

Research Article

A Method of Amino Acid Terahertz Spectrum Recognition Based on the Convolutional Neural Network and Bidirectional Gated Recurrent Network Model

Tao Li ¹, Yuanyuan Xu ², Jiliang Luo,² Jianan He,³ and Shiming Lin ⁴

¹School of Mechanical Engineering, Hebei University of Technology, Tianjin 300401, China

²College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China

³Central Laboratory of Health Quarantine,

Shenzhen International Travel Health Care Center and Shenzhen Academy of Inspection and Quarantine,
Shenzhen Customs District, Shenzhen 518033, China

⁴School of Informatics, Xiamen University, Xiamen 361005, China

Correspondence should be addressed to Yuanyuan Xu; yyxu@hqu.edu.cn and Shiming Lin; xmuls@xmu.edu.cn

Received 22 April 2021; Accepted 24 May 2021; Published 29 May 2021

Academic Editor: Yi-Zhang Jiang

Copyright © 2021 Tao Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to improve the accuracy of amino acid identification, a model based on the convolutional neural network (CNN) and bidirectional gated recurrent network (BiGRU) is proposed for terahertz spectrum identification of amino acids. First, we use the CNN to extract the feature information of the terahertz spectrum; then, we use the BiGRU to process the feature vector of the amino acid time-domain spectrum, describe the time series dynamic change information, and finally achieve amino acid identification through the fully connected network. Experiments are carried out on the terahertz spectra of various amino acids. The experimental results show that the CNN-BiGRU model proposed in this study can effectively realize the terahertz spectrum identification of amino acids and will provide a new and effective analysis method for the identification of amino acids by terahertz spectroscopy technology.

1. Introduction

Terahertz (THz) waves are electromagnetic waves spanning frequencies between 0.1 THz and 10 THz, occupying most of the electromagnetic spectrum between microwaves and infrared light waves [1]. The combination of the vibration and rotation frequency transition of biomolecules is in the terahertz band, and thus, the terahertz spectrum can reflect the subtle changes in molecular species and structure; this is called molecular fingerprinting [2]. The energy of terahertz photons is low, and the photon energy of the electromagnetic wave with a frequency of 1 THz can be set to about 4.1 meV, which can directly detect biological samples without destroying their structure. Compared with other detection methods, terahertz spectroscopy can realize the label-free, fast, and nondestructive detection of biomolecules

[3]. Therefore, it has great application potential in the field of biomedicine, and terahertz radiation has recently been used to study DNA, RNA [4, 5], amino acids [6], proteins [7], and other biomolecules [8].

Amino acid is the basic building blocks of proteins, and various proteins in living organisms are composed of 20 basic amino acids. Amino acids can be used not only as a marker for diseases but also as a therapeutic drug. Amino acids are representative biomolecules, and their rapid nondestructive detection and quantitative analysis are particularly important [9].

The terahertz fingerprint spectra of a large number of substances are superimposed on each other, which make the qualitative and quantitative analyses based on the spectra extremely difficult. Researchers mainly use machine learning or multivariate analysis (stoichiometry) to achieve

qualitative and quantitative identifications of biological samples [10]. Such methods include multiple linear regression (MLR), principal component analysis, and partial least squares [11, 12]. Ueno et al. [13] conducted a quantitative analysis on a mixture of different amino acids in 2006. In 2016, Lu et al. used partial least squares (PLS) and interval partial least squares (IPLS) regression to quantitatively analyze binary amino acids [9]. In order to improve the accuracy of quantitative analysis, some research groups have proposed the use of machine learning methods [10, 14]. Yuan et al. [14] performed a spectral classification of three fluoroquinolones based on the back propagation neural network (BPNN) and obtained an accuracy rate of 80.56%. Meanwhile, Peng et al. [15] used wavelet filtering combined with support vector machines to quantitatively analyze the main components of brain tissue; the root mean square error of this approach was 0.4%. Furthermore, Liu et al. [16] used the random forest (RF) algorithm to distinguish genetically modified rice seeds from nontransgenic rice seeds, and the classification accuracy of their model reached 96.67%.

