IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 26, 2018, accepted December 23, 2018, date of publication January 9, 2019, date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2891548

Mass Spectral Substance Detections Using Long Short-Term Memory Networks

JUNXIU LIU[1], (Member, IEEE), JINLEI ZHANG[1], YULING LUO[1], SU YANG[2], JINLING WANG[3], AND QIANG FU[4]

[1]Faculty of Electronic Engineering, Guangxi Normal University, Guilin 541004, China
[2]School of Computing, Engineering and Intelligent Systems, Ulster University, Londonderry BT48 7JJ, U.K.
[3]School of Computing, Ulster University, Belfast BT37 0QB, U.K.
[4]College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

Corresponding author: Yuling Luo (yuling0616@gxnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61603104, in part by the Guangxi Natural Science Foundation under Grant 2017GXNSFAA198180 and Grant 2016GXNSFCA380017, in part by the funding of Overseas 100 Talents Program of Guangxi Higher Education, in part by the Doctoral Research Foundation of Guangxi Normal University under Grant 2016BQ005, and in part by the Innovation Project of Guangxi Graduate Education under Grant XYCSZ2018080.

ABSTRACT In this paper, mass spectral substance detection methods are proposed, which employ long short-term memory (LSTM) recurrent neural networks to classify the mass spectrometry data and can accurately detect chemical substances. As the LSTM has the excellent understanding ability for the historical information and classification capability for the time series data, a high detection rate is obtained for the dataset which was collected by a time-of-flight proton-transfer mass spectrometer. In addition, the differential operation is used as the pre-processing method to determine the start time points of the detections which significantly improve the accuracy performance by 123%. The feature selection algorithm of Relief is also used in this paper to select the most significant channels for the mass spectrometer. It can reduce the computing resource cost, and the results show that the network size is reduced by 28% and the training speed is improved by 35%. By using these two pre-processing methods, the LSTM-based substance detection system can achieve the tradeoff between high detection rate and low computing resource consumption, which is beneficial to the devices with constraint computing resources such as low-cost embedded hardware systems.

INDEX TERMS Mass spectral substance detections, long short-term memory networks, chemometrics.

I. INTRODUCTION

Unusual substance detections are ubiquitous in daily life with significant importance especially in the security and safety applications, such as hazard detections, environmental monitoring, testing and identification of chemical or biological substances, analyzing inorganic, organic and biological aerosol components, and even explosives and drugs detections [1]. There are two major solutions in detecting the unusual substances: i) the sniffer dogs, and ii) the chemical detector such as mass spectrometer. Although using the sniffer dogs for detection has always been an effective solution [2], it still has some drawbacks, such as the resource cost for the training and feeding, the limited working hours. They are also vulnerable to the deliberate distractions and diseases. The other approach is to use mass spectrometer to measure the chemical or physical properties of the environment to produce time-series mass spectra that describe the current conditions, where the unusual substance is presented by the time related mass spectrometry data. Mass spectrometry data is recorded over time and contains specific patterns of temporal information. The unusual substances can be detected via the anomalies in the mass spectrometry data. Detecting outliers or anomalies in data has been studied from 19th century [3]. Anomaly detection refers to finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred as anomalies, outliers, discordant observations, or exceptions in different application domains [4]. Various anomaly detection techniques have been proposed for particular or general application domains [4]. However due to the data missing, irregular sampling, and different recording length, the anomaly detection has some challenges. For the mass spectrum applications, this problem becomes even worse since the sensors often change their properties over time, leading to increment of the complexity

degree in the mass-spectra. Therefore the identification of unusual substance in the mass spectrometry data is investigated in this paper.

Mass spectrum can be used to analyze a wide range of compounds. The development of different ionization techniques allows the analysis of gas, liquid and solid samples without considering the nature of samples (e.g. metallic, inorganic, organic, polymeric or biological) [1]. The mass spectrum is a plot of the relative abundance of each ion versus mass-to-charge (m/z) ratio, which can be generated from the mass spectrometer. The mass spectrometer is designed to convert neutral atoms or molecules into a beam of positive or negative ions, to separate the ions on the basis of their mass-to-charge (m/z) ratio, and to measure the relative abundance of each type of ion [5]. Using the mass spectrum, both the molecular mass and the molecular formula of an unknown compound can be determined. In the real applications, a small volume of substance is injected into the mass spectrometer, which then actuates the data acquisition. However, the mass spectrum data from a mass spectrometer is always affected by electronic noise, sensor drift etc. [6]. Simultaneously, the mass spectrum of each elemental distribution of the substances is not completely pure and is disturbed by the presence of isotopes. Some mass spectrometers also need manually calibrations after working for a period of time [6]. Hence, it is important to extract the key features from complete spectra, especially in the presence of noise and distortion. This is also beneficial to improve the system integration and portability. In addition, the ion abundances measured by the mass spectrum are not ideal inputs for classifiers, since they do not correlate well with the presence of structural features in different compounds [7]. It is therefore desirable to transform the raw spectrum into a more suitable feature set. Although it is assumed that using all the features as inputs to the neural network may give a good result, but in practice this leads to a quite large neural network and unnecessarily long training time [6]. Thus, to select and feed the features with significant characteristic for classification to the neural network is the key to improve the system performance. As a feature selection method, the Relief algorithm and its variants are known to be relatively efficient in practice, which can estimate features according to how well their values distinguish among the instances that are near to each other [8]. The Relief algorithm is initially proposed and used for binary classification [9]. The aim is to seek the two nearest neighbors from both the same and different classes, which are defined as nearest hit and nearest miss, respectively. Based on Relief, the feature selection process is improved in the approach of Kononenko *et al.* [8] which is known as the ReliefF algorithm, i.e. an variant of the Relief algorithm. It can cope with incomplete and noisy data, and solve multi-class problems [10].

