Research in Astronomy and Astrophysics, 23:035003 (12pp), 2023 March https://doi.org/10.1088/1674-4527/acafc5

Investigation of Traffic Classification Applied to an Astronomical Data Transmission Network of the XAO Using Deep Learning

Jie Wang[1,2], Hai-Long Zhang[1,2,3], Na Wang[1,3], Xin-Chen Ye[1,2], Wan-Qiong Wang[1], Jia Li[1], Meng Zhang[1,4], Ya-Zhou Zhang[1,4], and Xu Du[1,4]
[1] Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi 830011, China; wangjie@xao.ac.cn, zhanghailong@xao.ac.cn
[2] National Astronomical Data Center, Beijing 100101, China
[3] Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Nanjing 210008, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China
Received 2022 August 19; revised 2022 December 13; accepted 2022 December 30; published 2023 February 10

Abstract

A telecommunication network used for the transmission of astronomical observation data, telescope remote control and other astronomical research purposes is a critical infrastructure. The monitoring and analysis of network traffic, which help improve the network performance and the utilization of network resources, are a challenging task. The accurate identification of the astronomical data traffic will effectively improve transmission efficiency. In this paper, a classification method applied to types of traffic containing astronomical data using deep learning is proposed. The advantages of a convolutional neural network model in image classification are exploited to classify types of traffic containing astronomical data. The objective is to identify the mixed traffic in the network and accurately identify types of traffic containing astronomical data. The effectiveness of the model in improving classification accuracy is also discussed. Actual traffic data captured by Tcpdump and Wireshark are tested, and the experimental results indicate that the proposed method can accurately classify types of traffic containing astronomical data.

Key words: astronomical databases: miscellaneous – virtual observatory tools – surveys

1. Introduction

A transmission network is a critical infrastructure in astronomical data transmission, data archiving, publishing and network research. The Xinjiang Astronomical Observatory (XAO) relies on the China Science and Technology Network to ensure a stable Gigabit-bandwidth Zhang et al. (2019). This network employs a data transmission fiber line ($b = 300\,\mathrm{Mb\,s^{-1}}$) between Nanshan Station and XAO headquarters ($\Delta D = 100\,\mathrm{km}$). It serves the transmission of data obtained from observation equipment, such as the Nanshan 26 m radio telescope (NSRT) and the Nanshan 1 m wide-field optical telescope (NOWT). It also provides disaster backup of the observation data using the data storage system of Nanshan Station and XAO Zhang et al. (2022). To solve the data transmission problems encountered in site monitoring and optical telescope observation equipment, a data transmission fiber line ($b = 100\,\mathrm{Mb\,s^{-1}}$) has been established between the Qitai Station and XAO headquarters ($\Delta D = 200\,\mathrm{km}$).

The astronomical data transmission traffic in the XAO data transmission network can be roughly divided into the following two types Wang et al. (2022): (1) Bandwidth-sensitive traffic generated by batch transmission and astronomical data storage. (2) Delay-sensitive traffic generated by high-performance computing, data retrieval and other processes. Due to the rapid growth of high throughput demand in the bandwidth-sensitive traffic and low-latency demand in the delay-sensitive traffic, as well as the emergence of new Internet applications and the interaction among various terminals, the complexity and diversity of transmission networks have considerably increased, making the classification of types of traffic containing astronomical data a complex problem. Therefore, it is necessary to use network resources reasonably and effectively to identify different types of data transmission traffic.

The main contributions of this paper are as follows:

(i) By comparing the four main classification methods, we propose a traffic classification method that employs deep learning. This method can be applied to the data communication network of the XAO. As far as we know, this is the first study on the classification of types of traffic containing astronomical data.

(ii) In the absence of public astronomical traffic data sets, and no studies on the classification of traffic containing astronomical data, we capture real traffic data transmitted in an astronomical network by Tcpdump and Wireshark using the port mirroring mode in the core switch, and private astronomical transmission traffic data sets are constructed for conducting experiments.

Table 1
Comparison of Network Traffic Classification Methods

Methods	Classification Basis	Advantages	Disadvantages
Port matching	port number	fast detection time, high accuracy, low complexity, strong practicability	poor generality, high error rate
Deep packet inspection	payload of packets	high recognition accuracy, considering the port randomness	requires packet capturing, and consumes resources
Statistics characteristics	statistical features	does not need to obtain time or size for individual packets	difficult to select features a false alarm can be caused
Machine learning	feature selection	does not depend on computational resources, has few model parameters	depends on correct training data and network flow features

(iii) The four main traffic classification methods require manual operation, especially the method based on machine learning, which first requires manual extraction of network-traffic-related features by experts. Therefore, we propose a method that does not require manual extraction of network traffic-related features, and apply the method to astronomical data transmission. As a result, the cumbersome steps of feature search and extraction are avoided, and the data processing efficiency is improved.

The rest of this paper is organized as follows. In Section 2, some basic theoretical knowledge related to network traffic classification and deep learning is introduced. The proposed classification method of data transmission traffics, including data acquisition, data pre-processing and model training, which is the core of this paper, is presented in detail in Section 3. In Section 4, the hardware and software experimental parameters and the selection of model evaluation indices are described. Also, the experimental results are further discussed to investigate the effectiveness of the model in improving the accuracy of data transmission traffic classification. Finally, Section 5 concludes the paper.

2. Theoretical Background

2.1. Network Traffic Classification

Due to the continuous development of network applications, the identification of specific network traffic or applications has become very important in network control and management. Current network traffic classification methods can be classified into the following four types according to the different technologies used: (i) The method based on port matching Wang et al. (2015); this method has the advantages of fast detection time, high accuracy, low complexity and strong practicability in a traditional network environment, it is known to be among the fastest and simplest methods for classification of network traffic. However, it also exhibits some limitations due to its poor generality and high error rate Rezaei & Liu (2018). Madhukar & Williamson (2006) proposed only 30%–70% of the current Internet traffic can be classified using this method. Thus, this method is not used because of the low accuracy of traffic classification. (ii) The method based on deep packet inspection Guo et al. (2017); this method has the advantage of high recognition accuracy. It can recognize a certain number of protocols by considering the port randomness. However, the extraction of traffic characteristics requires packet capturing and consumes resources El-Maghraby et al. (2017). Hence, this method is not used because it is time-consuming and has limited traffic classification capability. (iii) The method based on traffic-statistics characteristics Zhang et al. (2012); this method does not need to obtain time or size information for individual packets to identify specific applications. However, statistical features are difficult to select, a false alarm can be caused, and the performance of the method is relatively poor in real time Yang (2019). Accordingly, this method is not used because it takes a lot of time to select statistical features. (iv) The method based on machine learning Shafiq et al. (2016); this method does not depend on computational resources, is not greedy for training samples and has few and well-defined model parameters. However, it depends on large-scale correct training data and network flow features Akinsola (2017). Consequently, this method is not used because the feature extraction and feature selection phases are essentially done with the assistance of the expert, and it is time-consuming and prone to human mistakes. A comparison of the advantages and disadvantages of the above four methods is shown in Table 1. In summary, we will use deep learning methods applied to astronomical data traffic classification for reducing time-consuming and manual feature selection.

