Classification for Glucose and Lactose Terahertz spectrums based on SVM and DNN methods

Kaidi Li, Xuequan Chen, Rui Zhang and Emma Pickwell-MacPherson

***Abstract*— In recent decades, terahertz (THz) radiation has been widely applied in many chemical and biomedical areas. Due to its ability to resolve the absorption features of many compounds non-invasively, it is a promising technique for chemical recognition of substances such as drugs or explosives. A key challenge for THz technology is to be able to accurately classify spectral measurements acquired in unknown complicated environments, rather than those from ideal laboratory conditions. Support vector machine (SVM) and deep neural networks (DNN) are powerful and widely adopted approaches for complex classification with a high accuracy. In this paper, we explore and apply the SVM and DNN methods for classifying the frequency spectra of glucose and lactose. We measured 372 groups of independent signals under different conditions to provide a sufficient training set. The classification accuracies achieved were 99% for the SVM method and 89.6% for the DNN method. These high classification accuracies demonstrate great potential in chemical recognition.**

***Index Terms*— SVM and DNN methods, Terahertz spectrum**

I. Introduction

Detection using THz radiation (0.1-10 THz) is attractive as many visibly opaque materials, especially most packaging materials are transparent in the THz range[1]. Additionally, many drugs, explosives and crystalline chemicals have characteristic spectra in this region [2-4]. THz spectroscopy is thus a promising tool to be applied in material recognition and classification, especially for security applications. Despite the great potential, there is still a long way to go to realize efficient practical applications of THz spectroscopy. A robust data analysis method should also be developed to enable accurate recognition. This is because most practical detection, especially for security checks, cannot be performed in optimized laboratory conditions. The investigated samples typically contain multiple ingredients with the component of interest having a low concentration. Irregular sample shapes, high attenuation, large thickness, Fabry-Perot oscillations and absorption from the packaging materials, will all strongly disrupt the sample fingerprints [5-8]. In addition, THz spectra contain dense and strong water-vapor absorption lines, which are difficult to remove in a practical situation. These feature lines can interfere with the sample characteristic absorptions to further increase the recognition difficulty [9]. Therefore, it is very challenging to establish a simple evaluation function to quantitatively identify the material from the spectral features. Manual recognition requires professional knowledge and adequate experience, which prevents the broad application of this technique. Instead, machine learning based algorithms can be a good solution to accurately and automatically classify different substances in a complicated environment by training the data measured under various imperfect conditions approximating the real situations. In this work, we have investigated SVM and DNN methods for the classification of glucose and lactose THz spectra in various imperfect conditions. Similar methods have been proposed to realize accurate classification for other THz applications. For example, Yin X. et al proposed to use SVM to do the binary and multiple classifications [10]. They combined SVM with feature extraction methods to realize 72% accuracy for classifying two types of RNA samples. However, these two RNA samples are easily discernible in transmission THz imaging which is not a good demonstration for the excellent classification ability of SVM. The data collected from six powder substances lack independence and diversity as only densities and concentrations are taken into account. Shi J. et al introduced a machine learning method to classify the biological THz images of traumatic brain injury (TBI) into sham, mild, moderate and severe degrees [11]. The accuracy is up to 87.5%, which is a great improvement. However, the number of the collection for each type is below 20, which is deficient in terms of diversity to draw a strong conclusion. Y. Sun et al performed a machine learning method to classify different concentrations of bovine serum albumin (BSA) solutions based on THz time-domain spectroscopy. The accuracy can be up to 91% [12]. However, samples of each concentration level were measured 7 times, which reduces the independency of the data for machine learning method to learn the spectroscopic features. The major difference between the reported works and our study is that, all the previous experiments were performed in optimized conditions. Here, we utilize the great generalization ability of SVM and DNN methods for material classifications in complicated and unpredictable environments.

SVM is a powerful and widely adopted algorithm aiming to solve the binary classification problem based on a linear classifier with the largest interval defined in the feature space

This work was supported in part by the Hong Kong Research Grants Council (project numbers: 14201415 and 14205514). Corresponding author: Emma Pickwell-MacPherson.

K. Li, C. Chen are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong (e-mail:1155100870@link.cuhk.edu.hk; swench@qq.com)

R. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China (e-mail: zhangruigt@126.com).