After the mass spectrometry data is pre-processed, they can be feed into the detection system. The artificial neural network (ANN) processes information by imitating the structure of the neural network in the brain. It is an important machine learning technique and has excellent performance for the classification tasks [11]. It is composed of input, hidden and output layers, where the data transmission between layers has one-way propagation. The multi-layer feed forward artificial neural networks (FFNNs) have been used for spectroscopy [12], where the data reduction, robust regression, and instrumental drifts were also considered. However the ANNs and other types of feed forward neural networks such as FFCC [13] also have some constraints [14]. For example, it is a challenge to design an ANN with appropriate size and structure, and this should be based on three aspects: the complexity of the solution, the desired prediction accuracy, and data characteristics [15]. For the first two aspects, i.e. precision and complexity, good performance can be obtained by using the FFNNs. However, for the data characteristics, Recurrent Neural Networks (RNNs) is found to be more suitable than FFNNs [16]. FFNNs face over-fitting, convergence problems, and are difficult for implementation when they are applied in the time series data analysis [17]. In the conventional neural network model, the flow of information is from the input to the hidden layers and then to the output layer. The pre and post-layer are fully connected, and there is no connections between the neurons in the same layer. This neural network architecture cannot achieve a good performance for some applications such as the time series data processing tasks [11]. Compared to the conventional neural network, the RNN adds a weighted sum of the hidden layer with the previous input when calculating the output of the hidden layer. Therefore the input of the hidden layer includes not only the output of the pre-layer, but also the previous output of the hidden layer. This introduces a feedback mechanism in the hidden layer to learn the context-related information, which can effectively process the sequence data (e.g. time series). The RNN has been applied to many applications, such as action recognition [18], multilingual machine translation [19] etc. As it is capable of dealing with time series data, this work aims to investigate and design substance detection systems based on the RNNs. As a type of the RNNs, Elman networks use simplified derivative calculations but have some drawbacks for the reliable learning. Recent research show that the long short-term memory recurrent neural networks (LSTMs) [20] achieve a better performance than Elman networks. The main contributions of this work are as follows:

(a). Novel substance detection methods are proposed, which are based on the LSTMs. A good detection performance of time series mass spectrometry data and low computing resource cost are achieved.

(b). By using the differential operation and ReliefF algorithm, the performances of classification accuracy, speed and required computing resources are improved.

(c). Results demonstrate that the detection accuracy of 81.81% is achieved by one of the proposed substance detection system where the dimension of the raw dataset is significantly reduced from 270 to 50, the training speed of the neural network is increased by 35% and the network size is reduced by 28.46%.

The remainder of this paper is organized as follows. Section II provides the motivation and related works. Section III describes the differential operation, ReliefF algorithm and the proposed substance detection system. Section IV provides the results and performance analysis. Section V concludes the paper.

II. MOTIVATION AND PREVIOUS STUDIES

Various learning tasks in practice require dealing with sequential data [21]. Sequence classification is closely related to the sequential supervised learning problem, which is different from the classical supervised learning problem [22]. The sequence imposes an order on the observations which must be preserved during the model training and decision-making. Most of the existing research on detecting anomalies in discrete sequences focus on one of the following three problem formulations [23]: (a) Sequence-based anomaly detection, i.e. detecting anomalous sequences from a database of test sequences; (b) Contiguous subsequence-based anomaly detection, i.e. detecting anomalous contiguous subsequence within a long sequence; (c) Pattern frequency-based anomaly detection, i.e. detecting patterns in a test sequence with anomalous frequency of occurrence. These formulations are fundamentally different, and hence require exclusive solutions. Using neural networks to solve the problems in the chemical application domains has been proposed and implemented in spectroscopy (mass, infrared, nuclear magnetic resonance, ultraviolet), and structure/activity relationships etc. [24]. Results show advantages of high precision and low computing complexity [25]. Therefore in this paper, we focus on the substance detections using the time-series mass spectrometry data.

In previous research, neural network has been used in the fields of mass spectra [6], [26], where good performance is achieved. The FFNNs can classify low-resolution mass spectra of unknown compounds [7]. A method for identification of the structural features of compounds from mass spectrometry data is proposed in the approach of Eghbaldar *et al.* [27], which uses an optimized artificial neural network. Ion mobility spectra is successfully classified through the neural networks [28], which uses a combination drift times, number, intensity and shape of peaks. Based on the selection of the relevant input data, an optimized ANN model is used to analyze instrumentation spectra, where the reduction of the input dimension improves the robustness of the model [29]. However, all these aforementioned networks are based on the feed-forward structure with slight variations, and they are not suitable for the temporal correlated data. In addition to the conventional neural networks, deep neural networks (DNNs) have also been applied to spectral data. Deep learning [30] is a method which can extract features directly from original data. Deep belief network, one of the deep learning methods, has been used to predict molecular substructure in the mass spectral data [31]. They can approximate arbitrary nonlinear functions, which overcomes the limitations of using classical linear methods and is beneficial in time series processing [32]. However the DNNs can only be applied to problems whose inputs and targets are encoded with fixed dimensional vectors. For the sequence classification, the input sequence of the DNN is required to be divided into small overlapping sub-sequences. The time steps of the input sequence become features to the network and the subsequences overlap to simulate a window along the sequence. The limitations include (a). the size of the sliding window is fixed and must be imposed on all inputs, and (b). the size of the output is also fixed. The DNN has capability for sequence classification but still suffer from this key limitation, i.e. to specify the scope of temporal dependence between observations which needs to be done before the model development. This is a constraint as many problems are expressed by sequences whose lengths (dimensions) are unknown in advance [14]. For time-series mass spectrometry data, observing only one mass spectrum at a specific time (i.e. point anomaly detection) is difficult to classify the substances. Time-series data has been extensively investigated by the contextual anomaly detection strategies [33]–[35]. Compared to the point anomaly detection techniques, the contextual anomaly detection can achieve a better performance [4]. The RNN is one of contextual anomaly detection method, which is able to exploit a dynamically changing contextual window over the input sequence history [36]. Furthermore, the LSTMs can solve many time series tasks which are impossible for the FFNNs with fixed time window sizes [37]. The LSTM networks do not need a pre-defined time window and are capable of accurately modeling complex multivariate sequences [38].

