S & M 3382

Machine-learning-assisted Bacteria Identification in AC Nanopore Measurement

Maami Sakamoto, Kosuke Hori, and Takatoki Yamamoto*

Mechanical Engineering, Tokyo Institute of Technology,
Ishikawadai 1-314, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan

(Received March 21, 2023; accepted June 12, 2023)

Keywords: nanopore, AC, bacteria, machine learning, lock-in detection

The AC nanopore method can measure the impedance of single nanoparticles to obtain information on their material properties as well as their size. One of the technical challenges in applying this capability to bacterial sensing lies in the realization of an analytical method to identify bacteria from measured values. In this study, we improved the bacteria identification performance of the AC nanopore method by using machine learning. Comparing four representative machine learning methods for the classification of bacterial groups that are nearly identical in size and difficult to classify based on size, we found that the random forest method has the best classification performance, achieving a classification accuracy of 78.6% for six different particles containing five bacterial species. The use of machine learning was demonstrated to be effective in improving the performance of the bacterial classification capability in the AC nanopore method.

1. Introduction

Nanopore sensing is an electrical measurement technique that utilizes the change in electrical current as a target passes through a nanometer-sized pore (nanopore) and is characterized by the ability to study the behavior and interaction of individual particles at the single-particle level.(1–3) Nanopore sensing has become an important tool in life sciences, medicine, and other fields, providing both spatial resolution at the single-particle, single-molecule level and temporal resolution in milliseconds or less, and is in some respects superior to optical single-particle observation using fluorescent molecules.(4,5) Although it has primarily been studied for applications in DNA sequencing, its use is expanding to the detection of bio-nanoparticles such as viruses and extracellular vesicles.(6–11)

The conventional nanopore method is a DC measurement method (DC nanopore method), and the measurement quantity obtained is limited to changes in the magnitude (pulse height) and time (pulse width) of the current.(1) The magnitude of the current pulse corresponds to the particle size and the current pulse width corresponds to the zeta potential of the particles. However, zeta potential has low measurement accuracy owing to various effects such as electro-

*Corresponding author: e-mail: yamamoto.t.ba@m.titech.ac.jp

https://doi.org/10.18494/SAM4398

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

osmosis and off-axis motion of particles, and practical measurements are often limited to particle size measurement.(12–14) Therefore, it was difficult to distinguish between particles of the same size. One solution to this problem is to improve the particle identification capability using machine learning, which learns and classifies the characteristics of the waveform itself without converting the measured waveform into a physical quantity.(15–17) In addition, machine learning algorithms that remove noise from noisy waveforms are beginning to be used, and this integration with information science is improving the sensing capabilities of nanopores.(18)

On the other hand, we are developing an AC-measured nanopore method (AC nanopore method), which is characterized by its ability to measure the impedance of particles because it can acquire two types of information, namely, magnitude and phase of the current, thus providing more information than DC. In this study, we aim to improve the particle identification performance over the conventional DC nanopore method by combining our AC nanopore method with machine learning.

First, single-particle detection was evaluated by measuring the size distributions of various bacteria. We then demonstrated that high classification performance can be achieved by using machine learning, even for bacteria of similar size.

2. Materials and Methods

2.1 Particles

Standard carboxylated polystyrene particles of 100 nm (CPC100) and 930 nm (CPC1000) diameter were purchased from IZON Science. These particles were dispersed in a measurement solution consisting of phosphate buffered saline (PBS) (17-516Q, Lonza) containing 0.01% final concentration of an artificial phospholipid-type surfactant (Lipidure BL206, Nichiyu Corporation) as an anti-adsorption agent. In all experiments, measurements were performed at pH 7.4 and conductivity of 17 mS/cm. To prevent aggregation, particles were diluted in PBS solution to a concentration of 10^6–10^8 particles/mL and agitated with a vortex mixer immediately before use.

2.2 Bacteria

Escherichia coli (EC) cells with an optical density of 0.5 at 600 nm wavelength were suspended in LB medium, incubated at 37 °C for 3 h, centrifuged at 5000 × *g* for 2 min, and the supernatant was replaced with the measurement solution to prepare an EC sample. *Lacrobacillus plantarum* (LP, AOK-L1315), *Enterococcus faecalis* (EF, AOK-L1390), *Leuconostoc mesenteroides* (LM, AOK-L1789), and *Pediococcus pentosaceus* (PP, AOK-L4140) were purchased from Akita Konno Shoten. For these bacteria as well, samples were prepared by suspending a bacterial solution with an optical density of 0.5 at 600 nm in MRS medium [BD Difco (TM) Lactobacillus MRS Broth, 63-6530-37, Becton Dickinson], incubating at 37 °C for 3 h, centrifuging at 5000 × *g* for 2 min, and replacing the supernatant with the measurement solution.