E. Pickwell-MacPherson is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong, and with Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, U.K. (e.pickwell.97@cantab.net).

[13].Gaussian and polynomial kernels have been found to give a good performance for two-class problems, depending on the feature of the input data. Deep learning architectures like DNN have been successfully applied in many fields such as material inspection, medical image analysis and bioinformatics, where they have obtained results comparable to, and in some situations better than human experts [14-15]. The experiments in this work demonstrate the potential of SVM and DNN in THz spectra recognition. Uncorrelated glucose and lactose spectra are collected by THz-time domain spectroscopy in transmission mode under different experiment conditions to simulate the situations in practical applications. The classification results show 99% accuracy for SVM and 89.6% for DNN method. To the best of our knowledge, this is the first work investigating classification approaches under numerous imperfect experimental conditions to develop a robust algorithm towards practical applications. We show the great potential of using THz spectroscopy combined with machine learning methods to realize highly accurate automatic classification, which is not limited to the chemical compound examples in this work, but can also be further extended to various drugs and explosives for security applications.

II. EXPERIMENT SETUP

The self-built THz time-domain spectroscopy system is set up in transmission to collect the spectra from glucose and lactose samples. As shown in Fig. 1, both pump and probe beams are produced by a Ti:Sapphire femtosecond laser. A delay stage is implemented to adjust the optical path difference between the pump and probe beams. The THz beam is generated by the LT-GaAs photoconductive antenna, which is then collimated and focused onto the sample by a pair of parabolic mirrors with a focal spot of about 3 mm diameter. The THz beam passing through the sample is recollimated and focused to the detector by another pair of parabolic mirrors. The THz time-domain waveform was sampled by moving the delay stage using a step scan mode. The step scan moves and stops the delay stage at every sampling position to integrate the waveform amplitudes at each time-position. Compared to the rapid scan mode which continuously shakes the delay stage forwards and backwards, it provides a better accuracy for the sampling position. The frequency-spectrum is acquired by applying a Fast Fourier Transform (FFT) to the time-domain signals. The 3ms time-constant used for the data integration provides up to 60 dB signal-to-noise ratio (SNR) for the air reference signal at around 1 THz.

The glucose and lactose powders were purchased from XILONG SCIENTIFC Company and SIGMA life science, respectively. The powders were pressed into pellets using a pellet press. The same pressure and pressing time were applied for preparing different samples. During the measurement, the pellet was mounted on an open aperture.

To obtain sufficient uncorrelated data that simulate the possible detecting environment in practical applications, we designed several conditions as shown in Table 1. The glucose and lactose pellets have the same diameter, thus the weight is proportional to the thickness, and larger weights result in greater attenuation to the THz light, giving rise to lower SNR. The doping level of polyethylene (PE) was set to be 0% or 50% during the fabrication process to introduce different sample concentrations.

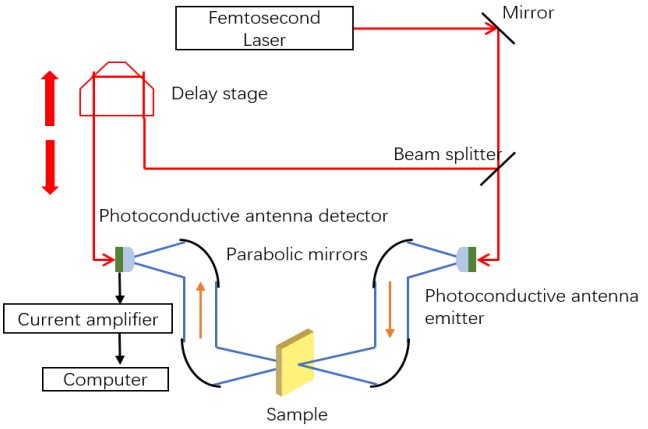


Fig.1. Diagram of the THz spectroscopy measurement in transmission geometry. The red line shows the route of the laser, the red orange arrows the route of the THz beam.