In summary, the ANN-based models are widely used in the chemical application domains, especially for analyzing spectra/structure correlations. However, the research by using contextual anomaly detection needs to be further investigated. Recent research show that the RNN is a powerful and practical tool for the supervised learning from sequences [21]. One key challenge of the RNNs is how to train the networks effectively, e.g. to avoid the vanishing and exploding gradients. The LSTM overcomes this challenge [39], thus it is employed in this approach. In the meantime, the computational complexity of an anomaly detection technique should also be considered, especially when it is deployed to the devices with limited computing resources [4]. Therefore, the LSTM-based substance classification system is proposed in this paper where some challenges such as improving detection rate and reducing computing resource overhead will be addressed. This work is a contiguous subsequence-based, multi-classification anomaly detection system [23], where the high accuracy, low overhead, fast response and sensitivity will be obtained. It will be described in detail in the following sections.

III. THE LSTM-BASED MASS SPECTRAL SUBSTANCE DETECTIONS

The chemometric community aims to analyze instrumentation spectra efficiently and accurately, and overcome the

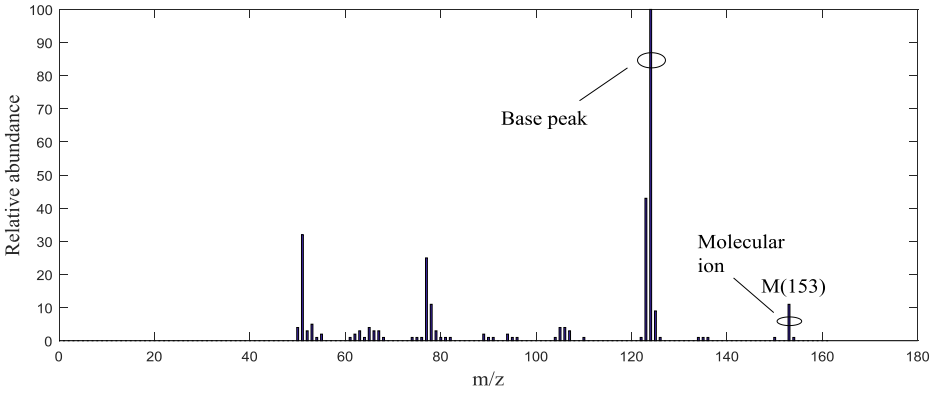


FIGURE 1. Mass spectrum of dopamine.

noise and instrumental drifts [29]. Chemometric uses mathematics, statistics and formal logic to design, select optimal experimental procedures, to provide maximum relevant chemical information by analyzing chemical data, and to obtain knowledge about chemical systems [40]. It is important to discriminate the composition of substance by estimating the material activity ratio in spectra. If the redundant and irrelevant elements of spectra are not completely separated, they will interfere with the determination of the substance type. Therefore, the overlapping of peaks makes the analysis of spectra and the interpretation of results difficult. In addition, in order to detect the substance types quickly and accurately, training speed of neural network need to be improved and system computing overhead need to be reduced. In the following subsections, mass spectrum and its characteristics are explained, then the proposed systems using the differential operation, the ReliefF algorithm and the LSTM are presented.

A. MASS SPECTRUM

The quality of mass spectrum obtained on a given instrument is highly dependent on the purity of the mass spectrum, and the condition of the mass spectrometer. The relationships between mass spectrum and resolution, the presence of isotopes, and the fragmentation of molecules and molecular ions are crucial in understanding a mass spectrum, which will be discussed in the following subsections.

The mass spectrum has different resolutions. Low-resolution mass spectrometry refers to instruments which are capable of distinguishing among ions of different nominal masses [5]. High-resolution mass spectrometry calculates the precise mass of each compound that can aid distinguishing nominal mass. For example, compounds with the molecular formulas $C_3\ H_6\ O$ and $C_3\ H_8\ O$ have nominal masses of 58 and 60, respectively, and can be distinguished by low-resolution mass spectrometry. However, the compounds $C_3\ H_8\ O$ and $C_2\ H_4\ O_2$ have the same nominal mass of 60 and cannot be distinguished by a low-resolution mass spectrometry, where high-resolution mass spectrometry is needed. The presence of isotopes also has an effect on identifying the compound type. Fig. 1 is the mass spectrum of dopamine ($C_8\ H_{11}NO_2$), where the molecular ion nominal masses appears at m/z 153. But there is a small peak at m/z 154, which is from an ion 1 atomic mass units (amu) that is heavier than the molecular ion of dopamine, and corresponds to the presence in the ion of a single heavier isotope of H, C, N, or O in dopamine.

It is common to use electrons with energies of 70 eV (approximately 6750 KJ/mol) for electron ionization. This energy is sufficient to dislodge one or more electrons from a molecule, and to cause extensive fragmentation [5]. These fragments may be unstable and, in turn break apart into even smaller fragments. The molecular ions of some compounds have a sufficiently long lifetime in the analyzing chamber. They are observed in the mass spectrum, sometimes as the base (most intense) peaks. Molecular ions of other compounds have a shorter lifetime and present in low abundance or not at all. As a result, the mass spectrum of a compound ionized consists of a peak for the molecular ion and a series of peaks for fragment ions. In mass spectrum, the peak resulting from the most abundant cation is defined as the base peak and it is assigned an arbitrary intensity of 100. The relative abundances of all other cations in a mass spectrum are reported as percentages of the base peak.

Based on the aforementioned three aspects, for a realistic mass spectrum, the resolution of the mass spectrometer for measurement can be high or low. Both the number of isotopes contained in the element of the measured substance, and the distribution of the fragmentation have effects for the analysis. Fig. 2 is a real mass spectrometry data of shower gel, where the mass-to-charge ranges from 1 to 270 (i.e. a total of 270 features). There is a challenge of that how the optimal analysis results can be achieved while analyzing the similar mass spectrums, which will be addressed in the next subsection.