Methods based on machine learning have a common problem, the need to manually extract features. This process is complicated and cumbersome, and some methods have limited nonlinear fitting capabilities, resulting in the lack of hidden information in the extracted features. In contrast, deep learning is very suitable for terahertz signal recognition, and they do not need manual extraction of features.

Among the deep learning algorithms, the bidirectional gated recurrent network (BiGRU) is a type of the bidirectional recurrent neural network model that can fully express the relationship between the current output of a sequence and previous information [17]. However, the characteristic dimension of the amino acid time series is too high, and the BiGRU is directly used to process the amino acid sequence parameters, which results in low efficiency. In addition, the convolutional neural network (CNN) has powerful feature extraction capabilities. To a certain extent, the more the layers, the more advanced the extracted features, the more the information contained, and the better the final classification result. In addition, the CNN requires fewer hyperparameters, has low computational complexity, and is widely used in image processing, semantic segmentation, and recognition.

Based on this, the present study uses the representative CNN and BiGRU to establish a CNN-BiGRU recognition model and makes full use of the advantages of the two networks to classify amino acids. First of all, we use the CNN to extract the abstract features of amino acid time series and then use BiGRU's dynamic timing information modeling ability to process the feature vectors of amino acid sequences.

Finally, experiments are carried out on the terahertz spectra of various amino acids. The experimental results show that the CNN-BiGRU model proposed in this study effectively realizes the terahertz spectrum identification of amino acids and avoids the tedious steps of artificial feature extraction, selection, and dimensionality reduction. Moreover, it demonstrates itself as a suitable terahertz spectrum technology for the identification of amino acids.

2. Methodology

2.1. CNN. Deep learning can ensure effective information extraction and feature expression and can complete tasks such as image recognition, time series prediction, and text prediction. Typical deep learning networks include the CNN, deep belief network, and recurrent neural network. Among them, the CNN can automatically learn filters and has developed into a mature feature extractor.

The CNN was proposed by Lecun et al. [18], and it is a feedforward multilayer neural network. It uses convolution operations to greatly reduce the dimensionality of the data and can achieve abstract expression of the original data [19, 20]. The basic structure of a CNN includes input, convolution, pooling, fully connected, and output layers, as shown in Figure 1.

The essence of the CNN is to construct multiple filters that can extract data features and the topological structure features hidden between the data using layer-by-layer convolution and pooling operations on the input data. Finally, these abstract features are merged through a fully connected layer, and the classification problem is solved through a Softmax or sigmoid activation function [21].

The convolution layer convolves the information in the receptive field by designing a convolution kernel of a suitable size and abstractly expresses the original data. When inputting data X , the feature map of the convolutional layer can be expressed as follows:

$$C = f(X \otimes W + b), \quad (1)$$

where \otimes is the convolution operation, W is the weight vector of the convolution kernel, b is the offset, and $f(\cdot)$ is the activation function, which can be tanh or ReLU.

The pooling layer downsamples the convolution output, extracts strong features and removes weak features, reduces the number of parameters, and prevents overfitting.

The fully connected layer performs regression classification on the features extracted from the previous layer. Through the weighted summation of the output of convolution and pooling layers and then through the response of activation function, the following formula is obtained:

$$x^l = f(w^l x^{l-1} + b^l), \quad (2)$$

where x^l is the network weight coefficient.

2.2. BiGRU. The BiGRU is a new type of the bidirectional recurrent neural network model. The recurrent neural network (RNN) effectively solves the problem that there is no operational connection between inputs in the hidden layer of traditional neural networks and can be used for the prediction of time series data and text semantic data. Its structural diagram is shown in Figure 2. However, the RNN uses a back propagation algorithm. When learning a long time sequence, the RNN may have gradient disappearance and gradient explosion problems, and thus, it is unable to grasp the nonlinear relationship of complex time series data with a long time span.