2.2. Deep Learning

Deep learning is a branch of machine learning. Compared with traditional machine learning, deep learning does not require experts to spend a large amount of time on feature selection, thus, avoiding the subjectivity and incompleteness caused by human feature selection. In addition, the deep learning ability in large data sets far exceeds that of traditional machine learning. The convolutional neural network (CNN) used in this paper is a feedforward neural network involving convolutional computation and has a deep structure. It is one of the representative algorithms in deep learning. The classic

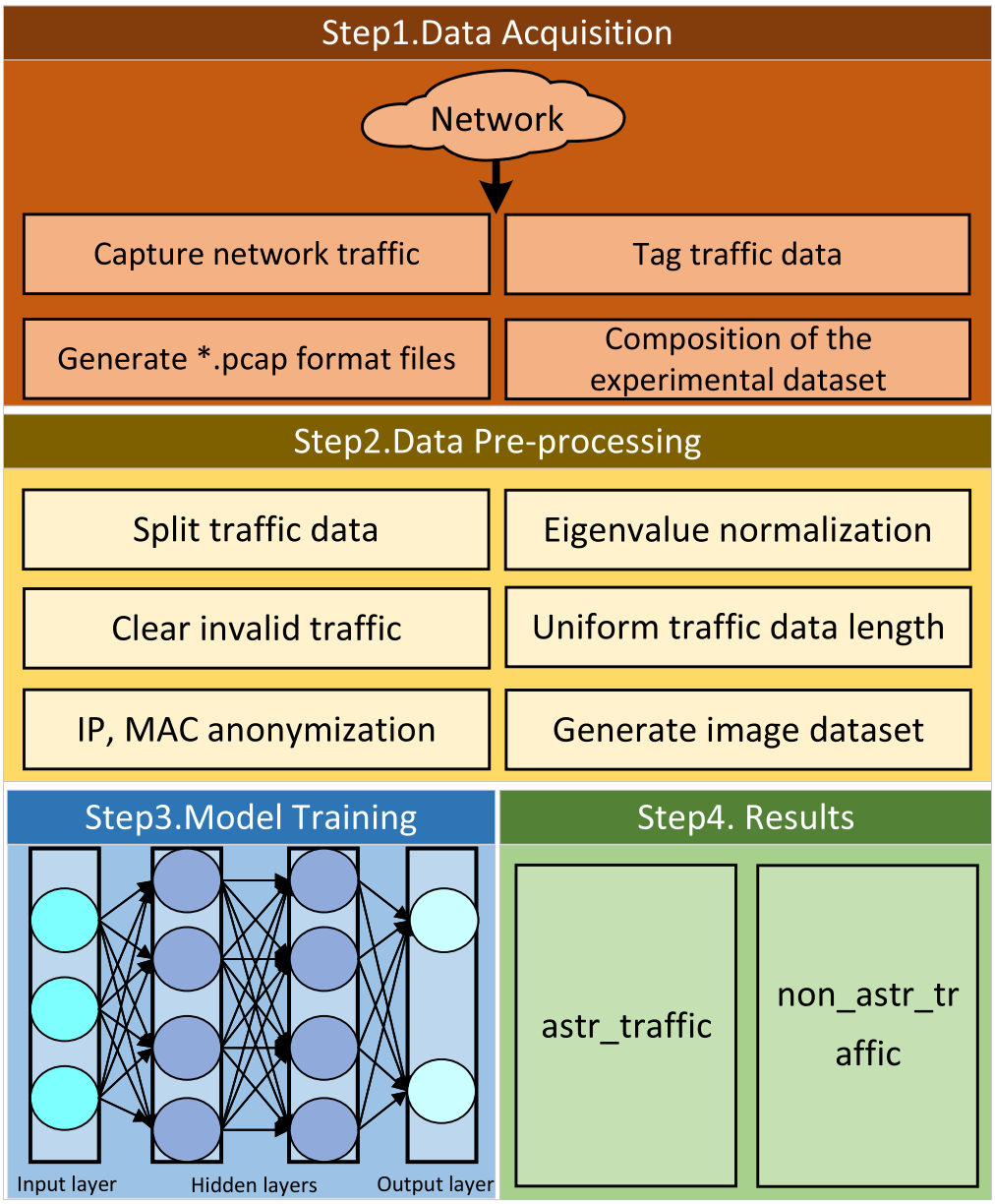

Figure 1. Flow chart of the proposed classification method for types of traffic containing astronomical data.

CNN model includes the following four types: (i) LeNet-5 Lecun et al. (1998); this model has the following advantages: (a) feature extraction and pattern classification are performed and generated simultaneously during training; (b) weight sharing reduces the network training parameters; (c) its structure is simpler than that of other network types. (ii) AlexNet Krizhevsky et al. (2012); this model adds a rectified linear unit (RELU) nonlinear activation function to enhance the nonlinear expression of the model and a dropout layer to prevent overfitting. (iii) ResNet He et al. (2016); this model introduces a batch normalization layer to increase the training speed and stability of the network during convergence as well as to increase the depth of the network in order to improve the feature extraction ability of the model. (iv) Xception Chollet (2017); this model achieves full cross-channel correlation decoupling, and spatial correlation can make full use of hardware resources. One-dimensional-CNN (1D-CNN) is commonly used in sequence models or natural language processing and has achieved some results. We use 1D-CNN to construct classification model for traffic containing astronomical data, because a traffic packet is a sequence of data bytes which is similar to a language sequence to some extent.

Table 2
Traffic Data in the XAO Data Transmission Network

Type	Wave Band	Date	Time	Volume
Pulsar data-I	Radio	2022-05-18	10:06:54	1.32 GB
Pulsar data-II	Radio	2022-07-16	19:10:38	1.58 GB
VLBI data	Radio	2022-05-26	12:30:50	3.78 GB
		2022-05-26	17:34:36	
Optical Data-I	Optical	2022-06-25	09:04:01	1.17 GB
Optical Data-II	Optical	2022-06-17	16:32:08	1.03 GB
Optical Data-III	Optical	2022-05-31	19:07:24	2.28 GB
		2022-06-17	16:11:29	

Two-dimensional-CNN (2D-CNN) has been applied in the computer vision and image processing. We also use 2D-CNN to construct a classification model for traffic containing astronomical data as a comparison experiment, because it is possible to convert the traffic packet into a two-dimensional image.

