OPEN

Machine learning for pattern and waveform recognitions in terahertz image data

Dmitry S. Bulgarevich^{1,2✉}, Miezel Talara², Masahiko Tani² & Makoto Watanabe¹

Several machine learning (ML) techniques were tested for the feasibility of performing automated pattern and waveform recognitions of terahertz time-domain spectroscopy datasets. Out of all the ML techniques under test, it was observed that random forest statistical algorithm works well with the THz datasets in both the frequency and time domains. With such ML algorithm, a classifier can be created with less than 1% out-of-bag error for segmentation of rusted and non-rusted sample regions of the image datasets in frequency domain. The degree of linear correlation between the rusted area percentage and the image spatial resolution with terahertz frequency can be used as an additional cross-validation criteria for the evaluation of classifier quality. However, for different rust staging measured datasets, a standardized procedure of image pre-processing is necessary to create/apply a single classifier and its usage is only limited to 1 ± 0.2 THz. Moreover, random forest is practically the best choice among the several popular ML techniques under test for waveform recognition of time-domain data in terms of classification accuracy and timing. Our results demonstrate the usefulness of random forest and several other machine learning algorithms for terahertz hyperspectral pattern recognition.

Terahertz/gigahertz time-domain spectroscopy (THz/GHz-TDS) is a promising nondestructive testing (NDT) technique for applications in various materials due to its harmless nature, longer radiation penetration depth for various materials, and the possibility of getting a rich structural and complex dielectric information in time and frequency domains^{1,2}. THz/GHz-TDS imaging is inherently a hyperspectral technique with possibility of sliced imaging in frequency as well as in time domains since each image pixel contains a THz/GHz waveform^{3,4}. In this respect, the typical spectral resolution of a classical THz-TDS system with mechanical delay stage is limited to ~ 0.8 GHz at ~ 0.5 THz due to the following reasons: (1) the duration of the temporal window brought by the laser repetition rate, (2) the length of the delay stage, and (3) the noise introduced by the fluctuations in THz radiation/detection⁵. In addition, time resolution is limited only by the fs-pulse duration and minimum mechanical delay stage step. Moreover, the spatial image resolution is governed by the Abbe diffraction limit with used THz wavelength and employed optics. In spite of the very attractive benefits and potentials of THz/GHz-TDS imaging, the associated large image data volumes and complex image contrasts could make the analysis and interpretation very difficult and time consuming.

In recent years, dramatic progress has been made in automated image pattern recognition techniques for automobile⁶, medical⁷, biological⁸, agricultural⁹, IT¹⁰, security¹¹, and other applications. With suitable algorithm, image database, and powerful computers, the large volumes of image data can be classified in a short time with high accuracy^{12,13}.

Apart from spectral recognition using various techniques in previous works^{14–18}, statistical pattern recognition with Feed-Forward Artificial Neuronal Network (ANN), Convolutional Neural Networks (CNN), and Gaussian Mixture Model (GMM) were applied for sorting of recycling plastics by using THz frequency domain data¹⁹. Classification of breast cancer, if benign or malignant, from biomedical THz images with ANN and *K*-Nearest Neighbor (KNN) pattern recognition algorithms was also reported²⁰. Just recently, KNN, Support Vector Machine (SVM), and Random Forest (RF) were compared in terms of total classification accuracy and recognition ability for the diagnosis of traumatic brain injury and results showed that RF was the best algorithm among the three algorithms under study²¹. In a previous study²², SVM was observed to be effective in distinguishing patterns associated with ribonucleic acid and various powdered substances hidden inside noisy biomedical measurements. Minimum Distance classifier and ANN methods were implemented for identification of explosives and

¹Research Center for Structural Materials, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan. ²Research Center for Development of Far-Infrared Region, University of Fukui, 3-9-1 Bunkyo, Fukui 910-8507, Japan. ✉email: BULGAREVICH.Dmitry@nims.go.jp

bio-chemical materials from THz images at different frequencies²³. Moreover, a deep learning algorithm was used for classification of THz images^{24,25}. In addition, a simple neural network for resolving of coating thickness²⁶ and a back-propagation (BP) neural network and SVM for conducting regression analysis for the characterization of thermal barrier coatings with THz-TDS were realized²⁷. Besides ML algorithms²⁸, various segmentation methods^{29,30} were applied for automated detection of concealed objects from THz images.