B. DIFFERENTIAL OPERATION

Differential operation can eliminate the noise interference and highlight the signal changes. In this approach, it is

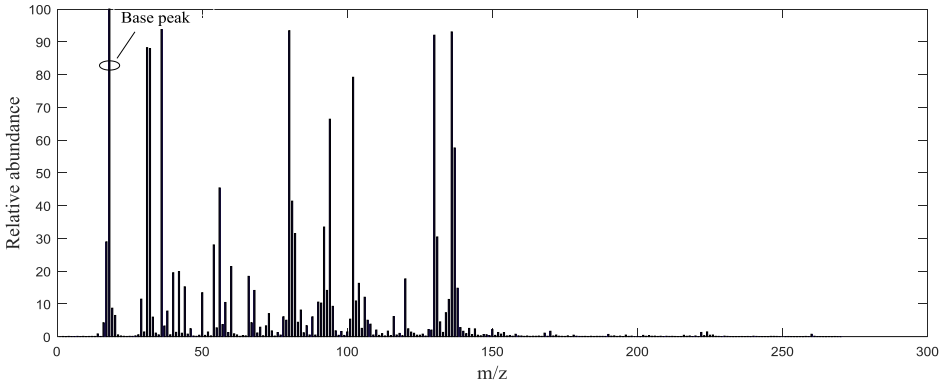


FIGURE 2. Mass spectrum of shower gel.

described by

$$x_d(n) = x(n) - x(n-1) \tag{1}$$

where $x_d(n)$ is the difference signal, $x(n)$ and $x(n-1)$ are the raw spectrum data at time step n and $n-1$, respectively.

When a mass spectrometer is used for detection, the substance can appear regularly or randomly. Once it is detected, the mass spectrometer produces a period of time-related mass spectrometry data. These data can be clearly observed after differential operation and the time point when the substance appears can be quickly located. Therefore, the differential operation is the first data processing step in this approach. In addition, the raw mass spectrometry data has a significant magnitude difference which is not conducive to neural network training. However the post-processing data is more uniform, which is beneficial to train the neural networks. By using the differential operation, the data that is used to input to the neural network becomes more concise and less interference.

C. FEATURE SELECTION OF THE MASS SPECTRUM

The raw mass spectrum has 270 features. Through the ReliefF algorithm, the raw mass spectrum was eventually reduced to 50 features. The specific operational process is as follows. Suppose there are K category tags in a given single dataset, the training dataset is defined as $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in R^p$ and $y_i \in R^k$ represent feature set and class label space of classified samples, respectively. If the sample x_i belongs to class k, then $y_i(k) = 1$, otherwise $y_i(k) = 0$. Thus the $p \times n$ feature matrix $X = [x_1, x_2, \ldots, x_n]$ and the $k \times n$ label matrix $Y = [y_1, y_2, \ldots, y_n]$ constitute the classification sample D. According to the ReliefF algorithm, the features of the raw mass spectrum are rearranged. The ReliefF algorithm chooses an instance R_i randomly and seeks for k of its nearest hits H_j and nearest misses $M_j(C)$ respectively. The basic idea of ReliefF is to assign the weights to each feature in the feature set of the classification samples, and then iterate to update the weights. Secondly, the feature subsets are selected according to the weight of feature, which makes the good features discrete different samples and aggregates similar samples. Finally, according to the features rearranged by the ReliefF algorithm, the top 50 most weighted features are selected as the final input of the neural network.

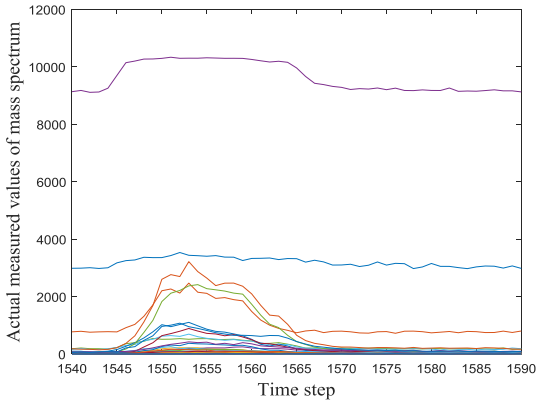


FIGURE 3. Mass spectrum channel selection by using the ReliefF algorithm for the DSTL spectrum dataset.

For the DSTL spectrum dataset, by using the ReliefF algorithm, an example of processed dataset is shown by Fig. 3. The x-axis corresponds to the time step, and the y-axis is the actual measured values of mass spectrum. The details of DSTL spectrum dataset will be provided in the Section IV. The processed dataset is ultimately used as input to the LSTM-based substance classification system. The ReliefF algorithm is used in this approach because not all mass spectrometer channels are used. It can find the channels that have high correlations with the output result. Mass spectrometer channels with weak correlations (e.g. isotope-induced) can be removed. This can also reduce the computing overhead of the classification systems.

D. RECURRENT NEURAL NETWORK

One problem of the RNN is the vanishing (or exploding) gradient problem. In this section, the architecture of RNNs and the gradient vanishing problem are briefly discussed. Then the LSTM that can address this problem is introduced.

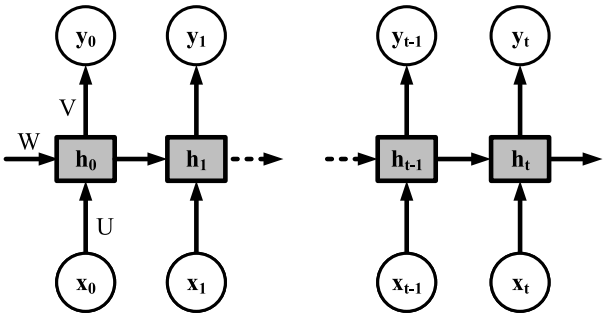


FIGURE 4. The architecture of RNNs [16].

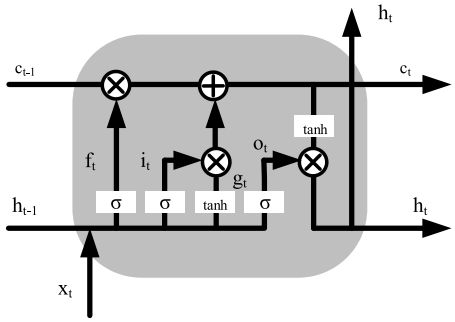


FIGURE 5. The architecture of LSTM [20].

1) RECURRENT NEURAL NETWORKS

The RNN is the extension of the conventional FFNNs in the time scale. Assuming that an input sequence, the hidden state sequence, and output vector sequence denoted by x, h and y, respectively. RNNs combine the input vector with the previous state vector to produce a new state vector. Hidden state h_t is described as

$$h_t = f(Ux_t + Wh_{t-1} + b_1) \tag{2}$$

where U, W are the weights for the connections from the input layer to the hidden layer, hidden layer to the hidden layer, respectively. Hidden state h_t equivalent to a memory container captures information which happened in all the previous time steps. Similar to the FFNNs, the output vector of RNN y_t is described as

$$y_t = g(Vh_t + b_2) \tag{3}$$

where V is the weights for the connections from the hidden layer to the output layer. In (2) and (3), the f and g are activation functions that squash the dot products to a specific range. The function f is usually tanh or ReLU. The g can also be a softmax. The b_1 and b_2 are biases that help offset the outputs from the origin. Fig. 4 shows a common topology of RNNs.