2.3 Nanopores

We used the commercially available nanopores NP100 and NP2000 (Izon Science, Inc.) with tunable pore sizes. The nanopores were attached to the stretching jaws of a commercially available QNano system (Izon Science Ltd.), and measurements were performed under 42 mm stretching conditions. All measurements were performed under the same nanopore and measurement conditions, except for the sample.

2.4 AC nanopore measurement

In this study, a lock-in amplifier was used to realize AC measurements for various single bacteria. Lock-in detection is a technique that applies a specific frequency modulation to a sample and extracts only signals of the same frequency that pass through the sample by homodyne detection, as shown in Fig. 1. A commercial lock-in amplifier (MFIA 5M, Zurich Instruments AG) connected with a current amplifier (SA-604F2, NF Corporation) with a gain of 10^7 V/A was used to measure the ultralow AC current required by the AC nanopore method. To reduce noise from the power supply, batteries (Eneloop, Panasonic) were used to power the preamplifier. Because the shielded cage provided with QNano reduced the noise to a level that was acceptable for measurement, no shield box or similar device was used. With these various noise reduction measures, the baseline current noise was reduced to 10 pA_{rms} under the present measurement conditions, and the measurement was performed at a signal-to-noise ratio (SNR) of 5 or better throughout. In all experiments, the particle suspension was added to the upper cell with a 3 mm (30 Pa) pressure head, and the particle suspension was hydrostatically driven.

A minimum of 500 particles were measured for machine learning. The same nanopore chip and stretching conditions were used as for the measurements, and only the sample solution was changed to perform a series of measurements under identical conditions across samples. Each measurement was performed within 10 min to avoid the effect of pressure fluctuations caused by decreasing solution volume.

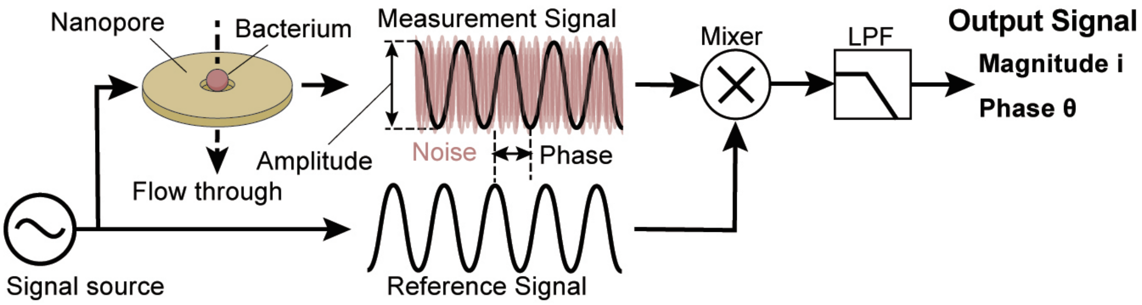

Fig. 1. (Color online) AC nanopore measurement system with lock-in amplifier.

2.5 Particle sizing

The magnitude of the current pulse is proportional to the size of the particles.(19) Therefore, the size of the bacteria was determined by calibrating standard particles of known size (CPC1000) with excellent monodispersibility against the size of the current. As for the measurement conditions, we used the following conditions, which gave the best SNR in our previous study using a similar measurement system. The measurement conditions were applied voltage of 200 mVpp at measurement frequency of 10 kHz, 3rd-order low-pass filter, bandwidth of 2 kHz, and particle driving pressure head of about 30 Pa, which gave the best SNR results in our previous studies.(20) For nanopores, NP1000 was used to cover a bacterial diameter range of approximately 1 to 2 μm. These measurement conditions were optimized in advance on the basis of the particle velocity generated by the applied pressure, and the measurement frequency and low-pass filter settings, which affect quantitation, were closely adjusted and checked. Since the measurement frequency in this measurement is equivalent to the time resolution, the measurement frequency must be sufficiently high relative to the particle velocity. In addition, the low-pass filter setting must be set to a time constant (bandwidth) that is sufficiently fast relative to the particle velocity. However, if the measurement frequency or LPF setting is set higher than necessary, noise will increase and SNR will decrease, so the balance between measurement speed and SNR is important. For example, in the above measurement, the pulse widths of the current and phase were several milliseconds, and the speed was their reciprocal. The measurement frequency of 10 kHz corresponds to 0.1 ms in the time domain, which is a profiling condition of several tens of points for a pulse width of several milliseconds.