However, there is still very limited literature on the application of ML techniques in nondestructive testing (NDT) using pattern or waveform recognition in THz-TDS datasets. Just recently, a classical BP neural network algorithm and a novel extreme learning machine (ELM) algorithm were employed for NDT of measuring film thickness from THz time domain data³¹. In this respect, we demonstrate the feasibility of pattern recognition in THz-TDS images of rusted steel with RF machine learning algorithm^{32,33}. Such analysis was conducted in frequency domain which is more widely used in NDT applications due to higher achievable spatial resolution compared to time domain imaging. Time domain imaging is typically employed for NDT measurements of sample thickness. Hence, the possibility of THz waveform recognition is also outlined with various ML methods. Here, it should be noted that applications of such techniques are novel for THz-TDS imaging, NDT, and materials science fields.

Experimental

For pattern recognition, sample preparations were carried out by wetting the steel plate, with cross-cuts in paint layer, with 3% NaCl solution and sealing it in container with 100% humidity at 23 °C to rust. The prepared samples were periodically taken out and dried naturally. THz-TDS imaging was performed by using a broadband Pulse IRS-2000 (AISPEC) THz-TDS system. The linearly polarized THz beam was focused on the sample surface using an optical reflection stage ($\approx \pm 10.3^\circ$ maximum cone angle). The THz beam profile with our THz-TDS system was previously reported³⁴. The samples were then raster scanned with 0.2×0.2 mm² spatial pixel resolution using a mapping mechanical stage in the focal plane of such reflection stage. The waveforms were collected for spectral analysis with 4096 data points at ~ 4.2 fs time intervals, i.e. with ~ 58 GHz spectral image resolution. Note that we studied the non-hidden rust patterns with THz-TDS in order to compare the image segmentation results with that of visible light optical images. For waveform recognition experiments, we used the aperture array of Siemens-star apertures microfabricated in thin-film Al on a thick Si substrate, which was previously described in literature³⁵.

The open source FIJI software package with Trainable Weka Segmentation plugin was applied for image dataset pre-processing, analysis, and segmentation on two-CPU 6128 Opteron Workstation with 128 GB RAM^{36–40}. ML on training/test datasets with THz waveforms was performed by using the open source WEKA software package with collection of ML algorithms⁴¹.

Results and discussion

Machine learning in frequency domain. Figure 1 shows the general outline of the RF algorithm for image segmentation on several classes (j), i.e. on rusted and non-rusted regions in our images. The basic idea is to generate the forest from individual decision tree (k) classifiers $[h(\mathbf{x}, \Theta_k)]$, which will store the information on appropriate sequences of image filter applications, optimized threshold parameters for split function $[S(j)]$ and leaf node highest probabilities (P) used to assign each image pixel to particular j . Here, Θ_k are the feature vectors obtained by transforming the image input vectors ($\mathbf{x}$) at each k node with image filters/kernels. The collection of such $[h(\mathbf{x}, \Theta_k)]$ from all k is the RF classifier $[\{h(\mathbf{x}, \Theta_k); k = 1, \dots\}]$.

In this ensemble-type supervised learning, the user has to choose the image or images from the dataset for training process in order to create the RF classifier: the program for automatic segmentation/classification of other images with the same j present. This step requires several user interactions: providing the training data set (T) with good examples of image areas/pixels for every j , forming the image filter dataset, and defining the forest scale such as number of trees, their maximum depth or calculation precision. In RF algorithm, the bootstrap data (T_k) for each decision tree (k) construction are chosen from T at random, but with overlays. During this process, $\sim 1/3$ of T are left unused to form the out-of-bag (OOB) dataset for later estimation of RF classifier error. Note that levelling-off the classification accuracy for forests with more than ~ 100 – 200 trees^{13,42,43} is well-documented.