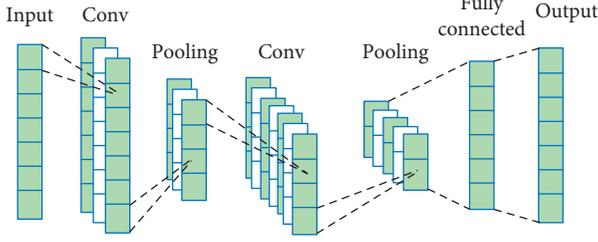


FIGURE 1: Structural diagram of the CNN.

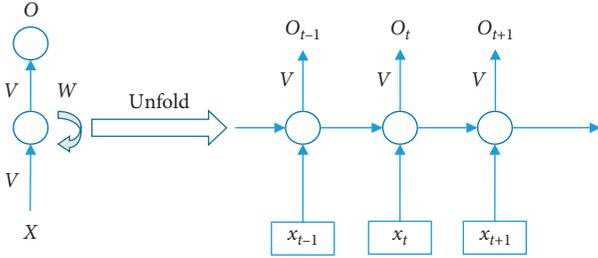


FIGURE 2: Structural diagram of an RNN unit.

When the recurrent neural network processes time series data, the transmission of its state is from front to back. However, in most complex time series data, the output at the current moment is not only related to the previous state but also related to the subsequent state. Schuster [22] proposed the bidirectional RNN (BiRNN) to solve the above problems. The basic idea is that each training sequence is completed by two forward and backward recurrent neural networks, and these two RNNs are connected to the same output layer. The output layer contains the complete information of the past and the future of each point in the input sequence, forming a double-loop network structure. The BiRNN brings a certain improvement compared with the ordinary RNN [23].

The bidirectional gated recurrent unit is a BiRNN based on the gated recurrent unit. In the BiGRU, two GRU inputs in opposite directions are provided at the same time at each time t . GRUs in the two directions are not directly connected, and the output is jointly determined by two unidirectional GRUs. The BiGRU model has good prediction performance in nonlinear time series data [24, 25]. The structure of the BiGRU is shown in Figure 3.

In the forward layer, we calculate the forward direction from time 1 to time t and obtain and save the output h_t of the forward hidden layer at each time. In the backward layer, the reverse calculation is performed from time t to time $t-1$, and the output h'_t of the backward hidden layer at each time is obtained and saved. Finally, the final output o_t is obtained by combining the output results of the forward layer and backward layer at each time. The mathematical expressions are presented as follows.

- (1) In the forward layer, the output h_t from front to back is

$$h_t = f(w_1 x_t + w_2 h_{t-1}). \quad (3)$$

- (2) In the backward layer, the output h'_t from back to front is

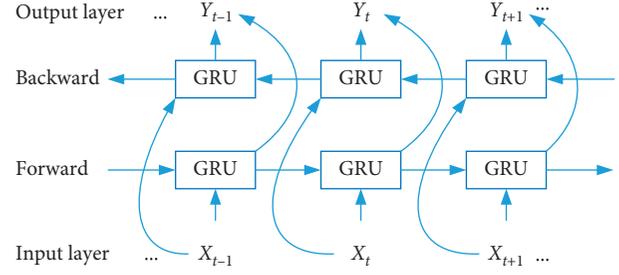


FIGURE 3: Structural diagram of the BiGRU.

$$h'_t = f(w_3 x_t + w_5 h'_{t-1}). \quad (4)$$

- (3) Combining the forward layer and the backward layer, the final output of the BiGRU is

$$o_t = g(w_4 h_t + w_6 h'_t), \quad (5)$$

where w_1 , w_2 , w_3 , w_4 , w_5 , and w_6 , respectively, represent the weights corresponding to the forward and reverse hidden states.