The conventional RNNs use Back Propagation Training Time (BPTT) to handle a variable-length sequence input [41]. There are two widely known issues during the training of RNNs, the vanishing and the exploding gradient problems [42], [43]. Unfortunately, the range of contextual information that standard RNNs can access is quite limited in practice [44]. Similar to RNN, the LSTM has recurrent connections so that the state from previous activations of the neuron in the previous time step is used as context for formulating an output. But unlike other RNNs, the LSTM has a unique formulation that allows to avoid two aforementioned problems.

2) LONG SHORT-TERM MEMORY

The LSTM is an architecture which was first proposed by Hochreiter and Schmidhuber [39] and refined by many other researchers. Fig. 5 shows a single LSTM cell.

For each time step t, x_t is the input to the memory cell layer, σ is the logistic sigmoid function, i_t, f_t and o_t are values of the input, forget and output gates, respectively. They are described by

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{4}$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{5}$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{6}$$

In LSTM, three gates control the information flow. The input gate decides which values will be updated. The forget gate defines how much of the previous state h_{t-1} are allowed to pass through, and the output gate defines how much of the internal state are exposed to the next layer. The candidate value g_t is computed by the current input x_t and the previous hidden state h_{t-1}. The key to LSTM is the cell state c_t, as the i_t, f_t and g_t interact with c_t. The g_t and c_t are described by

$$g_t = tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \tag{7}$$
$$c_t = f_t c_{t-1} + i_t g_t \tag{8}$$

where W_{xi}, W_{xf}, W_{xo} and W_{xg} are the weights for the connections. These weights propagate from the input data to the input gate, forget gate, output gate and candidate value. Similarly, this also applies to W_{hi}, W_{hf}, W_{ho}, W_{hg}, which are the connection weights from the hidden layer (at previous time-step) to the input gate, forget gate, output gate and candidate value (at the current time-step), and b_i, b_f, b_o and b_g are the corresponding bias.

Finally, the hidden state h_t at time t is computed by multiplying the $\tanh(c_t)$ with the output gate. This can be described by

$$h_t = o_t \tanh(c_t) \tag{9}$$

In the LSTM, the three gates (input, forget, output) are used to solve the vanishing and exploding gradient problems. In the conventional RNNs, the recurrent hidden layer is replaced by LSTM cell.

IV. EXPERIMENTAL RESULTS

The performance of the proposed LSTM-based substance classification systems are analyzed and discussed in this section. The experimental environment is the Anaconda platform using Python, i3-6100 CPU @ 3.70GHz, 12.0GB RAM. To demonstrate that the proposed classification systems can

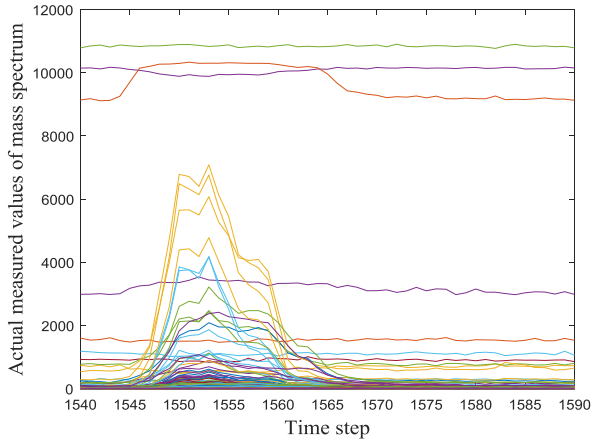


FIGURE 6. Mass spectrum data of brewed coffee.

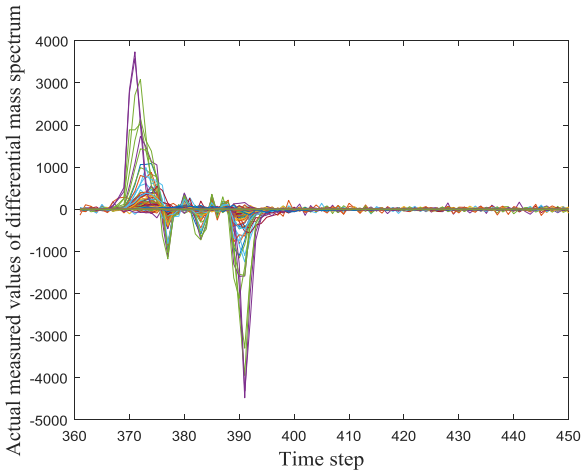


FIGURE 7. Rapid detection mechanism where the spectrum data is shown from time step 360 to 450.

be easily applied to spectrometry data, the DSTL spectrum dataset [2] is used in the experiment. It is a mass spectrometry dataset which was collected by using a highly sensitive time-of-flight proton-transfer mass spectrometer. At various intervals a number of different substances were introduced to the sensor which gradually change its properties over time at different distances and strengths. The strength was manually marked as weak, medium or strong. The mass spectrometry data within a time period represents a substance, and each substance has a unique profile. It has a total of 58,500 data samples where 20,000 samples are used for testing and the rest for training set, i.e. the ratio of testing and training dataset is about 7:3. It has various degrees of complexity and different levels of adulteration with anomalous substances. Fig. 6 shows the mass spectrometry data example of brewed coffee between time step 1,540 and 1,590. The color codes represent the features (i.e. channels) in mass spectrometer. The main advantages of this work include accurately detecting the presence of substances, reducing system computing resource overhead, and handling multiple time-related mass spectrometry data.