At the classification stage on training image/images, the pixels that arrives to L_k after all filtering and binary tests stages are assigned to $P(j|L_k)$ in each k , which were obtained at k construction stage. To make a decision on j assignment for each pixel, every k casts a unit vote for the most popular class. These votes from all k cast a unit vote too. As a result, each pixel is finally classified to the particular j . Then, the user can optimize the RF classifier by changing/adding the training areas, applied filters, and forest parameters to minimize the OOB error and to improve the visual segmentation quality. Finally, the RF classifier can be saved and used on other images. The ensemble effect and de-correlation of trees due to random bootstrapping and filtering make the RF classifiers unbiased and very stable to image noise compared to other statistical methods. However, they may also have some limitations, which will be discussed below.

Figure 2a shows the optical and THz images of the initial and rusted samples, respectively. To create the RF classifier with OOB error below 1% for particular rusting time, it is necessary to train it on THz images with different spatial resolutions (THz frequencies) due to different noise and contrast levels within the image spectral dataset. Then, T with examples of two classes (rust and paint) is selected from images at ~ 0.3 , ~ 1 , and ~ 2 THz. Figure 2b demonstrates the difference between several RF classifiers built with different pre-processing protocols to address such noise and contrast variations as well as with different number of used image filters for particular RF buildup. In the case of classifier with 4 filters, the following ones were employed: Gaussian Blur, Hessian, Sobel, and Difference of Gaussians. For the case with 17 filters, Membrane Projections, Mean, Maximum, Anisotropic

Figure 1. Schematic diagram of the principle behind RF classifier for pattern and waveform recognitions.

Diffusion, Lipschitz, Gabor, Entropy, Variance, Minimum, Median, Bilateral, Kuwahara, and Neighbors filters were added.

As shown in Fig. 2b, not only the absolute values of rusted area percentage at fixed frequencies were different between classifiers, but the shape of frequency dependence also varies. The minimum variations between classifiers were observed at ~ 1 THz since spatial resolution or signal-to-noise ratio decreased at lower or higher frequencies, respectively. In principle, due to the nature of RF, the use of larger number of different filters/kernels for Θ_k should lead to better quality classifier. However, by using OOB errors and visual inspection alone, it is difficult to judge whether the improvements of classifier quality depend on the preprocessing protocols or the number/type of image filters being used since segmentations seemed to be good anyway.

In this respect, the correspondence between frequency dependences of rusted area percentage and optical resolution could serve as an additional cross-validation criterion for classifier quality (see comparison in Fig. 2b and in insert). Image pre-processing protocols which include (1) background brightness correction, (2) background correction with sliding paraboloid of 150 pixel radius, (3) noise reduction by image smoothing with Gaussian blur filter ($\sigma = 1$), and (4) contrast normalization with analysis of the image histograms as well as the training of classifiers with 17 image filters³⁹ have resulted to reasonable classifier quality (see plots with Preprocessing-2 in Fig. 2b).

The performance of the resulted RF classifier is compared in Fig. 2c,d with the classifier built on the corresponding optical image. The difference in the obtained rusted area percentage estimated from both the optical and THz images was reasonable due to the higher optical resolution with visible light. Currently, the most accurate results were obtained at ~ 1 THz (see Fig. 2b). Note that the same RF classifier (with Preprocessing-2) created on a 4-month rusting dataset was applied on the THz image datasets for 1-month rusting time. This demonstrates the feasibility of creating a single classifier for analyses of different rust staging THz image datasets, which is practically useful for NDT.

Machine learning in time domain. Figure 3a shows the basic experimental setup used for THz-TDS imaging of Al aperture array described elsewhere³⁵. For ML, 500 waveforms across a 12×12 mm array on Si wafer (30×30 mm) in air were used. Each waveform with 16,384 data points corresponds to a single pixel with 0.1×0.1 mm spatial size from image line of 500 pixels. These waveforms belong to three classes: Air, Si wafer, and Aperture array. Such sample was used for our proof of concept since these classes have distinct waveforms as shown in Fig. 3b. For classifier training, 16 waveforms of each class were selected, which formed the training dataset.