2.3. Model Building. Deep learning-based models have the ability to automatically extract features. CNN models usually rely on the convolution kernel of the convolutional layer to extract features. However, the existence of the convolution kernel limits the long-term dependency problem of the CNN when processing time series data [26]. In this study, the introduction of the BiGRU can effectively solve this problem, and we can capture the dependencies before and after the time series. In view of the high feature dimension of amino acid time series, we first use the CNN to extract the features of amino acid time series. Then, the BiGRU is used to process the feature vector of the amino acid sequence. Finally, amino acid sequence identification is achieved through the fully connected network. The model structure is shown in Figure 4.

This model is composed of three modules: the CNN, the BiGRU, and the recognition network. The CNN consists of five convolutional layers and five average pooling layers. The two-way long- and short-term memory network is composed of three layers of BiGRUs, and each layer is composed of 512, 256, and 96 hidden units. The identification network consists of one discarding layer (the discarding rate parameter is set to 0.35), one fully connected layer, and one Softmax. Finally, an amino acid recognition model is obtained.

3. Experimental Equipment and Samples

3.1. Experimental Equipment and Sample Preparation. The experimental device uses a fiber-type terahertz time-domain spectroscopy system (THz-TDS) with a signal-to-noise ratio of up to 70 dB, as shown in Figure 5. The device consists of a femtosecond laser, an antenna that excites and detects terahertz radiation, a delay line, and a lock-in amplifier. The center wavelength of the ultrashort pulse fiber

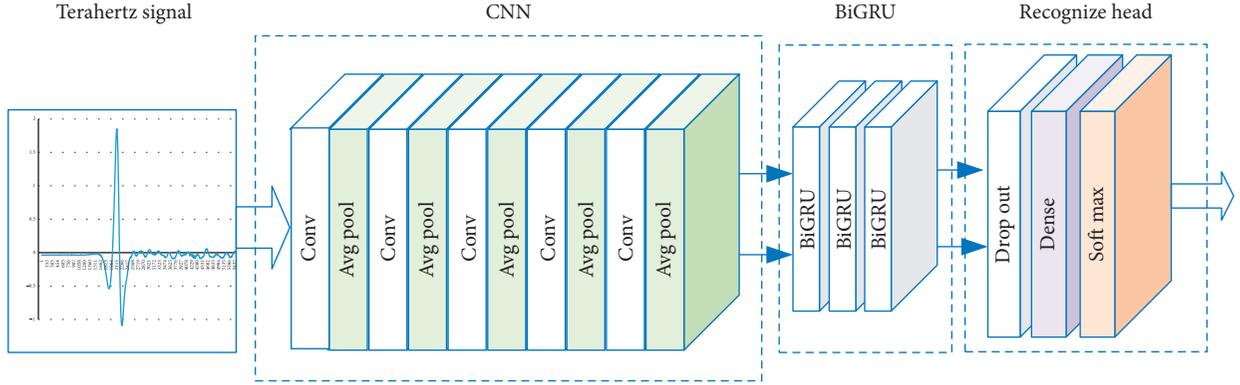


FIGURE 4: Structural diagram of amino acid terahertz spectrum recognition model based on the CNN and BiGRU.

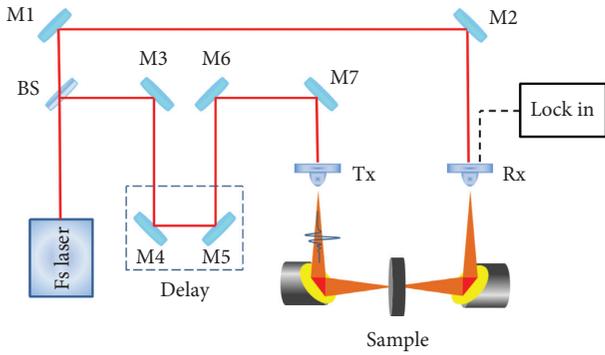


FIGURE 5: Schematic diagram of the terahertz time-domain spectroscopy system.

laser pulse is 1560 nm, and the repetition frequency is 100 MHz. In order to obtain higher resolution, the measurement time is 53 ps.