2.6 Initial data processing

To detect the small signal pulses obtained by the AC nanopore method, it is necessary to separate the signal pulses from the background noise. A threshold is defined to distinguish between spikes and noise, so that spikes are detected by considering them as spikes if they are higher than the threshold. The threshold was defined as five times the standard deviation of the signal intensity of the noise. However, the baseline current fluctuates over time due to increases in electrolyte concentration caused by solvent evaporation, electrode reactions, and so forth. Therefore, using data processing software (DIAdem, National Instruments), we flattened the data by finding the baseline trend at the moving median for every 3000 points and determined the difference from the original baseline. Since the baseline of the measurement data was horizontal after flattening, which allowed the threshold to be set, spikes were detected using the procedure described above. The spike detection algorithm provided by SciPy, a Python library, was used to detect spikes.

2.7 Machine learning

Supervised learning was used to classify bacteria from the pulse waveforms obtained in the experiment. In this learning method, the features extracted from each pulse wave are pretrained in the learning model as to which bacteria they are derived from. Six types of feature were used for learning: the height of the pulse waveform shown in Fig. 2(c) and the width of each height divided into five equal parts on the basis of the peak of the pulse waveform height. When two pulse waveforms of current magnitude and phase were used in pairs, there were 12 types of feature.

Machine learning was performed using SciPy, a Python library, to calculate the pulse height and width from the measured current and phase difference data, respectively. Four classification models were employed from the generic model: random forest, K-nearest neighbor method, support vector machine, and logistic regression.

As a preprocessing step for learning, the number of training data was first made the same for each classification method by undersampling, which randomly removes data from the majority group and aligns them with the number of data from the minority group. In addition, because the scale of the input data could affect the learning results, the data were normalized and the scales were adjusted. Furthermore, because some features may be unnecessary, all combinations were trained and the combination showing the highest accuracy was adopted. The accuracy of the training model was determined by k-partition cross-validation in order to prevent overtraining.

3. Results and Discussion

The lock-in amplifier converts the measured AC signal to a DC root-mean-square signal. This provides a measurement waveform similar to a DC measurement. Figure 2 shows a typical measurement waveform obtained by AC nanopore measurement, showing the magnitude change of the current and the phase change of the current with respect to the applied voltage. This example used CPC100 with a particle diameter of about 100 nm, a nanopore chip of NP100, and

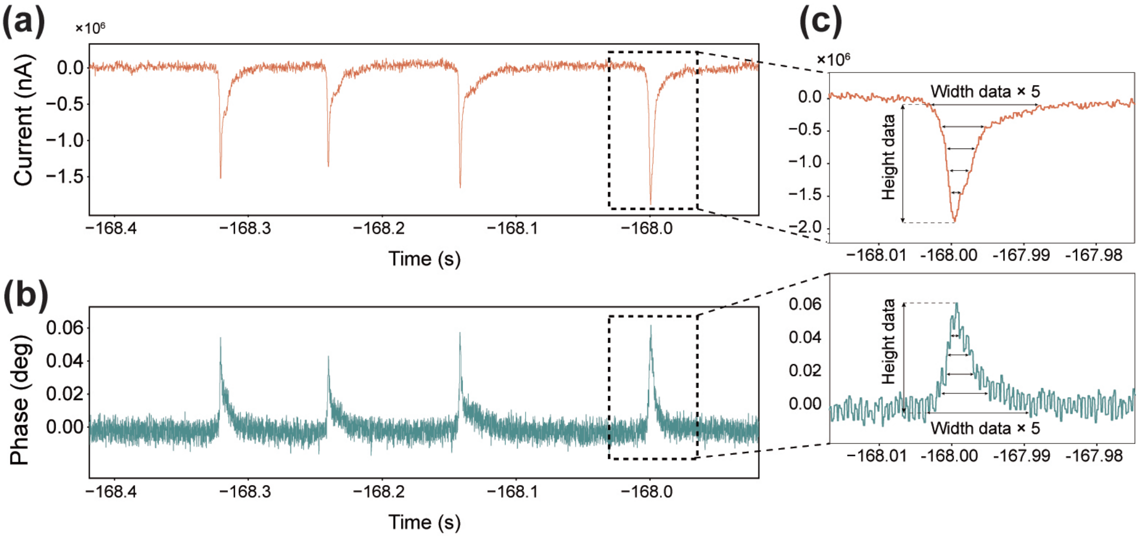

Fig. 2. (Color online) Representative measurement waveforms. (a) Current changes, (b) phase changes, and (c) magnified view of the current and phase waveform pairs.