Table 1 compares the performance of several ML techniques on such training dataset with tenfold stratified cross-validation, i.e. it invokes the learning algorithm 11 times, once for each fold and then a final time on the entire training dataset. The classifier settings were based on known behaviors of these algorithms. Note that diagonal and off-diagonal elements in the provided confusion matrixes correspond to the number of correct and incorrect classifications, respectively. Among the tested classifiers, RF demonstrated not only 100% correct

Figure 2. Analysis with ML in frequency domain. **(a)** Overlays of reflection THz and auto-segmented images of painted, cross-scratched, and subsequently rusted steel sample showing colored rusted and painted regions. The gray pictures are optical images of non-rusted sample and reflection THz images used for ML classifier training. The indicated frequencies correspond to a particular image in the spectral dataset. **(b)** Frequency vs % Rusted Area comparison of the different ML pre-processing conditions showing the classifier quality depends strongly on user-supervised training for lower OOB error, better overlays and correlation with optical resolution. **(c)** Sample optical, THz and segmented images showing good agreement in classifying rusted and painted regions. **(d)** Graphical representation of the good agreement between the optical resolution and the classifier quality.

classification of training dataset, but also exhibited the best overall training/testing time of the entire dataset with 500 waveforms.

Figure 3c shows the plot of attribute counts from 5000 RF trees, which reveals the internal attribute selection in the process of basic RF classifier build-up. The attributes between ~2–15 ps have higher counts in RF trees since in this delay time range, largest amplitude differences between waveform classes are present. The insert in Fig. 3c provides example of one logical tree from RF with used time attributes, selected thresholds, and assigned classes.

Figure 3d demonstrates the correspondence between RF estimated and experimental class boundaries. The bold colored line is RF results. Below this line, the waveform image is shown for comparison. Each pixel intensity in vertical column is the amplitude value at particular delay time for a single waveform. The 500 columns correspond to the number of waveforms in the test dataset. As it can be seen, RF placed the class boundaries quite accurately at diffraction boundaries in the experimental THz image.

To compare the accuracy of estimated class boundaries between different ML techniques, Fig. 4 plots class assignment for each waveform from the test dataset. In Fig. 4, each vertical color bar corresponds to the single waveform from each pixel in the image line, which is then assigned to a particular class with well-trained ML classifier. The vertical offsets were added for clarity. The k -M classifier exhibited the worst classification on train/test datasets for $k = j$ while the best classification was obtained for $k = 5$, i.e. for $k > j$ (see Table 1). However, this classifier is still behind other tested ML methods in terms of classification performance. Moreover, NB, which is another ML algorithm being tested in this study, also demonstrated poor accuracy in spite of good training

Figure 3. Analysis with ML in time domain. **(a)** Schematic diagram for waveform collection showing the Al aperture array filter on Si substrate configuration in the THz-TDS imaging system. **(b)** Examples of transmission waveforms taken from THz-TDS image of the Al aperture array filter on Si substrate. For classifier training, 16 waveforms for each class (Air, Si, and Array) were used. **(c)** Time attribute vs. Attribute count graph showing attribute importance evaluation with RF on train dataset. **(d)** Comparison of class assignments by RF classifier (top line image) and waveform amplitude image of the test dataset (lower grey image). The latter one was obtained by converting all 500 waveforms across the THz-TDS image (line of 1 pixel height) to the text image matrix with waveform amplitude value vs time in each column (waveform image). The dotted vertical lines are RF-estimated class boundaries.