The three amino acid samples of glutamic acid, glutamine, and asparagine were provided by Shanghai Aladdin Reagent Company. The samples were baked for 24 h at 50°C and then ground with agate. The particle size was less than 80 μm to reduce the scattering effect. Samples were then mixed with high-density polyethylene powder in different proportions (30 different proportions), and tablets with a diameter of 13 mm were put under a pressure of 20 MPa [20]. The weight of each sample is 120 mg, and the thickness is about 1.2 mm. Each sample was measured at different times to obtain 10 terahertz spectra.

3.2. Evaluation Index. This study uses three evaluation indicators, that is, accuracy, recall, and precision, to evaluate the performance of the model. Accuracy is the proportion of amino acids correctly classified in the total test sample. Accuracy is the most intuitive way to evaluate the performance of the model.

Recall is the true-positive rate (TPR), that is, recall is the proportion of real positive samples to all positive samples that are currently classified into the positive sample category:

$$R = \frac{tp}{tp + fn}. \quad (6)$$

Precision is a measure of accuracy, which represents the proportion of examples that are divided into positive examples:

$$P = \frac{tp}{tp + fp}. \quad (7)$$

F-score is the weighted harmonic average of precision and recall:

$$F = \frac{(a^2 + 1)P * R}{a^2(P + R)}, \quad (8)$$

when $\alpha = 1$, then the F-score is F1:

$$F1 = \frac{2PR}{P + R}. \quad (9)$$

Area under curve is defined as the area under the receiver operating characteristic (ROC) curve and surrounded by the coordinate axis. For the ROC, the abscissa is the false-positive rate (FPR), and the ordinate is the TPR; therefore, when the TPR is larger and the FPR is smaller, the classification result is better.

4. Results and Discussion

In order to verify the effectiveness of the CNN-BiGRU model proposed in this article, we use the BiGRU [27], PCA-SVM [28], PCA-LSTM [29], and CNN-LSTM [30] for comparison. In the experiment, the training and validation sets totaled 1000, including 80% of the training set, 10% of the validation set, and 10% of the test set. The accuracy rate, recall rate, precision rate, and F1-score for each model are given in Table 1.

The ROCs of each model are shown in Figures 6–9.

The nonlinear fitting ability of the traditional machine learning method is very limited, and it may not be able to extract high-level and high-resolution features accurately. On the contrary, it will omit important information in denoising and feature extraction. At the same time, the effect of the SVM in classification is general, and thus, the PCA-SVM model is the worst in all indicators.

LSTM is specially designed for time series, but LSTM alone does not achieve good results because it overfits the

TABLE 1: Summary of evaluation indices of each model.

Model/index	Accuracy rate	Recall rate (%)	Precision (%)	F1-score (%)
PCA-SVM	68.35%	68.35	68.35	56.08
BiGRU	91.56%	91.56	91.56	97.82
PCA-LSTM	97.89%	97.89	97.89	97.82
CNN-BiGRU	99.16%	99.16	99.16	99.17

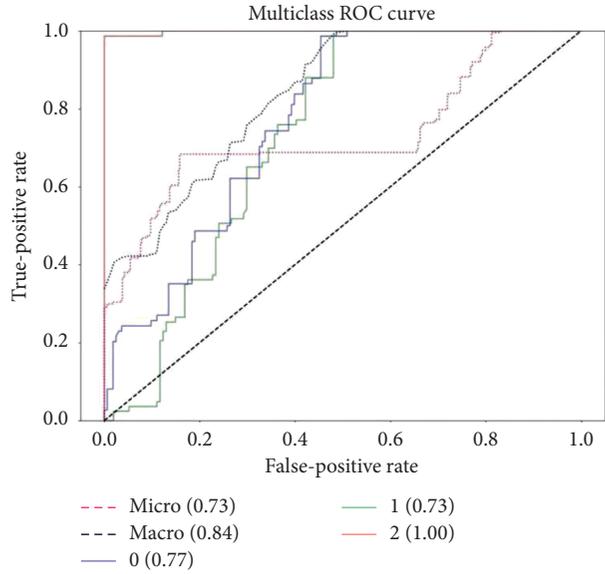


FIGURE 6: PCA-SVM ROC.