A. RAPID DETECTION MECHANISM

In order to classify the substances, the first step is to detect whether the substance is present at the mass spectrometer. To distinguish the substances from the background samples, a rapid detection mechanism is critical. In this work, the differential operation is used to detect whether the substance is present. If it is present, the spectrum data has an intense change which can indicate the starting time point of the substances. In this experiment, as the substance locations in DSTL spectrum dataset are randomly distributed, it becomes crucial to locate these substances. Selecting a suitable period to observe the data is a challenge. If the selected period is not appropriate, there is a risk that data integrity will be cut apart. However, the proposed rapid detection mechanism is suitable for either periodic or non-periodic distribution. Comparing the two examples in the DSTL spectrum dataset (time steps from 360 to 400 and from 400 to 440) in Fig. 7, it can be seen that the severe changes in amplitude corresponds to the presence of a substance, and the amplitude change is very small if no substance is present. By using the differential operations, when a substance is present, the rapid detection mechanism ensures that the location of the substance is clearly and accurately detected.

TABLE 1. Detection rates and training times of the LSTMs with different window sizes.

Window size [time step]	Detection rate [%]	Training time [%]
2	81.81	100*
5	78.78	74.03
10	81.81	67.8
20	81.81	69.61

*Training time of the LSTM with window size of 2 time steps is used as the baseline.

B. OPTIMAL WINDOW SIZE FOR THE LSTM-BASED SUBSTANCE CLASSIFICATION SYSTEMS

In the LSTM used in this approach, the recurrent connections add state or memory to the network and allow it to learn and harness the ordered nature of observations within input sequences. Due to recurrent connections, the states from previous activations of the neurons are used as context for formulating an output. The LSTM has internal states, and is explicitly aware of the temporal structures in the inputs and is able to model multiple parallel input series separately. It possesses memory which can overcome the issues of long-term temporal dependency with input sequences. Therefore, the number of memory cells (i.e. the optimal window size in this work) should be determined, as memory cells with different sizes produce different results. Table 1 shows the detection rates and training times of the LSTM-based substance classification systems with different window sizes. It can be seen that the lowest detection rate is 78.78% when the window size is 5 time steps, and the detection rates are same

TABLE 2. Different classification systems with and without ReliefF algorithm and differential operation.

Classification systems	ReliefF algorithm	Differential operation	Results
LSTM	No	No	Table 3
R-LSTM	Yes	No	Table 4
D-LSTM	No	Yes	Table 5
R-D-LSTM	Yes	Yes	Table 6

TABLE 3. Detection rates of the LSTM substance classification system.

Substance	Number of substances				Detection rate [%]
	Total	Matched	Missed	Misclassified	
Shower Gel	5	2	3	0	40
Shampoo	6	3	1	2	50
Shaving Gel	4	3	0	1	75
Coffee Beans	5	3	2	0	60
Brewed Coffee	4	2	0	2	50
Olive Oil	4	0	2	2	0
Smoked Ham	5	0	3	2	0
Overall	33	13	11	9	39.39

TABLE 4. Detection rates of the R-LSTM substance classification system.

Substance	Number of substances				Detection rate [%]
	Total	Matched	Missed	Misclassified	
Shower Gel	5	2	3	0	40
Shampoo	6	5	1	0	83.33
Shaving Gel	4	3	1	0	75
Coffee Beans	5	3	2	0	60
Brewed Coffee	4	0	2	2	0
Olive Oil	4	2	2	0	50
Smoked Ham	5	0	3	2	0
Overall	33	15	14	4	45.45

(81.81%) for the other window sizes. Table 1 also provides the results of training times, where the training time of the LSTM with the window size of 2 time steps is used as the baseline. For example if the window size is 5 time steps, the training time is shorter, i.e. 74.03% of the baseline. It can be seen that the LSTM with window size of 10 time steps achieves the lowest training time, 67.8% of the baseline. If the window size continues to increase, the training time increases again due to the intensive computing of large number of memory cells. Thus the window size of 10 time steps is optimal in this experiment due to the high detection rate and low training time.

C. DETECTION RATES OF THE LSTM-BASED SUBSTANCE CLASSIFICATION SYSTEMS

In this experiment, the detection rates of different classification systems are provided and these classification systems include the LSTMs with and without ReliefF algorithm and differential operation. Table 2 shows the four different classification systems and their performances under the optimal window size will be compared in this section.

1) LSTM SUBSTANCE CLASSIFICATION SYSTEM

In this experiment, the LSTM is used for the classifier and the raw DSTL spectrum dataset is used for the substance classifications. The results are shown by Table 3. The ReliefF algorithm and differential operation are not used in this experiment. It can be seen that only about one-third of the total substances can be detected. The numbers of matched (13) and missed (11) substances are almost the same. The number of missed and misclassified substances is relatively high. In total, 13 of the 33 substances are matched, which gives a detection rate of 39.39%.

2) R-LSTM SUBSTANCE CLASSIFICATION SYSTEM

The raw DSTL spectrum dataset is processed by ReliefF algorithm and then fed to the LSTM classifier in this experiment. The dataset dimension is reduced from 270 to 50 by using the ReliefF algorithm. Table 4 shows the result of the R-LSTM substance classification system. The proportion of matched (15) and missed (14) substances is still close to 1:1. However, the number of misclassified substances has been reduced to 4. It can be seen that the overall detection rate is 45.45%, which is better than the LSTM substance classification system in Table 3. This R-LSTM system has the advantages of high training speed and small network size which will be further discussed in the R-D-LSTM system.

3) D-LSTM SUBSTANCE CLASSIFICATION SYSTEM

In this experiment, differential operation is used only for the raw DSTL spectrum dataset and then the data is inputted to the LSTM. Table 5 shows the detection rates where a high

TABLE 5. Detection rates of the D-LSTM substance classification system.

Substance	Number of substances				Detection rate [%]
	Total	Matched	Missed	Misclassified	
Shower Gel	5	4	1	0	80
Shampoo	6	6	0	0	100
Shaving Gel	4	4	0	0	100
Coffee Beans	5	3	2	0	60
Brewed Coffee	4	4	0	0	100
Olive Oil	4	4	0	0	100
Smoked Ham	5	4	1	0	80
Overall	33	29	4	0	87.88

TABLE 6. Detection rates of the R-D-LSTM substance classification system.