3. Classification Method for Types of Traffic Containing Astronomical Data

The proposed classification method for types of traffic containing astronomical data includes the following three steps: (i) Data acquisition; traffic data are collected using port mirroring in the XAO core switch by employing the network traffic capture software to form the experimental data set. (ii) Data pre-processing; the collected traffic data are segmented, any invalid or duplicate traffic is removed, the IP is anonymized, and grayscale images are generated according to the CNN input. (iii) Model training; the generated images are inserted into the CNN model for training, and the classification results are finally obtained so that the data transmission traffic and non-data transmission traffic can be effectively distinguished. The flow chart of the proposed method is shown in Figure 1.

3.1. Data Acquisition

Tcpdump and Wireshark were used to capture the traffic data in the XAO data transmission network (between 2022 May and July). The captured data were stored in files (*.pcap format). Pcap is a commonly used packet storage file starting with a 24 byte pcap header, followed by the information related to each message (a 16 byte packet header records the message information, and the packet data record the message data). The captured packets were separated into different pcap files and marked according to the packets generated by the application (e.g., scp, rsync) and the type of astronomical data (e.g., pulsar data, very long baseline interferometry data, optical data) transmitted by the application during the traffic capture (e.g., "scp_vlbi.pcap," "rsync_pulsar.pcap"). The specific data set information is shown in Table 2. The total number of raw stream entries is 9.12×10^6, and the total data size is 11.16 GB. The specific packet length empirical probability mass function is shown in Figure 2.

3.2. Data Pre-processing

Network traffic data must be pre-processed to generate images that match the model input before being fed into the CNN model for training. Pre-processing mainly includes traffic data segmentation, invalid or duplicate traffic cleaning, IP address and MAC address anonymization, feature value normalization and unified traffic data length. Initially, the pcap file was segmented, as shown in Figure 3, into n files containing a global header, packet headers and packet data. Any traffic unrelated to data transmission traffic or invalid traffic (e.g., packet re-transmission, out-of-order packets, etc.) was removed.

On the Internet, the IP and MAC addresses can accept different categories of network traffic without affecting the classification results. However, the input to the neural network model not only affects the classification accuracy but also increases the model operation complexity and the training time. The IP and MAC addresses were anonymized during data pre-processing. To facilitate the model data, the traffic data were normalized, as shown in Equation (1)

$$\mathrm{NORM}[x] = \frac{x - \mathrm{MIN}[x]}{\mathrm{MAX}[x] - \mathrm{MIN}[x]} \tag{1}$$

where NORM[x] is the normalized eigenvalue of the flow data, x is the original flow data, MAX[x] is the maximum value in the original flow data, and MIN[x] is the minimum value in the original flow data. The model input is kept in the interval of [0, 1] by applying a linear transformation to the original flow data.

3.3. Network Model

CNN is one of the representative algorithms of deep learning, which belongs to deep feed-forward neural network, including convolutional layer and pooling layer. Based on the structural characteristics of traffic packets, and the advantages of CNN hidden layer with shared convolutional kernel parameters and sparsity of inter-layer connections, reducing the manual judgment of features, this paper uses CNN model to process the original packet data.

As shown in Figure 4, the input layer corresponds to the original traffic data, and the first 784 bytes are obtained to generate a grayscale map of 28×28 pixels. The core of a CNN is the convolution and pooling operation, where the convolution operation is based on convolutional kernels to extract pixel features, as shown in Equation (2). A single convolutional kernel extracts local features, and multiple convolutional kernels can extract data from multiple angles to form a

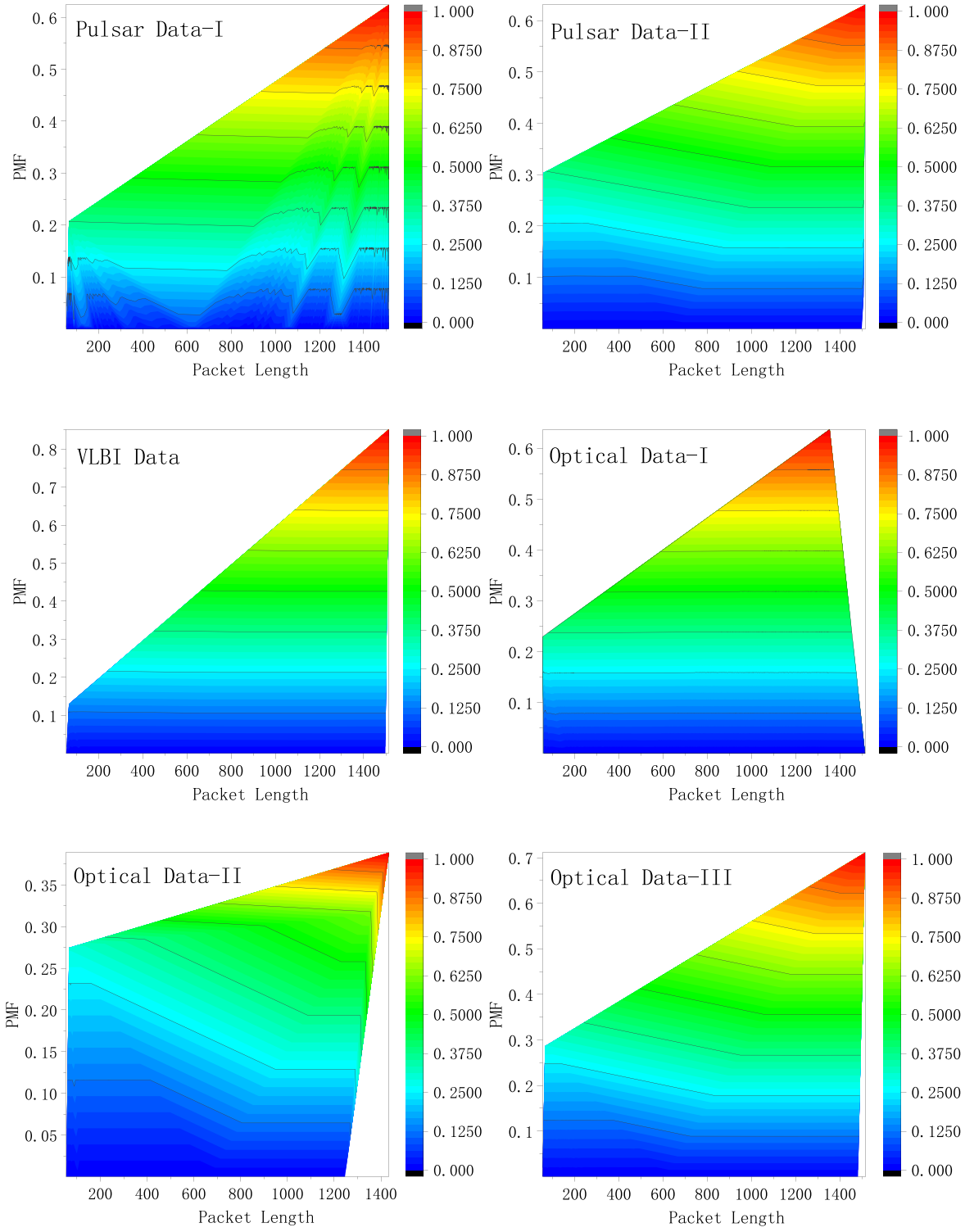

Figure 2. Visualization of the empirical probability mass function of packet length in astronomical data transmission traffic data sets.