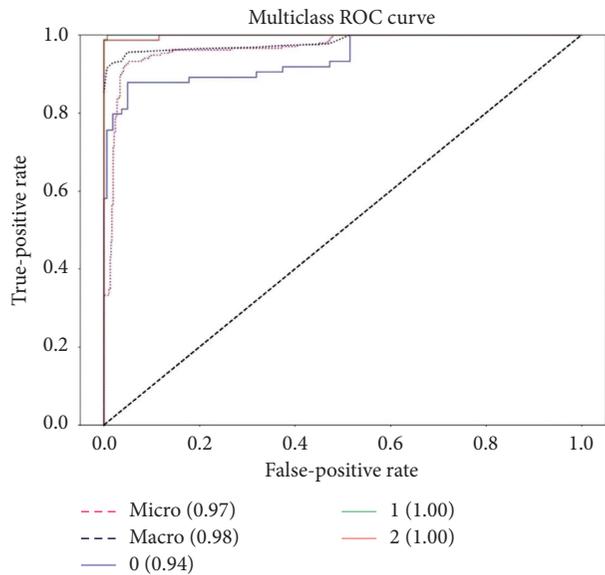


FIGURE 7: PCA-LSTM ROC.

sequence characteristics of amino acids. The main reason why PCA-LSTM is better than LSTM is that PCA compresses features and eliminates most redundant data.

The CNN-BiGRU model was used to input the amino acids after a simple pretreatment. The morphological features of basic acids were extracted by the CNN, and then, the temporal features of amino acids were extracted by the

BiGRU. The two features were combined to mine the hidden deep information in amino acids. The average accuracy of classification test results was 99.16%. It can be seen from table that our model achieved the best results in all indicators. The main reason for this is that the CNN can provide more features, and the BiGRU can consider the relationship between features.

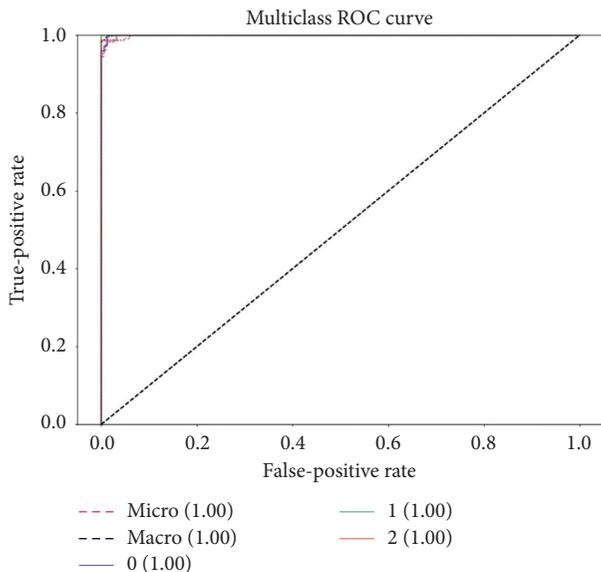


FIGURE 8: BiGRU ROC.