The frequency spectrum resolution is inversely proportional to the measured signal length in time-domain. In the experiment, 200ps and 100ps time length, corresponding to 5 GHz and 10 GHz frequency steps respectively, were set to verify the algorithm performance for different frequency resolutions. The sampling step size was set as 0.3ps and 0.15ps, which determines the highest frequency in the spectrum to be 1.67 THz and 3.33 THz, respectively. The pellets were also tilted with an angle of 0^0 and 30^0 respectively. The samples were placed either at the focus point or 3 cm after or before the focus point (the parabolic mirror has an f-number of 1 and a focal length of 10 cm). These settings test the algorithm performance on irregular sample shapes and poor alignments. Disturbances such as placing paper, quartz or a silicon wafer in front of the pellet were also introduced to investigate the robustness against packing blocks. All the measurements were carried out under room temperature (20^0c) and normal humidity (60%) to approximate the situation with dense water-vapor absorption lines. All the above variables simulate different complicated conditions that could occur in practical

TABLE I
SETUP FOR THE MEASUREMENT CONDITIONS

Conditions	Glucose		Lactose	
Weight(mg)	167,245, 91	90	90,150,242	113
Doping level	Pure	50%PE	Pure	50%PE
Measurement Time(ps)	100,200		100,200	
Step size(ps)	0.3, 0.15		0.3, 0.15	
Incident Angle	0^0, 30^0		0^0, 30^0	
Distance to focus point	3cm before,3cm after, at		3cm before,3cm after, at	
Disturbance	No, paper, silicon, quartz		No, paper, silicon, quartz	

applications, and they also greatly increase the diversity of the data.

III. Terahertz Data Representation

The structures of glucose and lactose are shown in Fig. 2. They have very similar chemical elements and structures. In the powder or pellet form, they cannot be distinguished by the naked eye. Although they have spectral fingerprints in the THz range, their THz responses under various complicated situations were greatly disturbed, making it very difficult to correctly classify them. Similar problems also occur for detecting many other substances using THz spectroscopy. For the acquired THz signals, spline interpolation was implemented in the frequency domain from 0.1-1.5 THz and a step of 5 GHz was set so that all the spectra have the same frequency axis. In this way, the length of the input vector was 280 for all data. Fig. 3(a) shows glucose and lactose frequency spectrum magnitudes measured in the following condition: pure pellet, at the focus point, no disturbance, 0^0 tilted angle. We refer to this as the standard condition. This condition is similar to the optimized condition set in most published works. But the data were measured in an ambient wet air environment to simulate the practical identification condition. The spectra can weakly display the absorption peaks at 1.41 THz for the glucose and more clearly resolve the peaks at 0.53 THz, 1.38 THz for the lactose. However, the water-vapor lines at 0.56 THz and 1.4 THz greatly reduce the visibility of the sample features. Fig. 3(b-d) show the same glucose and lactose pellets measured in some non-ideal conditions. In detail, Fig. 3(b) was measured with a crystal quartz placed in front of the sample, the sample being placed 3cm after the focus point with a 30^0 tilted angle. Fig. 3(c) was measured with a piece of paper in front of the sample, the sample being placed at the position 3 cm before the focus point with a 30^0 incident angle. Fig. 3(d) was measured with a high-resistivity silicon in front of the sample, the sample being placed at the position 3 cm before the focus point with a 30^0 incident angle. From the comparison of (a-d), we notice that even for the same glucose and lactose pellets, the spectra vary a lot under different situations. The glucose absorption feature at 1.41 THz became much less distinguishable in non-standard conditions. Furthermore, fluctuations from Fabry-Perot oscillations and un-optimized alignment were generated, which further disturbed the recognition. Similar influences can also be found for lactose. In practical detection applications, the measurement conditions can be even more complicated. For example, the sample could be in an absorptive container, or it could be mixed with other chemicals with absorption features. The purity could be low or the sample could be too thick to provide a good SNR. In these cases, manual identification could be very difficult, thus we expect SVM or DNN could work more efficiently to provide a reliable recognition.

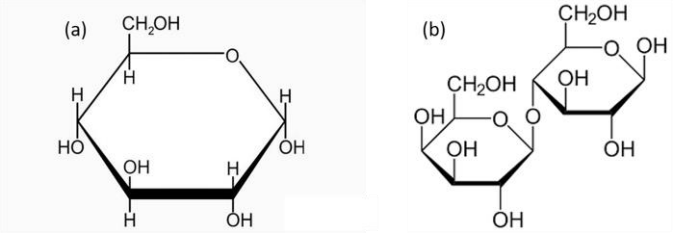


Fig.2. Structures of the glucose and lactose shown in (a) and (b), respectively.