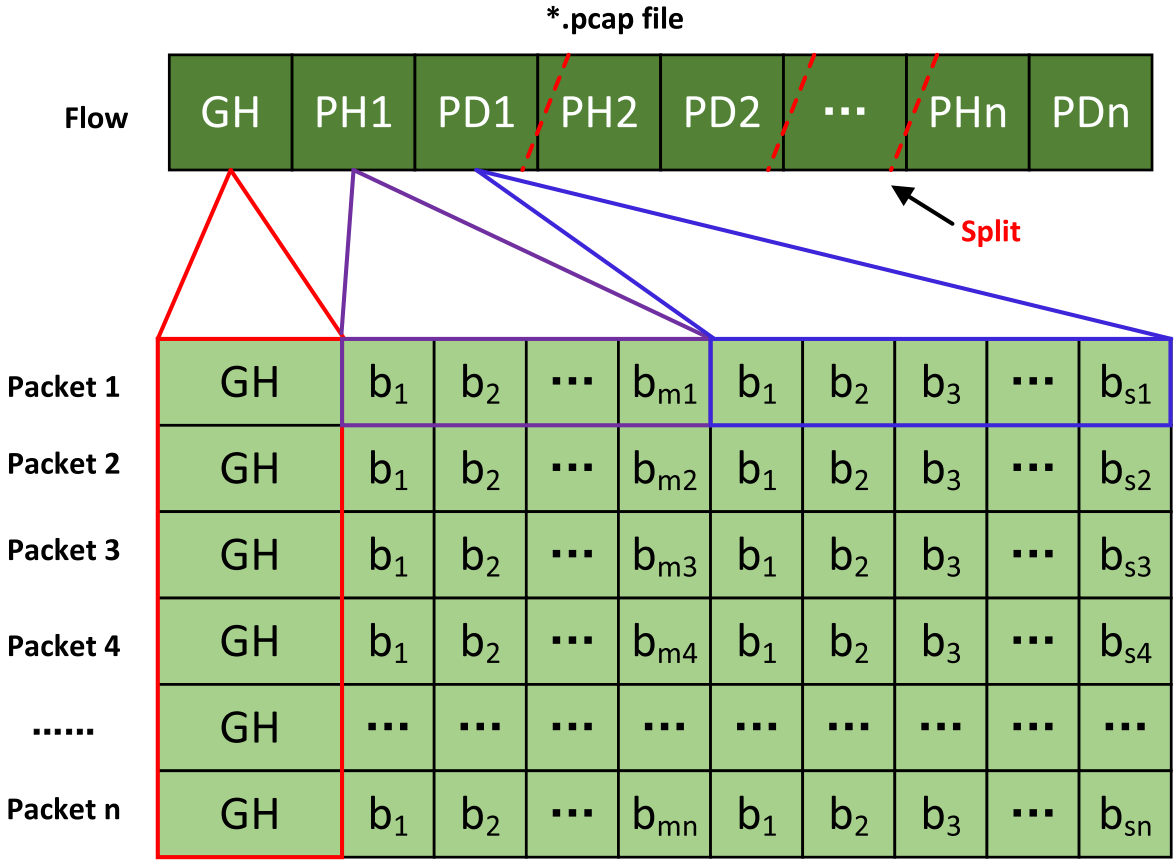

Figure 3. Network traffic data splitting. (Note: GH denotes the global header, PH denotes the packet header, PD denotes the packet data, bm and bs denote the m byte in the packet header and the s byte in the packet data, respectively, and n denotes the packet number.)

convolutional layer for extracting the overall features

$$y_{\text{conv}} = g\left(\sum_{a=0}^{m-1} \omega_a x_{i+a}^{(m-1)} + b_c^{(m)}\right), x \in \mathbb{R}^N \tag{2}$$

where m is the number of network layers in the CNN, i is the feature map, ω is the convolution kernel, b_c is the bias, RN is the set of feature vectors, and $g(\cdot)$ is the activation function RELU which injects nonlinear features into the model.

The pooling operation operates as a feature selection and information filtering as well as downsampling of the convolutional kernel output. The pooling layer converts point features into area features according to the predefined pooling function and further aggregates the features to reduce model overfitting. In this paper, maximum pooling is used to perform feature filtering in the convolutional layer output.

In the fully connected layer, the pooled feature vectors are integrated. Finally, multiple classification results are presented in the form of probabilities. The normalized exponential function softmax was used to obtain the classification results, as shown in Equation (3)

$$\text{softmax}(y_{\text{conv}}^{(m)}) = \frac{e^{y_{\text{conv}}^{(m)}}}{\sum_i e^{x_i^{(m)}}}. \tag{3}$$

4. Experimental Results

An Intel(R) Xeon(R) W-2145 CPU@3.70 GHz workstation with 64 GB RAM was to perform the experiments. python 3.7 was used as the programming language and Tensorflow 1.14.0 was used as the deep learning model framework.

4.1. Evaluation Index

The classification performance of the proposed traffic classification method applied to an astronomical data transmission network was validated using the commonly used classification metrics, i.e., accuracy, precision, recall and F1-score. Among them, the accuracy (i.e., the ratio of correctly classified traffic containing astronomical data to the entire network traffic in the experimental data set) was used to