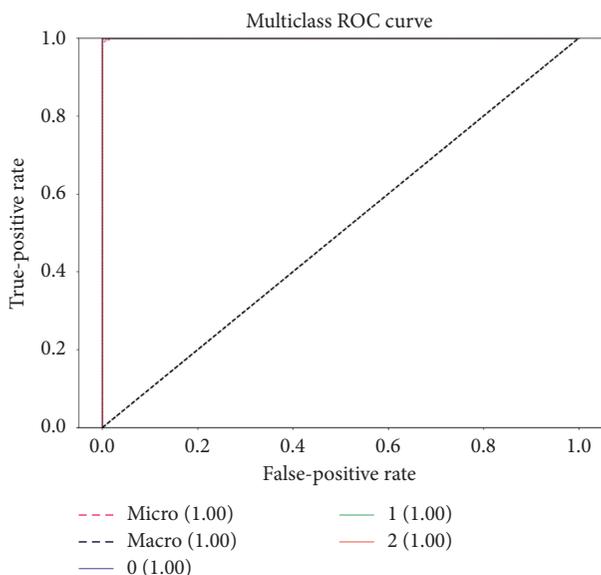


FIGURE 9: CNN-BiGRU ROC.

5. Conclusion

Because it is difficult to conduct a large number of experiments in the experimental environment, the number of samples of various amino acids is small in the present study, which has a certain impact on the experimental results. However, the relative effect of the tested models is not affected. In future, we will conduct more experiments and provide samples of more types of amino acids.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Key R&D Program of China (2018YFC0809200).

References

- [1] P. H. Siegel, "Terahertz technology," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 3, pp. 910–928, 2002.
- [2] Y. Xiang, Z. Xiang, Y. Ke et al., "Biomedical applications of terahertz spectroscopy and imaging," *Trends in Biotechnology*, vol. 34, no. 10, pp. 810–824, 2016.
- [3] K. Serita, E. Matsuda, K. Okada, H. Murakami, I. Kawayama, and M. Tonouchi, "Invited Article: terahertz microfluidic chips sensitivity-enhanced with a few arrays of meta-atoms," *Appl Photonics*, vol. 3, no. 5, Article ID 051603, 2018.
- [4] W. Zhang, E. R. Brown, M. Rahman, and M. L. Norton, "Observation of terahertz absorption signatures in microliter DNA solutions," *Applied Physics Letters*, vol. 102, no. 2, pp. 219–229, 2013.
- [5] E. R. Brown, E. A. Mendoza, D. Xia, and S. R. J. Brueck, "Narrow THz spectral signatures through an RNA solution in nanofluidic channels," *IEEE Sensors Journal*, vol. 10, no. 3, pp. 755–759, 2010.
- [6] M. R. Kutteruf, C. M. Brown, L. K. Iwaki, M. B. Campbell, T. M. Korter, and E. J. Heilweil, "Terahertz spectroscopy of short-chain polypeptides," *Chemical Physics Letters*, vol. 375, no. 3-4, pp. 337–343, 2003.
- [7] L. Xie, Y. Yao, Y. Ying et al., "The application of terahertz spectroscopy to protein detection: a review," *Applied Spectroscopy Reviews*, vol. 49, no. 6, pp. 448–461, 2014.
- [8] J. Neu, E. A. Stone, J. A. Spies et al., "Terahertz spectroscopy of tetrameric peptides," *The Journal of Physical Chemistry Letters*, vol. 10, no. 10, pp. 2624–2628, 2019.
- [9] S. Lu, X. Zhang, Z. Zhang, Y. Yang, and Y. Xiang, "Quantitative measurements of binary amino acids mixtures in yellow foxtail millet by terahertz time domain spectroscopy," *Food Chemistry*, vol. 211, pp. 494–501, 2016.
- [10] J. El Haddad, B. Bousquet, L. Canioni, and P. Mounaix, "Review in terahertz spectral analysis," *TrAC Trends in Analytical Chemistry*, vol. 44, pp. 98–105, 2013.
- [11] J. El Haddad, F. de Miollis, J. Bou Sleiman, L. Canioni, P. Mounaix, and B. Bousquet, "Chemometrics applied to quantitative analysis of ternary mixtures by terahertz spectroscopy," *Analytical Chemistry*, vol. 86, no. 10, pp. 4927–4933, 2014.
- [12] J. B. Sleiman, B. Bousquet, N. Palka, and P. Mounaix, "Quantitative analysis of hexahydro-1,3,5-trinitro-1,3,5, triazine/pentaerythritol tetranitrate (RDX-PETN) mixtures by terahertz time domain spectroscopy," *Applied Spectroscopy*, vol. 69, no. 12, pp. 1464–1471, 2015.
- [13] Y. Ueno, R. Rungsawang, I. Tomita, and K. Ajito, "Quantitative measurements of amino acids by terahertz time-domain transmission spectroscopy," *Analytical Chemistry*, vol. 78, no. 15, pp. 5424–5428, 2006.
- [14] L. Yuan, L. Bin, and L. Huan, "Analysis of fluoroquinolones antibiotic residue in feed matrices using terahertz spectroscopy," *Applied Optics*, vol. 57, no. 3, p. 544, 2018.

