1, ML2 and ML3 are mentioned in detail in Subsect. 3.1.

Table 3. Correlation coefficients of each element on the plots of Fig. 5.

Elements	Conventional	ML1	ML2	ML3
Si	0.590	0.135	0.685	0.581
P	0.935	0.061	0.373	0.387
S	0.984	0.893	0.878	0.855
Cl	0.957	0.364	0.947	0.883
K	0.994	0.303	0.988	0.649
Ca	0.993	0.924	0.959	0.959
Sc	0.965	0.945	0.970	0.952
Ti	0.986	0.859	0.917	0.810
V	0.694	-0.003	0.015	-0.002
Cr	0.992	0.910	0.883	0.851
Mn	0.929	0.843	0.482	0.700
Fe	0.992	0.654	0.970	0.905
Co	0.904	0.566	0.629	0.078
Ni	0.980	0.467	0.950	0.670
Cu	0.964	0.929	0.915	0.767
Zn	0.969	0.549	0.932	0.943
Sn	0.876	0.144	0.909	0.576
Ba	0.169	0.029	0.102	0.009
*W-Lβ	0.667	0.471	0.621	0.535

Table 4. The numbers of false detection cases from the validation samples from 1,634 points.

Element	Conventional	ML1	ML2	ML3
P	52	271	54	461
S	448	293	177	62
Cl	420	397	300	261
K	147	176	111	387
Ca	112	80	70	128
Sc	78	70	47	81
Ti	131	247	78	59
V	26	8	6	107
Cr	156	252	118	6
Mn	121	154	71	150
Fe	71	102	37	112
Co	119	110	71	52
Ni	156	169	103	103
Cu	79	106	58	117
Zn	124	103	71	80
Sn	309	436	193	87
Ba	16	43	20	221

independently from the above validation data. ML3 is also the fine-tuned ML model but it is tuned by only 100 sample data points. The base model of ML2 and ML3 were the trained ML model described in 3.1 using the data from the TXRF3760. The correlation plots, correlation coefficients and the number of false detection cases for ML1, ML2 and ML3 are shown in Fig. 5 (b), (c), (d) and in Tables 3 and 4.

Regarding the correlation coefficients for 60-second measurement results, both ML1 and ML3 gave generally inferior results to the conventional method, but ML2 gave results close to the conventional method results for most elements. ML2 definitely showed better results related to the number of false detection cases compared to the conventional method, and even in ML3 the results are not inferior to the conventional method on the whole.

Based on the results above, the ML method is an effective method for preventing false detections with short-time measurement. To obtain quantification results with a certain accuracy—which means obtaining good correlation results with those from long-time measurements—it is necessary to fine-tune the model with about 1,000 sample points of training data. Furthermore, we confirmed that the correlation coefficients and the number of false detection cases were greatly improved compared to the conventional method when the model was fine-tuned by 10,000 sample points of training data obtained by TXRF-V310, although this is not shown in this paper.

4. Conclusion

In this paper, we first introduced SWEEPING-TXRF, which is being used in actual semiconductor manufacturing processes, and stated that the demand for higher sensitivity for TXRF is increasing year by year due to technological progress in semiconductors. We then investigated the possibility of improving the detection performance for minor peaks in TXRF spectra, especially in short-time measurements using a Machine Learning approach.

We used the Convolutional Neural Network model as the ML model, which was classified as supervised learning. Each training data set consisted of input data and output data. The input data was the spectrum obtained by the TXRF measurement for 5 seconds. The output data consisted of the analyzed results obtained by measuring the same points for 60 seconds by the conventional peak fitting method for each element. After training the ML model, we applied it to the validation data prepared independently of the training data set and evaluated the results qualitatively and quantitatively.

First, we confirmed that the ML model trained by 10,000 or more sample data points showed good performances and gave good correlations with the analyzed results of the data of long-time measurements.

Next, we applied the ML model trained above to data from another spectrometer. When it was not fine-tuned or it was fine-tuned with a small number (e.g., 100) of sample points, correlations with the results of long-time measurements were inferior to those by the conventional method. However, some improvements were obtained in the number of false detection cases even with 100 sample data points. When 1000 or more sample data points were used to fine-tune the trained ML model, the ML results gave good performance both on the correlation coefficients and the number of false detection cases.

From the results above, we propose a hybrid method combining the ML method with the conventional method; that is, small peaks are analyzed by ML to reduce the false detection cases and large peaks are done

by the conventional method to quantify the amounts accurately, when the number of the re-training data samples are not many. Alternatively, the results of the ML method can be used as initial values for analyzing data by the conventional method. Analyzing by the hybrid method will be a challenge in the future.

We investigated the effectivities of using only the fine-tuning method to apply the trained ML model to the data from another spectrometer in this paper. The effects of the Transfer learning method will also be investigated in future.

References

(1) H. Aiginger: *Spectrochim. Acta Part B*, **46** (1991), 1313–1321.
(2) H. Schwenke, P. A. Beaven and J. Knoth: *Fresenius J. Anal. Chem*., **365** (1999), 19–27.
(3) D. Hellin, S. D. Gendt, N. Valckx, P. W. Mertens and C. Vinckier: *Spectrochim. Acta Part B*, **61** (2006), 496–514.
(4) Y. Mori: *Adv. X-ray Anal*. **45** (2002), 523–532.
(5) Y. Mori, K. Uemura, H. Kohno, M. Yamagami, H. Shimizu, Y. Onizuka and E. Iizuka: *Adv. X-ray Chem. Anal., Japan*, **36** (2005), 274–284 [in Japanese].
(6) A. Krizhevsky, I. Sutskever and G. E. Hilton: *Proceedings of the 25th International Conference on Neural Information Processing Systems*, **1** (2012), 1097–1105.
(7) C. Dong, C. C. Loy, K. He and X. Tang: *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2016), 2 295–307.
(8) TensorfFlow: An open source machine learning library for research and production. ⟨https://www.tensorflow.org/⟩.
(9) R. L. T. Hahnloser: *Neural Networks*, **11** (1998), 691–697.
(10) J. W. Muller: *Nucl. Instrum. Methods*, **112** (1973), 47–57.
(11) K. Omote: *Nucl. Instrum. Methods Phys. Res. Sect. A*, **293** (1990), 582–588.
(12) D. P. Kingma and J. Ba: Adam: A Method for Stochastic Optimization. (2015), arXiv:1412.6980v9.
(13) Rigaku Corporation: ⟨https://www.rigaku.com/fields-materials/semiconductors⟩.