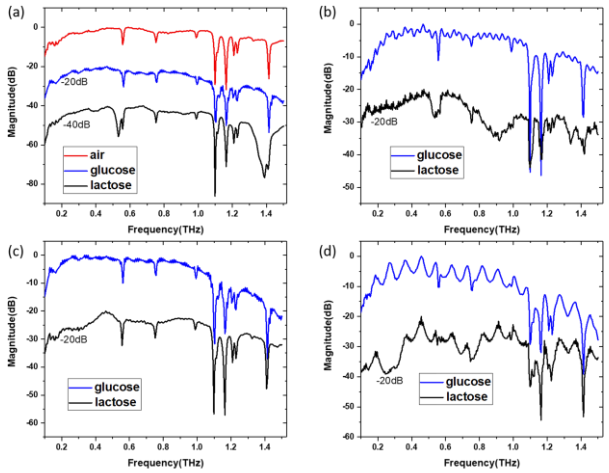


Fig.3. Magnitudes of glucose and lactose spectra measured under different conditions. (a) shows glucose and lactose frequency spectrum magnitude measured in the following condition: pure pellet, at the focus point, no disturbance, 0^0 tilted angle. (b) shows the spectrums measured with a crystal quartz placed in front of the sample, the sample being placed 3 cm after the focus point with a 30^0 tilted angle. (c) was measured with a piece of paper in front of the sample, the sample being placed at the position 3 cm before the focus point with a 30^0 incident angle. (d) was measured with a high-resistivity silicon in front of the sample, the sample being placed at the position 3 cm before the focus point with a 30^0 incident angle. The magnitude is normalized to the maximum value with an offset indicated in the figure for clarity.

IV. Methodology

A. SVM classification

SVM is a branch of machine learning theory, which maps the low dimension input data into a high dimension space where an optimal hyperplane boundary separating the two classes can be constructed [16]. Basically, SVM is a two-class classifier, while in many cases it can be designed to realize multi-classification [17]. In our cases, the output of the SVM is either glucose or lactose. The training dataset is a collection with the learning vectors and labels, denoted as (x_i, y_i) where x_i is the learning vector corresponding to the spectrum magnitudes of glucose or lactose and y_i is the corresponding label which is indicated by 0 or 1, standing for the glucose and lactose, respectively. By using the kernels, the computation of production of two kernels would accomplish in the low dimensional space, which replaces for all the occurrences of the dot product resulting from two mappings. In our study, the Gaussian kernel was implemented, because of its better classification performance compared to other kernels [10], as shown in the following equation

$$K(\overrightarrow{x_i},\overrightarrow{x_j}) = \exp(-\Upsilon\|\overrightarrow{x_i} - \overrightarrow{x_j}\|^2) \quad (1)$$

where $\overrightarrow{x_i},\overrightarrow{x_j}$ are two vectors, Υ is the parameter. Next, the boundary function built to separate two classes is defined by the support vectors in the mapped N-dimensional hyperplane in feature space, which are the subset of the training dataset, as shown in equation (2) .

$$f(\alpha) = R^N \rightarrow \pm 1 \quad (2)$$

where α corresponds to the weight and bias in the boundary function. It can be adjusted to label the output based on the input vector. Through the identification of support vectors, the parameters in the boundary function can be well adjusted for the prediction. While for other preset parameters, such as *c*, which is a punishment parameter to balance the aim of searching for the biggest margin hyperplane and guarantee minimum data point deviation, a parametric search needs to be implemented to find the most suitable parameters for the algorithm, thus the best performance can be achieved. Here we show the results of adjusting the parameter *c* as an example in Fig. 4, as the procedure is similar to adjusting the other parameters. With the increased number of the training data, the accuracies with different *c* values increase as expected. Notice that the accuracy for each training number is the averaged result from 20 runs. The best classification performance is obtained by *c* =5000 and Υ =0.008 with the accuracy higher than 99% when the training data covers larger than 20% of total data.

To evaluate the classification performance of SVM, the training data were randomly picked from a specific proportion, varying from 1/10 to 9/10 of the input population, which is 372 data in total for both samples. The SVM parameters were also adjusted based on the test performance.