a measurement frequency of 10 kHz. As the particles pass through the nanopores, downward pulsed current changes occur, as shown in Fig. 2(a). On the other hand, the phase changes upward (phase advances), as shown in Fig. 2(b). Since the phase of the current is more advanced than the applied voltage, it is clear that the response is capacitive. It is also clear that the current and phase changes originate from the same particle, since they occur synchronously. Figure 2(c) shows a magnified view of each waveform. The figure shows the pulse height and each width of the height divided into five equal parts on the basis of the peak of the pulse height; these were used as the parameters for machine learning.

Conventional studies revealed that the magnitude of the current pulse is proportional to the particle volume.(19) On the other hand, it has been shown that for nonspherical particles such as bacteria, i.e., ellipsoids, the longitudinal direction has a lower Brocade rate of current through the nanopore than the transverse direction.(21) Therefore, if nonspherical particles are assumed to be spherical, they must be freely rotating. However, in this study, we did not evaluate microscopic motion; thus, the analysis results include the possibility that the measured signal may vary depending on the direction in which the nonspherical particles pass through the nanopores. Assuming that the bacteria are spherical, although asymmetry effects may be included, the particle size distribution of each bacterium calibrated by the magnitude of the current pulse for a standard particle of known size is as shown in Fig. 3. The distribution results were in good agreement with the Gaussian distribution, confirming that the mean and mode diameters were nearly equal (Table 1). The coefficient of variation (CV) was found to be 7.1% for CPC1000, the calibration particle, whereas the bacteria were dispersed at a size of 7–10%.

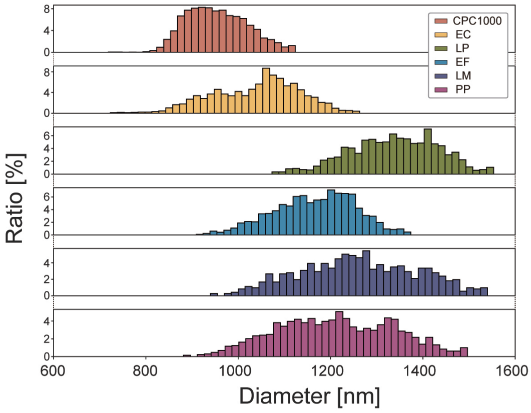

Fig. 3. (Color online) Size distribution of bacterial particle.

Table 1
Statistics on measured bacterial particles, taken from Fig. 3.

	Mean dia. (nm)	Mode dia. (nm)	CV (%)
CPC1000	937.7	928.7	7.1
EC	1008.3	1033.0	9.8
LP	1314.6	1333.1	7.2
EF	1141.3	1119.0	7.7
LM	1236.0	1249.9	9.9
PP	1179.7	1191.5	10.4

Because of the high dispersion and overlapping particle size distributions, it was difficult to identify specific bacteria when evaluated on the basis of particle size alone, as in the DC nanopore method. Therefore, we attempted to classify bacteria by machine learning. Pulse waveform pairs of current magnitude and phase obtained by the AC nanopore method were used for training. These waveforms were classified using four different classification methods. Random forest, logistic regression, *k*-nearest neighbor, and SVM are all widely used machine learning algorithms. In this study, we evaluated these algorithms for use in edge AI, where the actual classification is performed by sensors in the field, using a pretrained classification model. Therefore, the time and computational cost required for training were excluded from the evaluation, and only the classification performance was evaluated. Figure 4 shows the sample size dependence of accuracy for each method. From this figure, it is clear that the random forest method has the highest accuracy among the four methods. However, despite parameter tuning, convergence was not reached with 720 samples, the maximum number of samples in this study. Therefore, there remains a possibility that the accuracy can be further improved by increasing the number of samples in the future. Figure 5(a) shows representative measured waveforms of each bacterium used as training data for machine learning. Figures 5(b) to 5(e) are confusion matrices showing the classification performances of random forest [Fig. 5(b)], k-nearest neighbor algorithm [Fig. 5(c)], support vector machine [Fig. 5(d)], and logistic regression [Fig. 5(e)]. The darker the blue color of the right-down diagonal, the higher the classification performance. The number in the matrix is the number of particles classified. Comparison of the four methods revealed that the random forest method has the highest accuracy of 78.6%, with an F1 score (harmonic mean of accuracy and reproducibility) of 78.4%, as shown in Fig. 4.