- [15] Y. Peng, C. Shi, M. Xu et al., "Qualitative and quantitative identification of components in mixture by terahertz spectroscopy," *IEEE Transactions on Terahertz Science and Technology*, vol. 8, no. 6, pp. 696–701, 2018.
- [16] W. Liu, C. Liu, X. Hu, J. Yang, and L. Zheng, "Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics," *Food Chemistry*, vol. 210, no. 1, pp. 415–421, 2016.
- [17] Y. Xu, G. Yang, J. Luo, and J. He, "An electronic component recognition algorithm based on deep learning with a faster SqueezeNet," *Mathematical Problems in Engineering*, vol. 2020, Article ID 2940286, 11 pages, 2020.
- [18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [19] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," *Proceedings of International Conference on Machine Learning*, vol. 37, no. 7, pp. 2048–2057, 2015.
- [20] Y. Xu, W. Yan, G. Yang, J. Luo, T. Li, and J. He, "CenterFace: joint face detection and alignment using face as point," *Scientific Programming*, vol. 2020, Article ID 7845384, 8 pages, 2020.
- [21] K. Ping-Huan and H. Chiou-Jye, "A high precision artificial neural networks model for short-term energy load forecasting," *Energies*, vol. 11, no. 1, pp. 213–226, 2018.
- [22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions On Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [23] D. She and M. Jia, "A bigru method for remaining useful life prediction of machinery," *Measurement*, vol. 167, no. 1, Article ID 108277, 2020.
- [24] Q. Zhu, F. Zhang, S. Liu et al., "A hybrid VMD-bigru model for rubber futures time series forecasting," *Applied Soft Computing*, vol. 84, no. 1, Article ID 105739, 2019.
- [25] Q. Zhu, F. Zhang, S. Liu, and Y. Li, "An anticrime information support system design: application of K-Means-VMD-Bigru in the city of chicago," *Information & Management*, vol. 17, no. 11, Article ID 103247, 2019.
- [26] D. U. Yongping, X. Zhao, and B. Pei, "Short text sentiment classification based on CNN-LSTM model," *Journal of Beijing University of Technology*, vol. 45, no. 7, pp. 662–670, 2019.
- [27] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, and A. Elmaghraby, "Ensemble deep learning models for heart disease classification: a case study from Mexico," *Information*, vol. 11, no. 4, p. 207, 2020.
- [28] L. I. Yun-Fei, Y. Q. Huang, and G. L. Jiang, "Short-term load forecasting based on PCA-SVM," *Proceedings of the Chinese Society of Universities for Electric Power System and Automation*, vol. 19, no. 5, pp. 66–70, 2007.
- [29] M. Zimmermann, M. Mehdipour Ghazi, H. K. Ekenel, and J. P. Thiran, "Combining multiple views for visual speech recognition," in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP)*, Stockholm, Sweden, August 2017.
- [30] T. N. Nguyen, D. Q. Tran, T. N. Nguyen, and H. Q. Nguyen, "A CNN-LSTM architecture for detection of intracranial hemorrhage on CT scans," arXiv preprint arXiv:2005.10992, 2020.