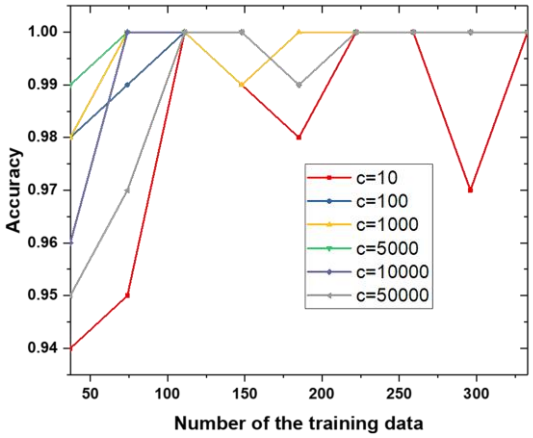


Fig.4. Classification accuracy versus the number of input training data, corresponding to the different values of the parameter c.

B. DNN classification

DNN is another method which has the potential to classify samples with a high accuracy. In our cases, as shown in Fig. 5, the connection of the full 3 layers is constructed to do the classification. The first layer has 280 neurons corresponding to the 280 spectral magnitudes as the input at the beginning. The output of the first layer contains 140 neurons, which is also the input for the second layer. The output for the second layer contains 70 neurons, being the input of the third layer. Finally, 2 neurons are outputted from the third layer as the classification result. The neuron with the higher value is recognized as the prediction result. For example, if the value of the first neuron is larger than the second one, the prediction would be glucose, otherwise, the result is lactose. The activation function, Relu, which introduces the nonlinearity into the network for the first two layers is shown in equation (3):

$$h(x) = \begin{cases} x & (x > 0) \\ 0 & (x \leq 0) \end{cases} \quad (3)$$

The criterion for loss function is Crossentropy Loss, which is used to optimize the parameters in a neural network model, defined in the following equation:

$$E = \sum -t_k log y_k \quad (4)$$

where t_k in a format of one-hot is the label of the correct output, y_k is the corresponding output from the DNN. K stands for the number of neurons in the output layer.

The training process is based on the loss function. The weights and biases were adjusted to get the minimal loss to fit the training data via the gradient descent, thus realizing a high classification accuracy. It can be divided into 3 steps. Firstly, a part of the data named as mini-batch is randomly selected from the training data. Secondly, to minimize the loss function, the gradient of all the weights and biases based on the loss function should be calculated. Thirdly, the weights and biases will be updated towards the direction of the gradient. These three steps will be repeated until the algorithm reaches the maximum iteration time or the training accuracy reaches a certain value.

The aim of the DNN is to classify data beyond the training data, that is the generalization ability. The final performance of DNN is evaluated based on the classifying accuracy on the test data.

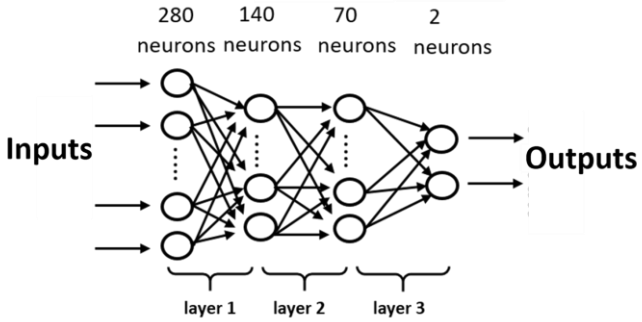


Fig.5. Structure of the deep neural network implemented in our classification task.

V. PERFORMANCE ASSESSMENT OF THE CLASSIFICATION

All the classification programs were performed using a computer with a 3.6 GHz Inter(R) Core(TM) i7-4790 CPU. The average testing time for classifying the two-class samples was shorter than 0.01s for the SVM and 0.5s for DNN respectively, which indicates a super high speed for the classification.

To evaluate the classification performance, the accuracy is calculated by

$$\text{accuracy} = \frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FN}+N_{FP}} \quad (5)$$

where N_{TP} stands for the true positives, N_{TN} stands for true negatives, N_{FN} stands for the false negatives, N_{FP} stands for the false positives.