Substance	Number of substances				Detection rate [%]
	Total	Matched	Missed	Misclassified	
Shower Gel	5	4	0	1	80
Shampoo	6	5	1	0	83.33
Shaving Gel	4	4	0	0	100
Coffee Beans	5	3	2	0	60
Brewed Coffee	4	4	0	0	100
Olive Oil	4	4	0	0	100
Smoked Ham	5	3	2	0	60
Overall	33	27	5	1	81.81

overall detection rate of 87.88% is obtained by this system. The D-LSTM effectively detects most instances of shampoo, shaving gel, brewed coffee and olive oil. It misses 1, 2 and 1 substances for the shower gel, coffee beans and smoked ham, respectively. In total, 4 of the 33 substances are missed, which gives a successful detection rate of 87.88% and no false-positive detection occurs.

4) R-D-LSTM SUBSTANCE CLASSIFICATION SYSTEM

In this experiment, the raw DSTL dataset is pre-processed by the ReliefF algorithm and differential operation. The R-D-LSTM system integrates the advantages of the R-LSTM and D-LSTM systems. The results are shown by Table 6. To improve computing efficiency, the ReliefF algorithm is used to select most significant features from the dataset and reduce the input dimensions for the neural network. The dimension of the raw dataset is significantly reduced from 270 to 50 giving a reduction rate of 81.48%. Table 6 shows detection rates of the R-D-LSTM system. Compare to the results of D-LSTM in Table 5, the detection accuracy of 81.81% is slightly lower due to that two substances are not matched. One of the two substances is missed for the smoked ham and the other is misclassified for the shower gel. However the advantage of the R-D-LSTM is that as the dimension of the raw dataset is significantly reduced, the training speed of the neural network is increased by 35% and the network size is reduced by 28.46%. This will be beneficial for the substance classification systems with limited computing resources such as the embedded hardware systems which are the typical platforms for the mass spectrometer.

By comparing the performances of these four different classification systems, the D-LSTM substance classification system has the best overall detection rate of 87.88%. The LSTM substance classification system has the lowest detection rate of 39.39% under the raw dataset. Compared to the LSTM system, R-LSTM improves the detection rate to 45.45% as the weak contribution features in the dataset have been removed. R-D-LSTM system combines the advantages of differential operation and the RelieF algorithm where the former is used to improve the detection rate, and the latter is used to reduce computing resource overhead. Although its overall detection rate of 81.81% is lower than the D-LSTM but the computing resource overhead is greatly reduced.

The DSTL spectrum dataset is also used in other approaches, such as the receptor density algorithm (RDA) in [45]. The RDA is inspired by T-cell signaling, and it has been used for the substance classifications of the DSTL spectrum dataset. It achieved 86.5% of detection rate and 3.2% of false-positive rate [2]. Comparing to the RDA, the D-LSTM and R-D-LSTM systems in this paper achieve the detection rates of 87.88% and 81.81%, false-positive rates of 0% and 3.57%, respectively, where the performance of D-LSTM is better than the RDA and the R-D-LSTM is lower. Based on the trade-off between the detection performance and the computing resource requirements, the D-LSTM and R-D-LSTM can be selected for different application domains.

V. CONCLUSION

In this work, LSTM based substance detection methods are proposed, which consist of two parts, i.e., the pre-processing of mass spectrometry data and classifications of the chemical substances. For the former, differential operation and ReliefF algorithm are used to improve the classification accuracies and reduce the feature dimensions and computing cost,

respectively. For the latter, LSTM based substance detection systems are designed, where the optimal parameters of the LSTM model are obtained through experiments. Results show that the D-LSTM substance classification system has the best overall detection rate and the R-D-LSTM system achieves the balance between high classification accuracy and low computing resource requirement by combining the advantages of differential operation and the RelieF algorithm. It is desirable for these detection systems to be implemented in embedded hardware systems of different real applications. Future work include further optimize the substance detection systems such as the neural network parameters, and reduce the requirements of the hardware computing resources.

REFERENCES

[1] F. Aubriet and V. Carré, "Potential of laser mass spectrometry for the analysis of environmental dust particles—A review," *Anal. Chim. Acta*, vol. 659, nos. 1–2, pp. 34–54, 2010.

[2] J. A. Hilder *et al.*, "Parameter optimisation in the receptor density algorithm," in *Proc. Int. Conf. Artif. Immune Syst.*, vol. 6825, 2011, pp. 226–239.

[3] F. Y. Edgeworth, "On discordant observations," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 23, no. 143, pp. 364–375, 1887.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, 2009, Art. no. 15.

[5] W. H. Brown, B. Iverson, E. V. Anslyn, and C. Foote, *Organic Chemistry*, 8th ed. Boston, MA, USA: Cengage Learning, 2017.

[6] C. S. Tong and K. C. Cheng, "Mass spectral search method using the neural network approach," *Chemometrics Intell. Lab. Syst.*, vol. 49, no. 2, pp. 135–150, 1999.

[7] B. Curry and D. E. Rumelhart, "MSnet: A neural network which classifies mass spectra," *Tetrahedron Comput. Methodol.*, vol. 3, nos. 3–4, pp. 213–237, 1990.

[8] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Appl. Intell.*, vol. 7, no. 1, pp. 39–55, 1997.

[9] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. 10th Nat. Conf. Artif. Intell.*, 1992, pp. 129–134.

[10] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artif. Intell.*, vol. 176, no. 1, pp. 2291–2320, 2012.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[12] D. A. Cirovic, "Feed-forward artificial neural networks: Applications to spectroscopy," *TrAC Trends Anal. Chem.*, vol. 16, no. 3, pp. 148–155, 1997.

[13] H. M. Azamathulla, M. C. Deo, and P. B. Deolalikar, "Alternative neural networks to estimate the scour below spillways," *Adv. Eng. Softw.*, vol. 39, no. 8, pp. 689–698, 2008.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 3104–3112.

[15] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *Int. J. Forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[16] M. Hüsken and P. Stagge, "Recurrent neural networks for time series classification," *Neurocomputing*, vol. 50, pp. 223–235, Jan. 2003.

[17] M. Dixon, "Sequence classification of the limit order book using recurrent neural networks," *J. Comput. Sci.*, vol. 24, pp. 277–286, Jan. 2018.