It is assumed that this accuracy ranking reflects the following characteristics of each method. Logistic regression has difficulty with nonlinear classification and is not good at multiclass classification along with k-nearest neighbor and SVM. SVM and k-nearest neighbor are vulnerable to noise and outliers; moreover, hyperparameter tuning is a difficult feature, leaving

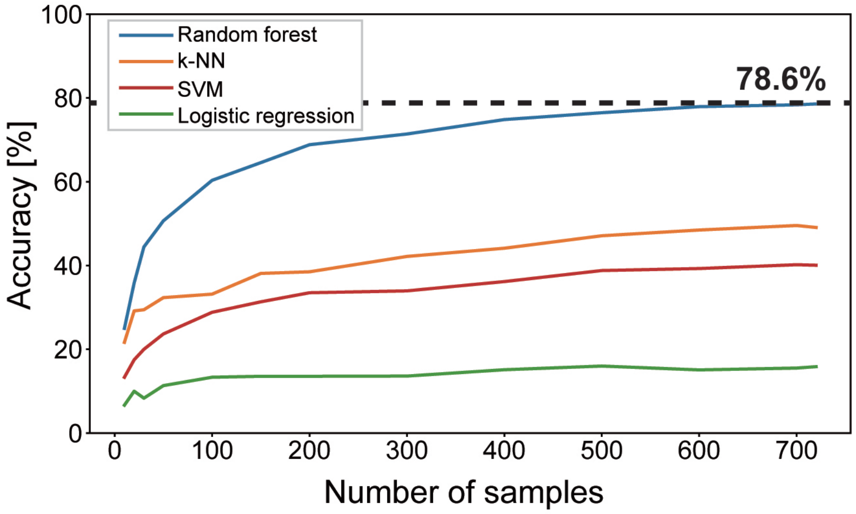

Fig. 4. (Color online) Dependence of accuracy on the number of samples for four methods.