A. *SVM results analysis*

Fig. 6 shows the results of the SVM classification with different portions of data being trained and tested. For each portion setting, the algorithm was run 20 times independently, thus the training/testing data in each run were randomly selected. The results show the improved accuracy with the increased training numbers in a good agreement with our expectations. The smallest training set of 37 populations used only 1/10 of the input to realize an accuracy over 95%. The testing data from 9/10 of the input were mostly measured in conditions not considered in the training set. The high accuracy demonstrates the excellent generalization ability. The small fluctuations for all the results indicate the high robustness of the SVM method. The training size required for an acceptable accuracy and robustness depends on how complicated the practical measurement would be. In our experiment, 7 disturbing factors were considered as shown in Table. 1 with 372 data collected in total. We show in this case that 1/10 of the input data was sufficient to provide a favorable accuracy, while a drop on the accuracy can be observed compared to the groups with more training numbers. Therefore, the accuracy may further reduce if fewer training data are used, as the algorithm may not be able to extract the features from a very small training dataset. In measurements with more disturbing factors, the training size should be increased appropriately to adapt with the increased complexity.

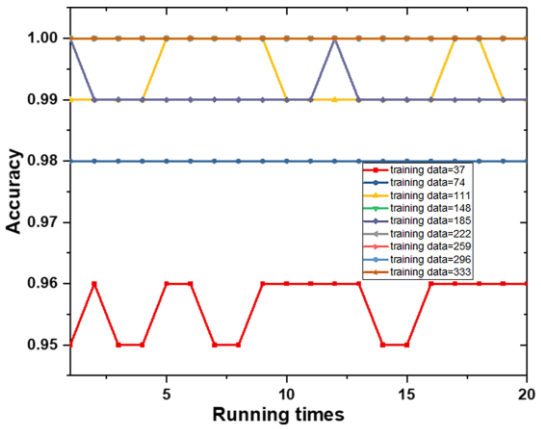


Fig.6. With the best selected parameters, the test classification accuracy versus the number of running times under different number of training data.

B. *DNN results analysis*

For the assessment of the DNN performance, 40% of the both glucose and lactose data (149) were randomly selected as the test data while the others were used as the training data. The accuracy is calculated after every epoch training. Note that one epoch means all the training data have been used to train the network once. Fig.7 is the Cross-entropy Loss value versus the epoch in the training process. It clearly shows that loss value decreases after every training epoch and gets nearly saturated after 70 epochs, indicating that the parameters in the network have been adjusted properly to correctly classify the training data. Tests were executed after every epoch and the accuracy was calculated in Fig. 8. The accuracy increases with the training epochs, which matches well with our expectation. The accuracy is nearly saturated after 70 epochs, with an average accuracy of 0.897 and a standard deviation of 0.0242 from 70th epoch to 100th epoch. The number of misclassified glucose and lactose spectrums in the last ten epochs is shown in Fig.9. As the test dataset contain nearly identical number for both glucose and lactose data, the almost equal misclassified results for glucose and lactose indicate the DNN has no specific tendency in the classification accuracy of the two samples.

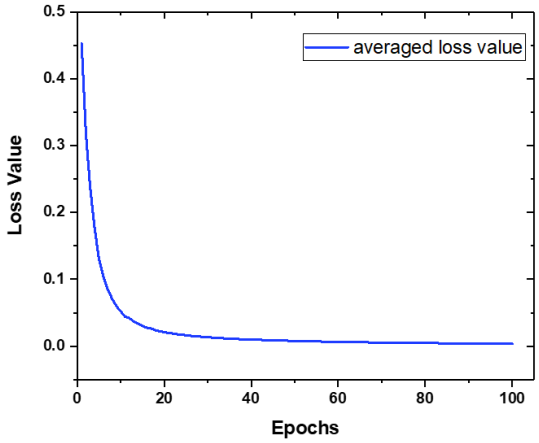


Fig.7. Loss value versus the epochs in the DNN training process.

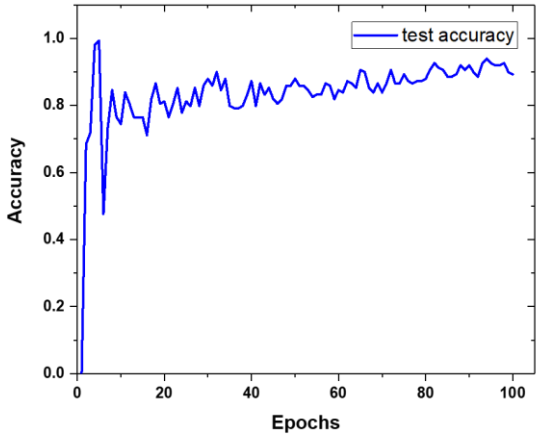


Fig.8. The test accuracy versus the epochs. After every epoch, the test data is tested once on the DNN.