[18] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[19] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

[20] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[21] Z. C. Lipton, J. Berkowitz, and C. Elkan. (2015). "A critical review of recurrent neural networks for sequence learning." [Online]. Available: https://arxiv.org/abs/1506.00019

[22] T. G. Dietterich, "Machine learning for sequential data: A review," in *Structural, Syntactic, and Statistical Pattern Recognition*. Berlin, Germany: Springer, 2002, pp. 15–30.

[23] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823–839, May 2012.

[24] J. Zupan and J. Gasteiger, "Neural networks: A new method for solving chemical problems or just a passing phase?" *Anal. Chim. Acta*, vol. 248, no. 1, pp. 1–30, 1991.

[25] A. Sagheer and M. Kotb, "Time series forecasting of petroleum production using deep LSTM recurrent networks," *Neurocomputing*, vol. 323, no. 5, pp. 203–213, 2019.

[26] Y. Kalegowda and S. L. Harmer, "Classification of time-of-flight secondary ion mass spectrometry spectra from complex Cu–Fe sulphides by principal component analysis and artificial neural networks," *Anal. Chim. Acta*, vol. 759, pp. 21–27, Jan. 2013.

[27] A. Eghbaldar, T. P. Forrest, and D. Cabrol-Bass, "Development of neural networks for identification of structural features from mass spectral data," *Anal. Chim. Acta*, vol. 359, no. 3, pp. 283–301, 1998.

[28] S. Bell, E. Nazarov, Y. F. Wang, and G. A. Eiceman, "Classification of ion mobility spectra by functional groups using neural networks," *Anal. Chim. Acta*, vol. 394, nos. 2–3, pp. 121–133, 1999.

[29] Z. Boger, "Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis," *Anal. Chim. Acta*, vol. 490, nos. 1–2, pp. 31–40, 2003.

[30] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," *Large Scale Kernel Mach.*, vol. 34, pp. 321–360, 2007.

[31] Z.-S. Zhang, L.-L. Cao, J. Zhang, P. Chen, and C.-H. Zheng, "Prediction of molecular substructure using mass spectral data based on deep learning," in *Intelligent Computing Theories and Methodologies*, vol. 9226. Cham, Switzerland: Springer, 2015, pp. 520–529.

[32] G. Dorffner, "Neural networks for time series processing," *Neural Netw. World*, vol. 6, no. 1, pp. 447–468, 1996.

[33] B. Abraham and G. E. P. Box, "Bayesian analysis of some outlier problems in time series," *Biometrika*, vol. 66, no. 2, pp. 229–236, 1979.

[34] A. M. Bianco, M. G. Ben, E. J. Martínez, and V. J. Yohai, "Outlier detection in regression models with ARIMA errors using robust estimates," *J. Forecasting*, vol. 20, no. 8, pp. 565–579, 2001.

[35] B. Abraham and A. Chuang, "Outlier detection and time series modeling," *Technometrics*, vol. 31, no. 2, pp. 241–248, May 1989.

[36] H. Sak, A. Senior, and F. Beaufays. (2014). "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition." [Online]. Available: https://arxiv.org/abs/1402.1128

[37] F. A. Gers, D. Eck, and J. Schmidhuber, "Applying LSTM to time series predictable through time-window approaches," in *Proc. Int. Conf. Artif. Neural Netw.*, 2001, pp. 669–676.

[38] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2015, pp. 89–94.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] M. Esteban, C. Ariño, and J. M. Díaz-Cruz, "Chemometrics for the analysis of voltammetric data," *TrAC Trends Anal. Chem.*, vol. 25, no. 1, pp. 86–92, 2006.

[41] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

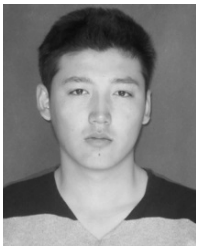
[42] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[43] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[44] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[45] N. D. L. Owens, A. Greensted, J. Timmis, and A. Tyrrell, "The receptor density algorithm," *Theor. Comput. Sci.*, vol. 481, pp. 51–73, Apr. 2013.

JUNXIU LIU, photograph and biography not available at the time of publication.

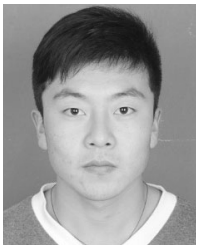


JINLEI ZHANG received the B.E. degree from the Harbin University of Science and Technology, China, in 2013. He is currently pursuing the master's degree with the Faculty of Electronic Engineering, Guangxi Normal University. His research interests include data analytics, and long short-term memory networks and its application.

YULING LUO received the Ph.D. degree in information and communication engineering from the South China University of Technology, Guangzhou, China. She is currently an Associate Professor with the Faculty of Electronic Engineering, Guangxi Normal University, Guilin, China. Her research interests include information security, image processing, chaos theory, and embedded system implementations.

SU YANG received the B.A. degree in mechanical engineering from the Changchun University of Technology, Changchun, China, in 2008, the M.Sc. degree in information technology from the University of Abertay Dundee, Dundee, U.K., in 2010, and the Ph.D. degree in electronic engineering from the University of Kent, Canterbury, U.K., in 2015. He was with the Intelligent Interactions Research Group, School of Engineering and Digital Arts, where his research was focused on using EEG for biometric person recognition. He was a Postdoctoral Research Associate with the College of Engineering, Temple University, Philadelphia, PA, USA, from 2016 to 2017. He is currently a Senior Research Associate with the Intelligent Systems Research Centre, Ulster University, Londonderry, U.K. His current research interests include signal processing, pattern recognition, EEG-event detection, and MEG source reconstruction/localization.

JINLING WANG received the M.Sc. degree (Hons.) in computing and information systems from the School of Computing and Intelligent Systems, University of Ulster, in 2003, and the Ph.D. degree from the Intelligent Systems Research Centre, University of Ulster, in 2016. She is currently a Research Associate of computing science with the University of Ulster. Her major research interests include intelligent data analysis, bio-inspired adaptive systems, spiking neural networks, artificial neural networks, online learning, machine learning, data mining, and pattern recognition for a wide range of datasets. She serves as a Reviewer for several international conferences and journals.



QIANG FU is currently pursuing the Ph.D. degree with the Computer Science and Technology College, Harbin Engineering University, China. His major research interests include intelligent information processing and computational intelligence.