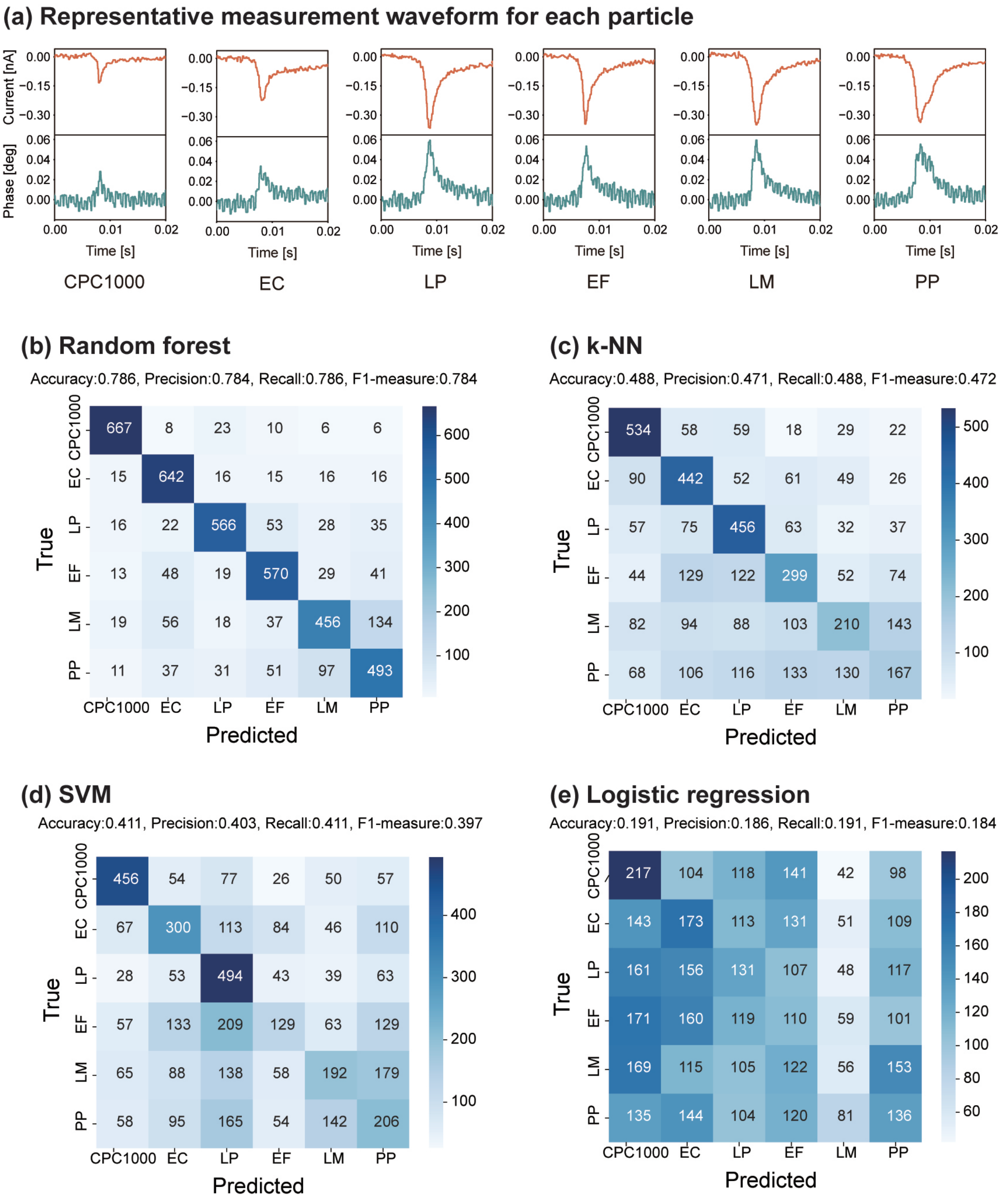

Fig. 5. (Color online) Identification results of bacteria using four classification methods. (a) Representative waveform for each particle. Classification of five bacterial species by (b) random forest, (c) k-nearest neighbor algorithm, (d) support vector machine, and (e) logistic regression. CPC1000 is the calibration particle for each particle size. Accuracy is the percentage of correct answers relative to the total data. Precision is the percentage of predicted positives that are actually positive, indicating fewer false positives. Recall is the percentage of predicted positives that are actually positive, indicating fewer false negatives. F1-measure is the harmonic mean of precision and recall, which indicates the balance between the two. The maximum value for all is 100, with closer to 100 indicating better performance.

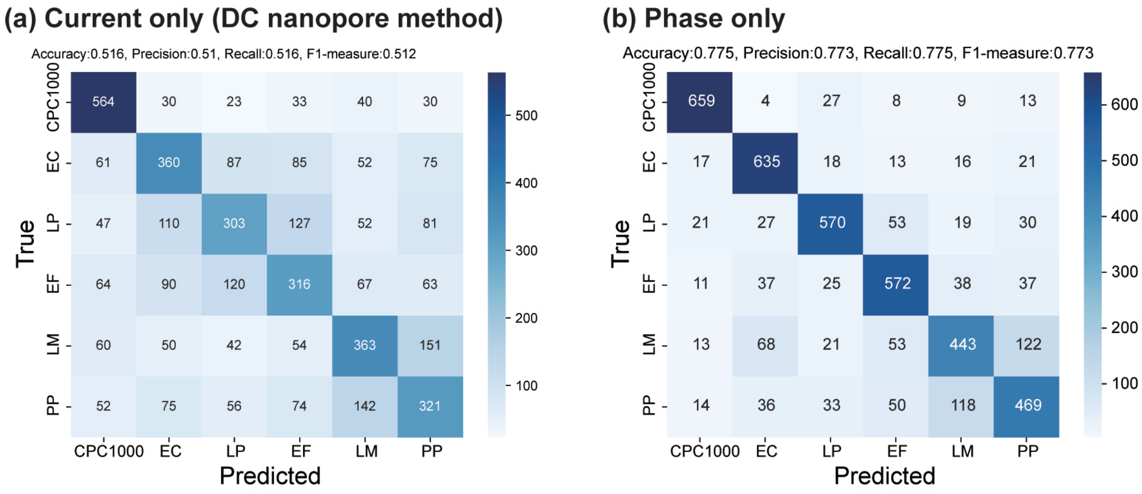

Fig. 6. (Color online) Classification results when only current magnitude and phase are used in the random forest method. (a) Result of current-only classification; (b) result of phase-only classification.

the possibility that they cannot be tuned to optimal values. On the other hand, random forests are suitable for data sets that are nonlinear and have a large number of features, and they are highly resistant to noise and outliers. It is assumed that the advantages of random forests were utilized in classification based on a large number of nonlinear features with a lot of noise and outliers, as in the present case.