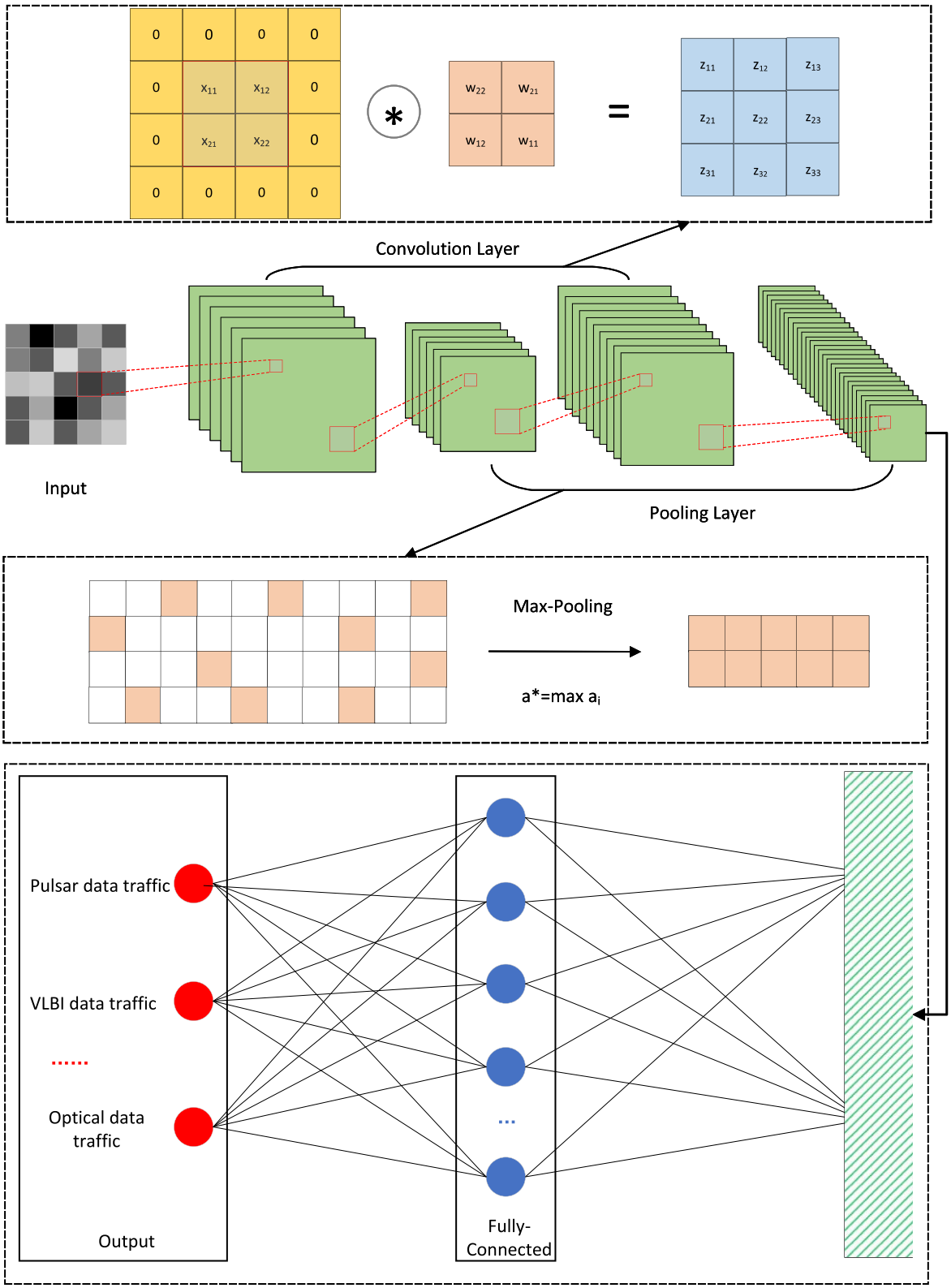

Figure 4. Convolutional neural network model.

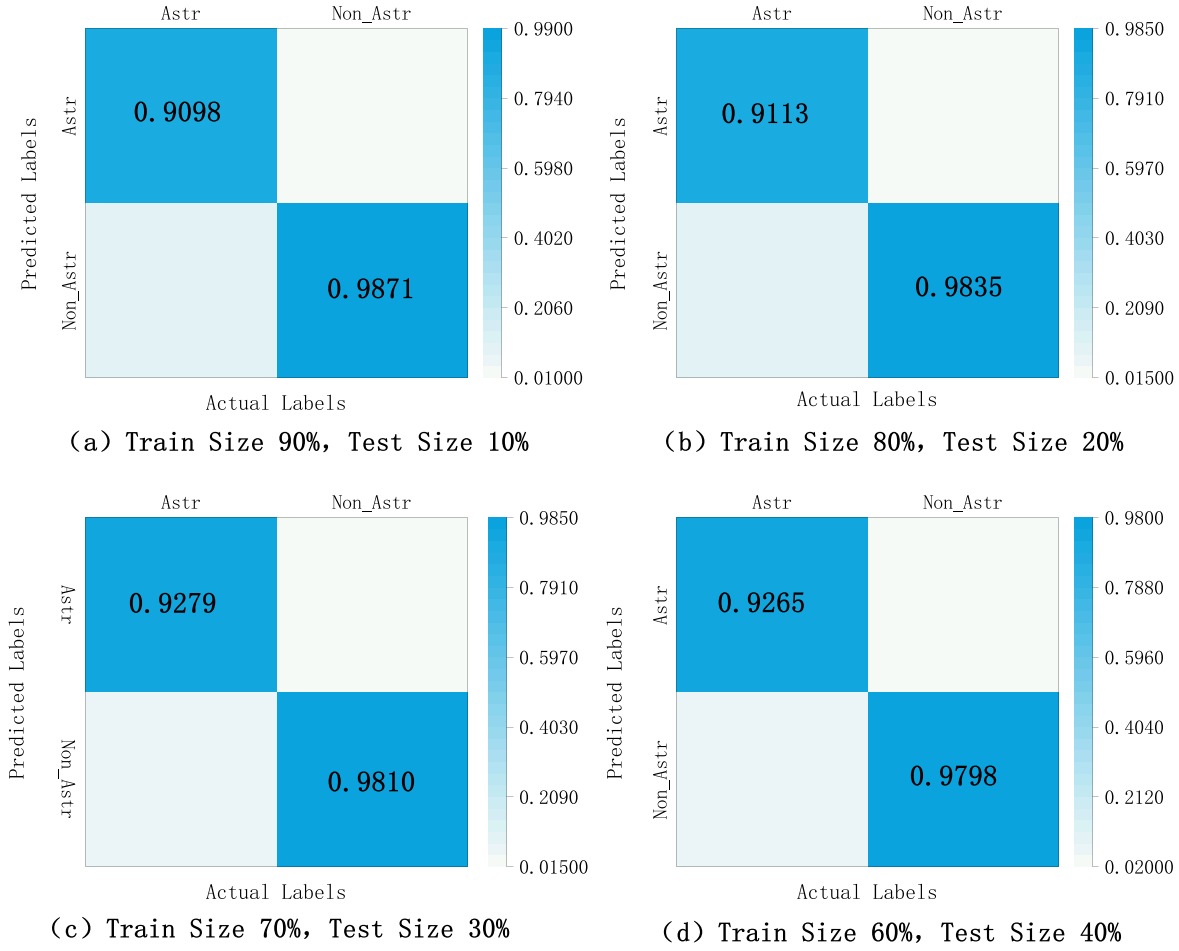

Figure 5. Two-class confusion matrix(1D-CNN).

evaluate the overall performance of the classification method, as expressed in Equation (4)

$$\text{Accuracy} = \frac{\sum T_i}{\sum (T_i + F_i)},\ i \in \text{Class}(\cdot). \tag{4}$$

The accuracy, recall and F1-score represent the ratio of correctly classified traffic containing astronomical data to the total number of the same type of traffic in the classification result, the ratio of correctly classified traffic containing astronomical data to the total number of the same type of traffic in the test set, and the overall classification evaluation of traffic containing astronomical data, respectively, as shown in Equations (5)–(7)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

where T_i is the correctly classified network traffic, F_i is the incorrectly classified network traffic, TP is the correctly classified astronomical traffic, FP is the incorrectly classified astronomical data transmission traffic, FN is the incorrectly classified non-astronomical data transmission traffic and i is the number of classifications.

