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

Kaidi Li[1], Xuequan Chen[1]and Emma Pickwell-MacPherson[1,2]
[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2] Department of Physics, The University of Warwick, Coventry, CV4 7AL, United Kingdom

***Abstract*— We propose an approach based on support vector machine(SVM) and deep neural network(DNN) to classify the chemical substances under different experimental conditions in terahertz time-domain spectroscopy (THz-TDS). 372 groups of independent signals under different conditions were measured to provide a sufficient training set. 99% accuracy for the SVM and 89.6% for the DNN method are realized in the test set. These excellent classification results show the high potentials in chemical recognition, security detection or clinical diagnosis.**

I. INTRODUCTION

Due to the non-invasive property and the ability to resolve many compounds absorption features, terahertz technology has been promising for chemical recognition such as drugs or explosives[1]. A key challenge is to accurately classify detected spectra collected from unknown practical environments, rather than that from an ideal lab condition. To solve this, SVM and DNN methods are implemented and compared [2,3]. Uncorrelated glucose and lactose spectrums are collected by THz-time domain spectroscopy in transmission mode under different experiment conditions to simulate the situations in practical applications. These spectrums vary a lot under different conditions and can't be recognized by people. 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.

II. METHODS AND RESULTS

Table 1 shows our designed experimental conditions to acquire uncorrelated data which simulate the practical detection environment. During the fabrication process, the diameter of the pellet is set to be the same. So the weight of the pellet is proportional to the thickness, and larger weights result in greater attenuation to the THz light, causing lower SNR.

Polyethylene (PE) is the selected doping material and the doping level was set to be 0% or 50% during the fabrication process to simulate different sample concentrations. In the practical experiment, 100ps and 200ps time length, corresponding to 10 GHz and 50 GHz frequency resolution 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. 0^0 or 30^0 incident angle were also achieved by tilting the pellet. At the focus point or 3 cm after or before the focus point (the parabolic mirror has a f-number of 1 and a focal length of 10 cm) were also set to mimic the unideal measurement conditions. These settings test the algorithm performance on irregular sample shapes and poor alignments. Disturbances like placing paper, quartz or a silicon wafer in front of the pellet were also introduced to explore 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 applications, and they also greatly increase the diversity of the data.

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	

To evaluate the classification performance, the accuracy is calculated by

$$\text{accuracy} = \frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FN}+N_{FP}}$$

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.

In the SVM setups, a Gaussian kernel is selected because of the better classification performance after comparison with other kernels. Best suitable parameters are fixed by the parametric search. 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. As shown in Fig. 1, the small fluctuations and high accuracy indicate the stability and excellent performance of SVM method.

For the assessment of the DNN performance, 40% of the total data (149) were randomly selected as the test data while the others were used as the training data. The accuracy is based on the test data calculated after every epoch training. Note that one epoch means all the training data have been used to train the network once. As shown in fig.2, the accuracy increases with the training epochs, which matches well with our expectation. N The accuracy is nearly saturated after 70 epochs, with an average accuracy of 0.897, and a standard deviation of 0.0242 from 70^{th} epoch to 100^{th} epoch.

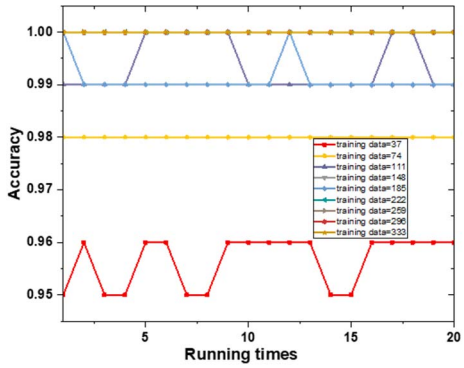

Fig. 1. With the best selected parameters, the test classification accuracy versus the number of running times under different number of training data.

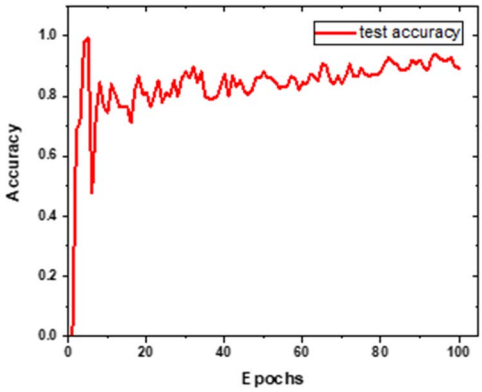

Fig.2. The test accuracy versus the epochs. After every epoch, the test data is tested once on the DNN.

III. ACKNOWLEDGEMENT

The authors would like to thank the research grants council of Hong Kong (project numbers 14206717 and 14201415) for partial support of this work.

REFERENCES

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

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

[3] D.Ciregan, U.Meier, and J.Schmidhuber, "Multi-column deep neural networks for image classification," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., no. February, pp. 3642–3649, 2012.