In contrast to the results of classification based on both current magnitude and phase waveforms in Fig. 5, Fig. 6 shows the classification results when only the current magnitude waveform (equivalent to the conventional DC nanopore method) or only the phase waveform is used, taking the random forest method as an example. The accuracy was 51.6% for current magnitude alone and 77.5% for phase alone. In this example, the classification accuracy was higher for phase than for current, and it was clear that the highest accuracy was achieved when both were used. In other words, under the conditions of this study, the AC method was 27.0% more accurate than the DC method.

The phase includes not only particle size but also information related to material properties such as dielectric constant, which may be one of the reasons for the higher accuracy, although clarifying the details of this is a future issue.

4. Conclusions

In this study, the bacterial classification ability of the AC nanopore method was improved by implementing machine learning data analysis. First, we evaluated the performance among four major machine-learning-based classification methods, and found that the random forest method gave the best accuracy. Classification by the random forest method provided a 78.6% accuracy and 78.4% F1-measure for bacterial samples that were difficult to classify on the basis of particle size alone, such as those with overlapping particle size distributions.

The AC nanopore method also improved classification accuracy by 27.0% compared with the

DC nanopore method using only current magnitude alone. This indicates that the AC nanopore method, with its multimodal measurement capabilities, such as current magnitude and phase, is superior to the DC nanopore method in classification using machine learning.

In the future, not only the measurement accuracy of the AC nanopore method, but also further improvements in the machine learning algorithm are expected to improve the identification performance. With these improvements, we expect this technology to have a wide range of applications, including the realization of identification at the virus, extracellular vesicle, and even molecular scales.

Acknowledgments

This work was partly supported by a grant-in-aid for scientific research (KAKENHI) (No. 20H02160) from the Japan Society for the Promotion of Science (JSPS) and Adaptable and Seamless Technology Transfer Program through Target-driven R&D (A-STEP) from the Japan Science and Technology Agency (JST) (Grant Number JPMJTR22RE).

References

1 L. Xue, H. Yamazaki, R. Ren, M. Wanunu, A. P. Ivanov, and J. B. Edel: Nat. Rev. Mater. **5** (2020) 931. https://doi.org/10.1038/s41578-02

2 H. Miller, Z. Zhou, J. Shepherd, A. J. M. Wollman, and M. C. Leake: Rep. Prog. Phys. **81** (2018) 024601. https://doi.org/10.1088/1

3 D. Kozak, W. Anderson, R. Vogel, and M. Trau: Nano Today **6** (2011) 531. https://doi.org/10.1016/j.nantod16/j

4 J. K. Rosenstein, M. Wanunu, C. A. Merchant, M. Drndic, and K. L. Shepard: Nat. Methods **9** (2012) 487. https://doi.org/10.1038/Nmeth.1932

5 A. J. W. Hartel, S. Shekar, P. Ong, I. Schroeder, G. Thiel, and K. L. Shepard: Anal. Chim. Acta **1061** (2019) 13. https://doi.org/10.1016/j.aca.2019.01.034

6 T. Albrecht: Annu. Rev. Anal. Chem. **12** (2019) 371. https://doi.org/10.1146/annurev-anchem-061417-125903

7 D. Deamer, M. Akeson, and D. Branton: Nat. Biotechnol. **34** (2016) 518. https://dx.doi.org/10.1038/nbt.3423

8 J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley: Nat. Nanotechnol. **4** (2009) 265. http://www.nature.com/doifinder/10.1038/nnano.2009.12

9 S. Akhtarian, S. Miri, A. Doostmohammadi, S. K. Brar, and P. Rezai: Bioengineered **12** (2021) 9189. https://doi.org/10.1080/21655979.2021.1995991

10 A. Darvish, J. S. Lee, B. Peng, J. Saharia, R. VenkatKalyana Sundaram, G. Goyal, N. Bandara, C. W. Ahn, J. Kim, P. Dutta, I. Chaiken, and M. J. Kim: Electrophoresis **40** (2019) 776. https://doi.org/10.1002/elps.201