4.2. Two-class Experimental Results

Figures 5 and 6 show the confusion matrix for the two-class test using 1D-CNN and 2D-CNN models, this is the result of random testing. It can be observed that, the correct classification of traffic containing astronomical data is above 90%, and the correct classification of the other types of traffic is above 97% in the remaining tests. There was no significant difference in the correct classification results of 1D-CNN and 2D-CNN models, which are −2.22%, 0.41%, 0.58%, 0.48% in the traffic

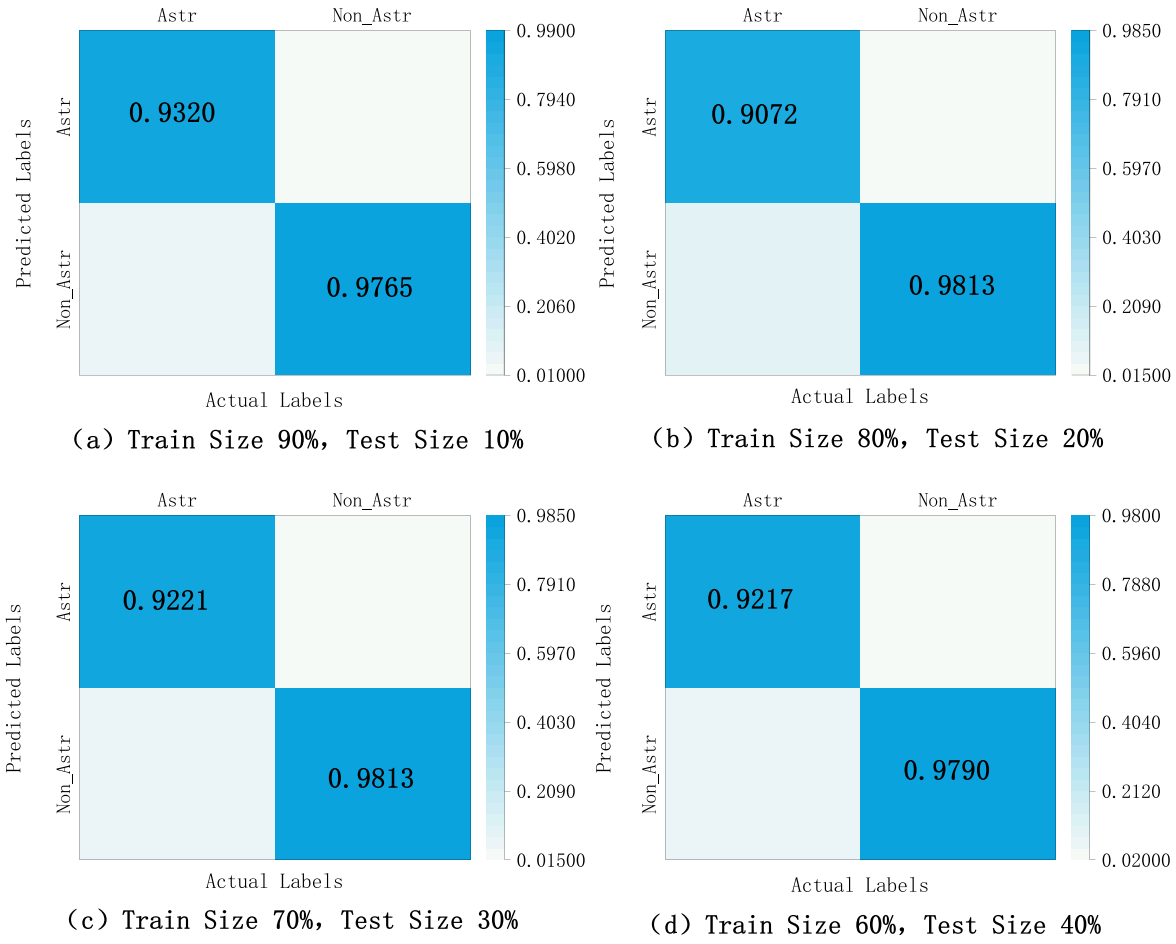

Figure 6. Two-class confusion matrix (2D-CNN).

Table 3
Results of Two-class Experiments (1D-CNN)

Data Set		Classes	Results			
Train Size	Test Size		Precision	Recall	F1-score	Total Accuracy
90%	10%	Astr	91.86% ± 1.43%	94.05% ± 1.07%	92.96% ± 0.54%	97.13% ± 0.22%
		Non_Astr	98.48% ± 0.27%	97.87% ± 0.43%	98.20% ± 0.14%	
80%	20%	Astr	91.68% ± 0.79%	91.53% ± 1.41%	91.59% ± 0.63%	96.97% ± 0.21%
		Non_Astr	98.14% ± 0.30%	98.18% ± 0.19%	98.15% ± 0.12%	
70%	30%	Astr	93.31% ± 0.68%	92.67% ± 1.01%	92.98% ± 0.29%	96.92% ± 0.12%
		Non_Astr	97.94% ± 0.27%	98.13% ± 0.22%	98.03% ± 0.08%	
60%	40%	Astr	92.18% ± 0.85%	92.65% ± 0.61%	92.30% ± 0.28%	96.86% ± 0.13%
		Non_Astr	98.12% ± 0.15%	98.00% ± 0.24%	98.03% ± 0.09%	

containing astronomical data and 1.06%, 0.22%, −0.03%, 0.08% in the other types of traffic, respectively.