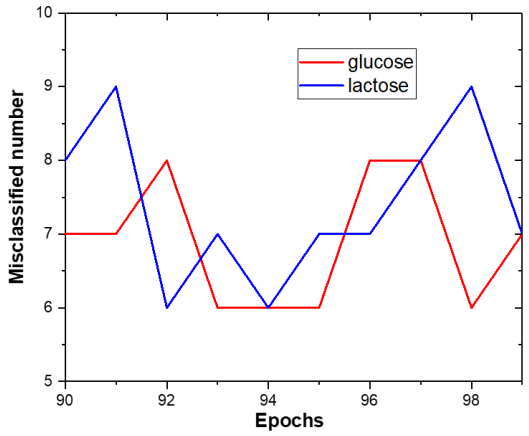


Fig.9. The number of misclassified glucose and lactose data in the last ten epochs.

VI. CONCLUSION

In this paper, we proposed the application of SVM and DNN methods for classifying glucose and lactose THz spectra under various imperfect measurement conditions. The introduced disturbance factors of different concentrations, thicknesses, frequency resolutions, alignments and blocking simulate non-ideal detection environments in practical applications. The parameters in the algorithms are finely adjusted to achieve a high accuracy and generalization ability. The experimental results show a testing accuracy of 99% for the SVM method and 89.6% for the DNN method. As the training is done offline, the testing is very fast for both methods and requires very small computational resource, thus instant classification can be achieved. In our current experiment, due to the relatively small amount of the data (372), DNN has a lower performance than the SVM. The performance of the DNN is expected to be further improved if more independent training data are measured under different conditions can be provided. Both the proposed SVM and DNN methods can be adaptively extended to many other types of samples that exhibit fingerprint features in the THz range. The promising experimental results demonstrate the capability of accurately classifying different THz spectra measured in complicated environments. As a proof of concept, transmission was used in this work to straightforwardly characterize the sample characteristic absorptions. The actual application of material recognition could be more feasibly conducted in a reflection mode. In this case, the reflectivity is mainly determined by the refractive index. Materials with fingerprint absorptions also exhibit feature in their refractive index, thus the proposed method can also be applied. In conclusion, the proposed technique offers an efficient solution for material recognitions in unpredictable and non-ideal situations, which is in high demand for various potential applications of THz spectroscopy, especially in the field of security.

REFERENCES

[1] W. L.Chan, J.Deibel, andD. M.Mittleman, “Imaging with terahertz radiation,” *Reports Prog. Phys.*, vol. 70, no. 8, pp. 1325–1379, 2007.

[2] W.Withayachumnankul, B.Ferguson, T.Rainsford, D.Findlay, S. P.Mickan, andD.Abbott, “T-ray relevant frequencies for osteosarcoma classification,” *Photonics Des. Technol. Packag. II*, vol. 6038, no. January 2006, p. 60381H, 2005.

[3] B.Ferguson, S.Wang, D.Gray, D.Abbott, andX. C.Zhang, “Identification of biological tissue using chirped probe THz imaging,” *Microelectronics J.*, vol. 33, no. 12, pp. 1043–1051, 2002.

[4] P. H.Siegel, “Terahertz technology in biology and medicine,” *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 10, pp. 2438–2447, 2004.

[5] H.Kita, K.Okamoto, andS.Mukai, “Dielectric properties of polymers containing dispersed TCNQ salts,” *J. Appl. Polym. Sci.*, vol. 31, no. 5, pp. 1383–1392, 1986.

[6] X.Chen andE.Pickwell-MacPherson, “A Sensitive and Versatile Thickness Determination Method Based on Non-Inflection Terahertz Property Fitting,” *Sensors*, vol. 19, no. 19, p. 4118, Sep.2019.

[7] M.Naftaly andR. E.Miles, “Terahertz time-domain spectroscopy of silicate glasses and the relationship to material properties,” *J. Appl. Phys.*, vol. 102, no. 4, 2007.

[8] M.Scheller, C.Jansen, andM.Koch, “Analyzing sub-100-µm samples with transmission terahertz time domain spectroscopy,” *Opt. Commun.*, vol. 282, no. 7, pp. 1304–1306, 2009.