Classifier name	Classifier main used/optimized parameters	Tenfold validation on train dataset				Time (s)																	
		Confusion matrix, correct classification				Training	Testing																
Random Forest (RF)	100 trees, int(log ₂ (features) + 1) randomly chosen attributes	<table><tr><td>Air</td><td>Si</td><td>Array</td><td>...</td></tr><tr><td>16</td><td>0</td><td>0</td><td>Air</td></tr><tr><td>0</td><td>16</td><td>0</td><td>Si</td></tr><tr><td>0</td><td>0</td><td>16</td><td>Array</td></tr></table>				Air	Si	Array	...	16	0	0	Air	0	16	0	Si	0	0	16	Array	1.20	5
Air	Si					Array	...																
16	0					0	Air																
0	16					0	Si																
0	0					16	Array																
Logistic (Log)	10 ⁻⁸ ridge value, full feature-class weights optimization	627	6																				
Support vector machine (SVM)	Log calibrator (see above), the linear polynomial kernel	0.45	7																				
Naive Bayes (NB)	None	0.27	36																				
k-Nearest neighbors (KNN)	1 neighbor, linear NN Euclidean distance search	0.02	22																				
Single-layer perceptron (SP)	1 hidden layer, 6 nodes, 0.3 learning rate, 0.2 momentum	100%	289	11																			
k-Means clustering (k-M)	5 clusters, Euclidean distance function	<table><tr><td>Air</td><td>Si</td><td>Array</td><td>...</td></tr><tr><td>14</td><td>0</td><td>0</td><td>Air</td></tr><tr><td>0</td><td>15</td><td>0</td><td>Si</td></tr><tr><td>0</td><td>0</td><td>16</td><td>Array</td></tr></table>				Air	Si	Array	...	14	0	0	Air	0	15	0	Si	0	0	16	Array	0.6	7
		Air	Si	Array	...																		
14	0	0	Air																				
0	15	0	Si																				
0	0	16	Array																				
~ 94% (with k= 5)																							

Table 1. List of tested ML classifiers for THz waveform recognition with provided training/testing performances.

Figure 4. Performance comparison of the different ML classifiers on the test dataset (see also Table 1) by comparing class boundaries of these classifiers with the actual Array and Si dimensions, which are shown by dotted and dashed black lines, respectively. The question marks indicate waveforms which cannot be assigned to any of three classes (Si, Array, or Air) with ML classifier.

results. This is actually expected due to the nature of NB classifier, that is, the assumption of independent and equal contribution to the outcome of each feature/time.

In principle, multi-parameter optimizations could still be performed for k -M classifier using MultiSearch meta-method in WEKA by optimizing the arbitrary number of user defined parameters and their ranges. However, we still need to use attribute selection/filtering with other tools. We did not proceed rigorously with k -M classifier tuning due to cumbersome efforts compared to already achievable 100% training accuracy with only simple settings established for RF and other well performing algorithms (see Table 1 and Fig. 4). Moreover, the state-of-the-art Auto-WEKA can be used for tuning of 789 hyperparameters from classification algorithms built into the WEKA software package^{44,45}. Depending on the requested accuracy, given time, and provided PC resources, the output is the best classifier or classifier list with information on found hyperparameters. These tools could be useful and they are available for other more challenging training datasets.

Conclusions

In principle, the results of this study could be used as a basis for more complex classification/analysis. It can be used for the automatic classification and analysis of painted/covered patterns and partial waveforms (wavelets). This may find its applications in industrial/security on-the-fly NDT together with THz/GHz and THz/GHz-TDS imaging/analysis. RF outperformed k -N and NB in overall accuracy on waveform training/testing datasets. It also outperformed all classifiers in overall training/testing time. Above all, RF typically does not require any parameter tuning to get good results if to use with 100–200 or more trees in a forest, making it as a user-friendly and a robust method while mostly outperforming other ML techniques on same training/testing datasets^{46,47}. Apart from that, Log, SVM, KNN, and SP (MP) are also very capable classifiers for waveform recognition task. Moreover, RF also produced very good results for image segmentation. However, detailed comparison with other ML methods, especially with convolutional neural networks (CNN), is still needed in the future.

Received: 31 July 2020; Accepted: 23 December 2020

Published online: 13 January 2021

References

1. Grischkowsky, D., Keiding, S., van Exter, M. & Fattinger, Ch. Far-infrared time-domain spectroscopy with terahertz beams of dielectrics and semiconductors. *J. Opt. Soc. Am. B* **7**, 2006–2015 (1990).
2. Hangyo, M., Tani, M. & Nagashima, T. Terahertz time-domain spectroscopy of solids: A review. *Infrared Millim. Terahertz Waves* **26**, 1661–1690 (2005).
3. Hangyo, M., Tani, M., Nagashima, T., Kitahara, H. & Sumikura, H. Spectroscopy and imaging by laser excited terahertz waves. *Plasma Fusion Res.* <https://doi.org/10.1585/pfr.2.S1020> (2007).