Random testing did not validate the results well. Then, we used a ten-fold cross-validation averaging method to test the accuracy of the classification model. A comparison between 1D-CNN and 2D-CNN models for various types of evaluation metrics in binary classification is shown in Tables 3 and 4. The Tables show that both groups of experiments achieve good classification results with classification accuracy mostly above 96.00%. The accuracy, recall and F1-score of the classification of traffic containing astronomical data are in the 90.65%–94.05%, and other types of traffic classification are in the 97.72%–98.48%. Specifically, the total accuracy of 1D-CNN are higher than 2D-CNN, as much as 0.34%, 0.13%, 0.16% and

Table 4
Results of Two-class Experiments (2D-CNN)

Data Set		Classes	Results			
Train Size	Test Size		Precision	Recall	F1-score	Total Accuracy
90%	10%	Astr	91.23% ± 0.82%	93.29% ± 1.52%	92.20% ± 0.50%	96.79% ± 0.19%
		Non_Astr	98.29% ± 0.38%	97.72% ± 0.27%	97.98% ± 0.11%	
80%	20%	Astr	91.56% ± 0.86%	90.65% ± 1.23%	91.21% ± 0.59%	96.84% ± 0.21%
		Non_Astr	97.95% ± 0.26%	98.16% ± 0.21%	98.07% ± 0.13%	
70%	30%	Astr	93.16% ± 0.28%	91.94% ± 1.28%	92.57% ± 0.65%	96.76% ± 0.26%
		Non_Astr	97.74% ± 0.35%	98.08% ± 0.09%	97.93% ± 0.16%	
60%	40%	Astr	91.67% ± 1.18%	92.46% ± 2.84%	92.02% ± 0.82%	96.75% ± 0.26%
		Non_Astr	98.08% ± 0.71%	97.84% ± 0.40%	97.95% ± 0.16%	

Table 5
Results of Multi-class Experiments (1D-CNN)

Classes	Precision	Recall	F1-score
Pulsar Data-I	79.33% ± 9.50%	68.02% ± 3.12%	72.83% ± 2.35%
Pulsar Data-II	62.00% ± 9.50%	76.51% ± 5.95%	67.49% ± 3.54%
VLBI Data	79.75% ± 3.75%	90.61% ± 3.02%	84.65% ± 1.24%
Optical Data-I	99.18% ± 0.27%	95.57% ± 1.76%	97.21% ± 0.91%
Optical Data-II	98.85% ± 0.64%	96.75% ± 0.73%	97.61% ± 0.51%
Optical Data-III	89.92% ± 3.42%	81.89% ± 2.23%	85.35% ± 0.65%
		Total Accuracy:	81.79% ± 0.77%

Table 6
Results of Multi-class Experiments (2D-CNN)

Classes	Precision	Recall	F1-score
Pulsar Data-I	75.83% ± 5.33%	63.82% ± 1.86%	69.14% ± 1.13%
Pulsar Data-II	58.00% ± 5.33%	74.63% ± 3.20%	64.98% ± 2.16%
VLBI Data	77.33% ± 3.00%	86.77% ± 3.23%	80.90% ± 1.00%
Optical Data-I	94.29% ± 1.90%	94.64% ± 2.05%	94.87% ± 1.07%
Optical Data-II	97.96% ± 0.76%	94.13% ± 1.13%	95.85% ± 0.43%
Optical Data-III	87.83% ± 3.50%	79.41% ± 2.87%	83.47% ± 1.10%
		Total Accuracy:	78.96% ± 0.55%

0.11%. For the classification of traffic containing astronomical data, the precision of 1D-CNN are higher than 2D-CNN, 0.63%, 0.12%, 0.15% and 0.51% higher on average; the recall of 1D-CNN are higher than 2D-CNN, 0.76%, 0.88%, 0.73% and 0.19% higher on average; the F1-score of 1D-CNN are higher than 2D-CNN, 0.76%, 0.38%, 0.41% and 0.28% higher on average. For other types of traffic classification, the precision of 1D-CNN are higher than 2D-CNN, 0.19%, 0.19%, 0.20% and 0.04% higher on average; the recall of 1D-CNN are higher than 2D-CNN, 0.15%, 0.02%, 0.05% and 0.16% higher on average; the F1-score of 1D-CNN are higher than 2D-CNN, 0.22%, 0.08%, 0.10% and 0.08% higher on average. In other words, 1D-CNN and 2D-CNN are suitable for solving the classification problem of traffic containing astronomical data, and 1D-CNN has better performance than 2D-CNN, because the traffic data is one-dimensional and 1D-CNN is more suitable for these data.

4.3. Multi-class Experimental Results

The effectiveness of a CNN in classifying traffic containing astronomical data was verified by conducting multiple classification experiments using 1D-CNN and 2D-CNN models. The results obtained from experiments using ten-fold cross-validation to evaluate the classification results of 1D-CNN and 2D-CNN models are shown in Figure 7. It can be observed that the 1D-CNN model performs better than the 2D-CNN model in terms of accuracy, recall and F1 score. For comparison, the mean values of the above three indicators are given in Tables 5 and 6. It can be observed that the 1D-CNN model exhibits better performance than the 2D-CNN model, with the Total Accuracy increase of 2.83%. Specifically, (i) for the traffic containing pulsar data-I, the precision, recall and F1-score of 1D-CNN are higher than 2D-CNN, 3.50%, 4.20% and 3.69% higher on average; (ii) for the traffic containing pulsar data-II, the precision, recall and F1-score of 1D-CNN are higher than 2D-CNN, 4.00%, 1.88% and 2.51% higher on average; (iii) for the traffic containing VLBI data, the precision, recall and F1-score of 1D-CNN are higher than 2D-CNN, 2.42%, 3.84% and 3.75% higher on average; (iv) for the traffic containing optical data-I, the precision, recall and F1-score of 1D-CNN are higher than 2D-CNN, 4.89%, 0.93% and 2.34% higher on average; (v) for the traffic containing optical data-II, the precision, recall and F1-score of 1D-CNN are higher than 2D-CNN, 0.89%, 2.62% and 1.76% higher on average; (vi) for the traffic containing optical data-III, the precision, recall and F1-score of 1D-CNN are higher than 2D-CNN, 2.09%, 2.48%