11 A. Arima, M. Tsutsui, I. H. Harlisa, T. Yoshida, M. Tanaka, K. Yokota, W. Tonomura, M. Taniguchi, M. Okochi, T. Washio, and T. Kawai: Sci. Rep. **8** (2018) 16305. https://doi.org/10.1038/s41598-0

12 R. Vogel, A. K. Pal, S. Jambhrunkar, P. Patel, S. S. Thakur, E. Reátegui, H. S. Parekh, P. Saa, A. Stassinopoullos, and M. F. Broome: Sci. Rep. **7** (2017) 17479. https://doi.org/10.1038/s41598-017-14981-x

13 N. Arjmandi, W. Van Roy, L. Lagae, and G. Borghs: Anal. Chem. **84** (2012) 8490. https://dx.doi.org/10.1021/ac300705z

14 W. R. Smythe: Rev. Sci. Instrum. **43** (1972) 817. https://doi.org/10.1063/1.1685770

15 J. Ko, N. Bhagwat, S. S. Yee, N. Ortiz, A. Sahmoud, T. Black, N. M. Aiello, L. McKenzie, M. O'Hara, C. Redlinger, J. Romeo, E. L. Carpenter, B. Z. Stanger, and D. Issadore: ACS Nano **11** (2017) 11182. https://doi.org/10.1021/acsnano.7b05503

16 A. Arima, M. Tsutsui, T. Washio, Y. Baba, and T. Kawai: Anal. Chem. **93** (2021) 215. https://dx.doi.org/10.1021/acs.analchem.0c04353

17 M. Taniguchi, S. Minami, C. Ono, R. Hamajima, A. Morimura, S. Hamaguchi, Y. Akeda, Y. Kanai, T. Kobayashi, W. Kamitani, Y. Terada, K. Suzuki, N. Hatori, Y. Yamagishi, N. Washizu, H. Takei, O. Sakamoto, N. Naono, K. Tatematsu, T. Washio, Y. Matsuura, and K. Tomono: Nat. Commun. **12** (2021) 1. https://doi.org/10.1038/s41467-021-24001-2

18 M. Tsutsui, T. Takaai, K. Yokota, T. Kawai, and T. Washio: Small Methods **5** (2021) e2100191. https://doi.org/10.1002/smtd.202100191

19 R. W. DeBlois and C. P. Bean: Rev. Sci. Instrum. **41** (1970) 909. https://doi.org/10.1063/1.1684724

20 K. Kitta, M. Sakamoto, K. Hayakawa, Akira Nukazuka, A. Nukazuka, K. Kano and T. Yamamoto: ACS OMEGA **8** (2023) 14684. https://doi.org/10.1021/acsomega.3c00628

21 C. Ying, J. Houghtaling and M. Mayer: Nanotechnology **33** (2022) 275501. https://doi.org/10.1088/1361-6528/ac6087

About the Authors

Maami Sakamoto received her B.E. degree from Tokyo Institute of Technology, Japan, in 2022. Currently, she is enrolled in the Mechanical Engineering Course, Department of Mechanical Engineering, School of Engineering, Tokyo Institute of Technology. Her research interests are in machine learning, data science, and lab-on-a-chip technology.
(sakamoto.m.ag@m.titech.ac.jp)

Kosuke Hori received his B.E. degree from Tokyo Institute of Technology, Japan, in 2022. Currently, he is enrolled in the Mechanical Engineering Course, Department of Mechanical Engineering, School of Engineering, Tokyo Institute of Technology. His research interests are in microfabrication, biosensing, and lab-on-a-chip technology. (hori.k.ak@m.titech.ac.jp)

Takatoki Yamamoto received his D.E. degree from Kyoto University, Japan, in 1999. From 1999 to 2000, he was a special postdoctor at RIKEN, Japan. From 2000 to 2008, he was an assistant professor at Institute of Industrial Science, The University of Tokyo, Japan. Since 2008, he has been an associate professor at Tokyo Institute of Technology. He is a member of the Institute of Electrical Engineers of Japan, the Japan Society of Mechanical Engineers, the Society of Chemistry and Micro-Nano Systems, the Japanese Society for Artificial Intelligence, and the Japanese Society for Environmental Infectious Diseases. His research interests are in bionanotechnology, micro/nano systems, biosensors, machine learning, and data science.
(yamamoto.t.ba@m.titech.ac.jp)