4. Bulgarevich, D. S., Shiwa, M., Furuya, T. & Tani, M. Gigahertz time-domain spectroscopy and imaging for non-destructive materials research and evaluation. *Sci. Rep.* <https://doi.org/10.1038/srep27980> (2016).
5. Mickan, S. P., Xu, J., Munch, J., Zhang, X.-C. & Abbott, D. The limit of spectral resolution in THz time-domain spectroscopy. *Proc. SPIE 5277, Photonics: Design, Technology, and Packaging* (2004).
6. Falcini, F., Lami, G. & Costanza, A. M. Deep learning in automotive software. *IEEE Softw.* **34**, 56–63 (2017).
7. Oliveira, R. B. *et al.* Computational methods for the image segmentation of pigmented skin lesions: A review. *Comput. Methods Programs Biomed.* **131**, 127–141 (2016).
8. Li, Y., Wu, F.-X. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* **19**, 325–340 (2018).
9. Maxwell, A. E., Warner, T. A. & Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **39**, 2784–2817 (2018).
10. Bhattacharjee, B. *et al.* IBM deep learning service. *IBM J. Res. Dev.* **61**, 1–10 (2017).
11. Arya, S., Pratap, N. & Bhatia, K. Future of face recognition: A review. *Procedia Comput. Sci.* **58**, 578–585 (2015).
12. Bulgarevich, D. S., Tsukamoto, S., Kasuya, T., Demura, M. & Watanabe, M. Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures. *Sci. Rep.* **8**, 2078 (2018).
13. Bulgarevich, D. S., Tsukamoto, S., Kasuya, T., Demura, M. & Watanabe, M. Automatic steel labelling on certain microstructural constituents with image processing and machine learning tools. *Sci. Technol. Adv. Mater.* **20**, 532–542 (2019).
14. Ryniec, R., Zagrajek, P. & Pałka, N. Terahertz frequency domain spectroscopy identification system based on decision trees. *Acta. Phys. Polon. A* **122**, 891–895 (2012).
15. Li, M. *et al.* Accurate determination of geographical origin of tea based on terahertz spectroscopy. *Appl. Sci.* **7**, 172–183 (2017).
16. Liu, J. *et al.* Identification of transgenic organisms based on terahertz spectroscopy and hyper sausage neuron. *J. Appl. Spectrosc.* **82**, 104–110 (2014).
17. Chen, T., Li, Z. & Moa, W. Identification of biomolecules by terahertz spectroscopy and fuzzy pattern recognition. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **106**, 48–53 (2013).
18. Hu, X. *et al.* A non-destructive terahertz spectroscopy-based method for transgenic rice seed discrimination via sparse representation. *J. Infrared Milli. Terahz. Waves* **38**, 980–991 (2017).
19. Brandt, C., *et al.* Sorting of black plastics using statistical pattern recognition on terahertz frequency domain data. *7th Sensor-Based Sorting and Control 2016*, Germany, February 23–24 (2016).
20. Motlak, H. J. & Hakeem, S. I. Detection and classification of breast cancer based-on terahertz imaging technique using artificial neural network and K-nearest neighbor algorithm. *Int. J. Appl. Eng. Res.* **12**, 10661–10668 (2017).
21. Shi, J. *et al.* Automatic evaluation of traumatic brain injury based on terahertz imaging with machine learning. *Opt. Express* **26**, 6371–6411 (2018).
22. Yin, X., Ng, B.W.-H., Fischer, B. M., Ferguson, B. & Abbott, D. Support vector machine applications in terahertz pulsed signals feature sets. *IEEE Sens. J.* **7**, 1597–1608 (2007).
23. Zhong, H., Redo-Sanchez, A. & Zhang, X.-C. Identification and classification of chemicals using terahertz reflective spectroscopic focal plane imaging system. *Opt. Express* **14**, 9130–9141 (2006).
24. Lin, X. *et al.* All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
25. Mitsuhashi, R., Murate, K., Nijima, S., Horiuchi, T. & Kawase, K. Terahertz tag identifiable through shielding materials using machine learning. *Opt. Express* **28**, 3517–3527 (2020).
26. Zhong, S., Shen, Y., Evans, M. J., May, R. K., Zeitler, J. A. & Dey, D. Neural Network-based non-destructive quantification of thin coating by terahertz pulsed imaging in the frequency domain. *35th International Conference on Infrared, Millimeter, and Terahertz Waves*, Italy, September 5–10 (2010).