[9] M.Mikerov, J.Ornik, andM.Koch, “Removing Water Vapor Lines from THz TDS Data Using Neural Networks,” *IEEE Trans. Terahertz Sci. Technol.*, no. c, pp. 1–1, 2020.

[10] X.Yin, B. W. H.Ng, B. M.Fischer, B.Ferguson, andD.Abbott, “Support vector machine applications in terahertz pulsed signals feature sets,” *IEEE Sens. J.*, vol. 7, no. 12, pp. 1597–1607, 2007.

[11] J.Shi *et al.*, “Automatic evaluation of traumatic brain injury based on terahertz imaging with machine learning,” *Opt. Express*, vol. 26, no. 5, p. 6371, 2018.

[12] Y.Sun *et al.*, “Quantitative characterization of bovine serum albumin thin-films using terahertz spectroscopy and machine learning methods,” *Biomed. Opt. Express*, vol. 9, no. 7, p. 2917, 2018.

[13] K. I.Kim, K.Jung, S. H.Park, andH. J.Kim, “Support vector machines for texture classification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1542–1550, 2002.

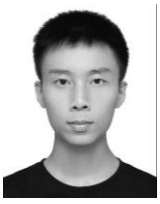
[14] D.Ciregan, U.Meier, andJ.Schmidhuber, “Multi-column deep neural networks for image classification,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. February, pp. 3642–3649, 2012.

[15] T. F.Gonzalez, “Handbook of approximation algorithms and metaheuristics,” *Handb. Approx. Algorithms Metaheuristics*, pp. 1–1432, 2007.

[16] M. E.Mavroforakis andS.Theodoridis, “A geometric approach to support vector machine (SVM) classification,” *IEEE Trans. Neural Networks*, vol. 17, no. 3, pp. 671–682, 2006.

[17] K. H. J. H. M.Sameer Pradhan Wayne Ward andD.Jurafsky, “Shallow Semantic Parsing using Support Vector Machines,” *Naacl-Hlt 2004*, 2004.

Kaidi Li received the B.Eng. degree in electronic information engineering from Hunan University, Hunan, China, in 2017. Since 2017, he has been working in Prof Macpherson’s Terahertz group, Department of Electronic Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong.
His research interests include terahertz biomedical application and terahertz device development.



Xuequan Chen received the B.Eng. degree (Honors) from University of Electronic Science and Technology of China in 2014. After that, he joined Prof. Emma Pickwell-MacPherson's Terahertz group in the Chinese University of Hong Kong for his PhD

research and completed his PhD degree in 2018. He is now a postdoctoral fellow and continues his research in fast and accurate terahertz spectroscopy and imaging, ellipsometry and advanced terahertz devices.

Rui Zhang received the B.Eng. degree in electronic information engineering from Xidian University, Shanxi, China, in 2010. He received his Ph.D. degree in mechanics from Peking University, Beijing, China, in 2016.

From 2016-2018, he was a Postdoctoral Fellow with Prof. MacPherson's Terahertz Group, Department of Electronic Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong. His research interests include terahertz spectroscopy and imaging, biomedical application of terahertz technique, terahertz wave generation. He is now an Assistant Professor at Shenzhen Institute of Advanced Technology.

Emma Pickwell-MacPherson received the undergraduate degree in natural sciences and M.Sc. degree in physics (specialized in semiconductor physics) from Cambridge University, Cambridge, U.K. She started working toward the Ph.D. degree with the Semiconductor Physics Group, Cambridge University, and TeraView, Ltd., Cambridge, a company specializing in terahertz imaging, in 2002 and received the degree in 2005. Her Ph.D. work focused on understanding contrast mechanisms in terahertz images of skin cancer.

She was with TeraView, Ltd., as a Medical Scientist until moving to Hong Kong in 2006. She set up the first terahertz laboratory at The Chinese University of Hong Kong (CUHK) during her post between 2006 and 2009 as an Assistant Professor. Dr. MacPherson has been representing Hong Kong on the International Organising Committee for the Infrared and Millimeter Wave and Terahertz Wave (IRMMW-THz) Conference Series since 2009 and she was the General Conference Chair of the 2015 IRMMW-THz Conference held at CUHK. She joined the Physics Department at Warwick University in October 2017.