27. Ye, D. *et al.* Characterization of thermal barrier coatings microstructural features using terahertz spectroscopy. *Surf. Coat. Technol.* **394**, 125836 (2020).
28. Antsiperov, V. E. Automatic target recognition algorithm for low-count terahertz images. *Comput. Opt.* **40**, 746–751 (2016).
29. Shen, X., Dietlein, C. R., Grossman, E., Popovic, Z. & Meyer, F. G. Detection and segmentation of concealed objects in terahertz images. *IEEE Trans. Image Process.* **17**, 2465–2475 (2008).
30. Agustin, A. S., Vinsley, S. S. & Krishnan, N. Image segmentation of concealed objects detected by terahertz imaging. *IEEE International Conference on Computational Intelligence and Computing Research*, India, December 28–29 (2010).
31. Xu, Z., Ye, D., Chen, J. & Zhou, H. Novel terahertz nondestructive method for measuring the thickness of thin oxide scale using different hybrid machine learning models. *Coatings* **10**, 805–819 (2020).
32. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
33. Breiman, L. Technical note: Some properties of splitting criteria. *Mach. Learn.* **24**, 41–47 (1996).
34. Bulgarevich, D. S., Watanabe, M. & Shiwa, M. Single sub-wavelength aperture with greatly enhanced transmission. *New J. Phys.* **14**(053001), 1–13 (2012).
35. Bulgarevich, D. S., Watanabe, M. & Shiwa, M. Highly-efficient aperture array terahertz band-pass filtering. *Opt. Express* **18**, 7369–7375 (2010).
36. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
37. Collins, T. J. ImageJ for microscopy. *Biotechniques* **43**, S25–S30 (2007).
38. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
39. Arganda-Carreras, I. *et al.* Trainable Weka Segmentation: A machine learning tool for microscopy pixel classification. *Bioinformatics* **33**, 2424–2426 (2017).
40. Ferreira, T. & Rasband, W. ImageJ user guide IJ 1.46r. <http://imagej.nih.gov/ij/docs/guide> (2012).
41. Frank, E., Hall, M. A. & Witten, I. H. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques* 4th edn. (Morgan Kaufmann, Burlington, 2016).
42. Ko, B. C., Kim, S. H. & Nam, J.-Y. X-ray image classification using Random Forests with local wavelet-based CS-local binary patterns. *J. Digit. Imaging* **24**, 1141–1151 (2011).
43. Wright, M. N. & Ziegler, A. A fast implementation of Random Forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
44. Thornton, C., Hutter, F., Hoos, H. H., *et al.* Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. in: *KDD '13 Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2013 August 11–14; Chicago, Illinois: ACM New York, NY, 847–855.
45. Kotthoff, L. *et al.* Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
46. Fernández-Delgado, M. *et al.* Do we need hundreds of classifiers to solve real world classification problems?. *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).
47. Tatsis, V. A., Tjortjis, C. & Tzirakis, P. Evaluating data mining algorithms using molecular dynamics trajectories. *Int. J. Data Min. Bioinform.* **8**, 169–187 (2013).

Acknowledgements

This work is partially supported by the Cooperative Research Program of Research Center for Development of Far-Infrared Region, University of Fukui (H30FIRDM022B) and by Council for Science, Technology and Innovation, Cross-ministerial Strategic Innovation Promotion Program(SIP), “Materials Integration for Revolutionary Design System of Structural Materials”(Funding agency: JST).

Author contributions

D.S.B. wrote the manuscript and developed with M.T. and M.W. the basic concept of experiments and analysis. Mi.Ta. participated in drafting the manuscript. M.W. supervised this research and edited the manuscript. The manuscript was written up through contributions of all authors. All authors have given approval to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.S.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021