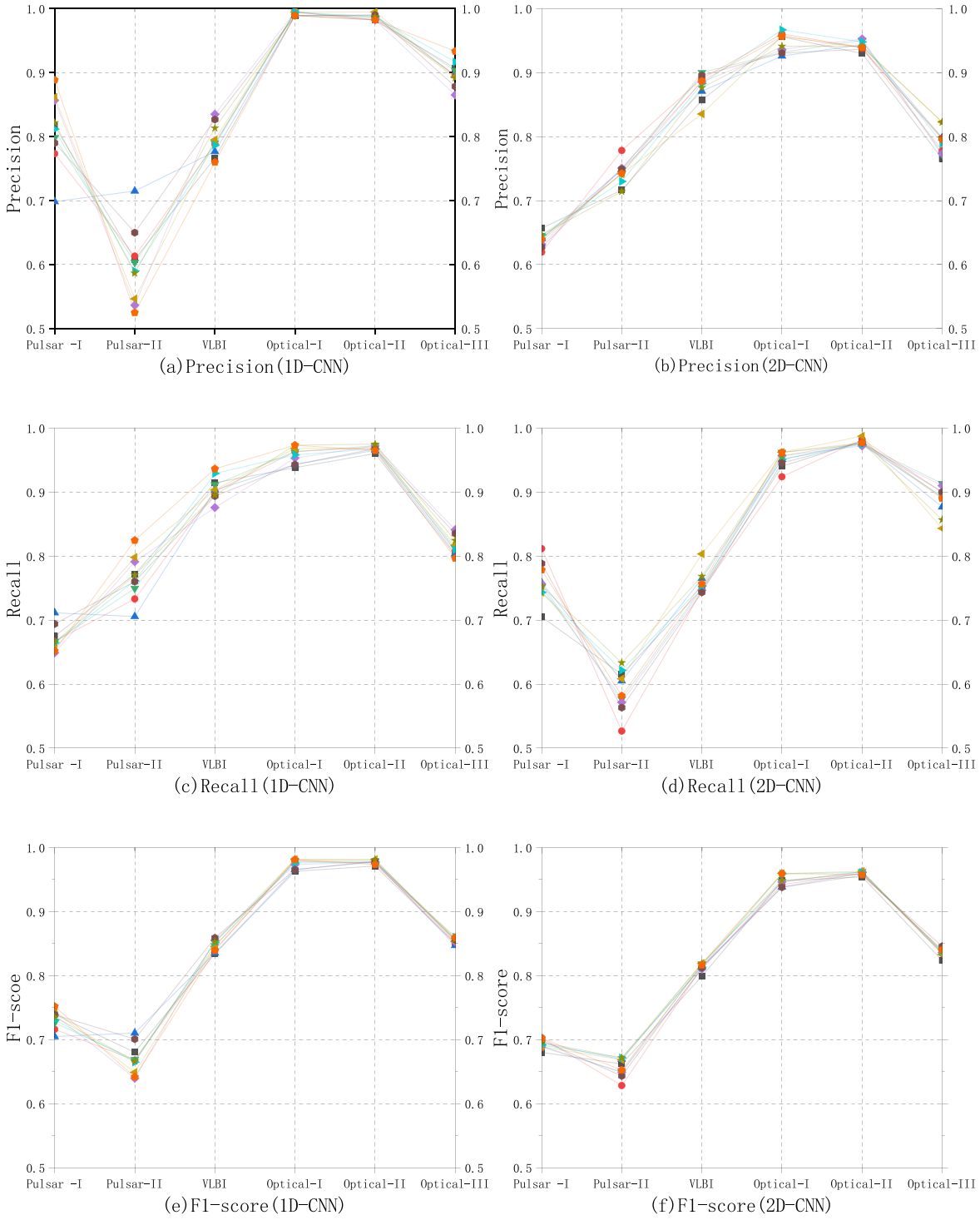

Figure 7. Ten-fold cross-validation compares the precision, recall and F1-score of 1D-CNN and 2D-CNN models in multi-class experimental results.

and 1.88% higher on average. That means, 1D-CNN has better performance than 2D-CNN, because 1D-CNN directly learns features from raw traffic automatically and directly outputs the predicted labels.

5. Conclusions

This paper proposes a classification method using deep learning for traffic containing astronomical data considering some network traffic classification methods, such as the method based on port matching, the depth packet detection method, the stream statistical feature method and the machine learning method. In this method, the identification problem of traffic containing astronomical data was transformed into a typical classification task by employing multiple deep feature learning and using CNNs to design a classifier that fits the characteristics of astronomical data transmission. The traffic containing an astronomical data set was formed by capturing the network traffic in the XAO data transmission network, whereas the non-astronomical transmission traffic data set was formed using part of the non-VPN data set in ISCX-2016 for classification experiments. The experimental results indicated that the proposed method can effectively distinguish the traffic containing mixed astronomical data from the traffic containing non-astronomical data in the network and achieve accurate identification of traffic containing astronomical data. In future work, we plan to capture more and more detailed astronomical data contained in transmission traffic in the XAO data transmission network to verify the effectiveness of the proposed method.

Acknowledgments

This work is supported by the National Key R&D Program of China (Nos. 2021YFC2203502 and 2022YFF0711502); Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01A360); the National Natural Science Foundation of China (NSFC) (12173077 and 12003062); the Tianshan Innovation Team Plan of Xinjiang Uygur Autonomous Region (2022D14020); the Scientific Instrument Developing Project of the Chinese Academy of Sciences (Grant No. PTYQ2022YZZD01); the Operation, Maintenance and Upgrading Fund for Astronomical Telescopes and Facility Instruments, budgeted from the Ministry of Finance of China (MOF) and administrated by the Chinese Academy of Sciences (CAS). This work is supported by Astronomical Big Data Joint Research Center, co-founded by National Astronomical Observatories, Chinese Academy of Sciences.

ORCID iDs

Xu Du https://orcid.org/0000-0001-6448-0822

References

Akinsola, J. E. T. 2017, IJCTT, 48, 128
Chollet, F. 2017, in IEEE CVPR (Honolulu, HI, USA: IEEE)
El-Maghraby, R., Elazim, N., & Bahaa-Eldin, A. 2017, in ICCES (Cairo, Egypt: IEEE)
Guo, Y., Gao, Y., Wang, Y., et al. 2017, Procedia Engineering, 174, 1309
He, K., Zhang, X., Ren, S., & Sun, J. 2016, in IEEE CVPR (Las Vegas, NV, USA: IEEE)
Krizhevsky, A., Sutskever, I., & Hinton, G. 2012, Advances in Neural Information Processing Systems, 2012, 25
Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proc. IEEE, 86, 2278
Madhukar, A., & Williamson, C. L. 2006, in 14th Int. Symp. on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2006) (Monterey, CA, USA: IEEE)
Rezaei, S., & Liu, X. 2018, IEEE Commun., 57, 76
Shafiq, M., Yu, X., Laghari, A., et al. 2016, in IEEE ICCC (Chendgu, China: IEEE)
Wang, J., Zhang, H., Wang, N., et al. 2022, RAA, 22
Wang, Y., Yun, X., & Zhang, Y. 2015, in ICNP, Vol. 22 (San Francisco, CA, USA: IEEE)
Yang, D. B. 2019, in ITNEC (Chengdu, China: IEEE), 1887
Zhang, H., Demleitner, M., Wang, J., et al. 2019, AdAst, 2019, 5712682
Zhang, H., Wang, J., Demleitner, M., et al. 2022, ACOM, 39, 100578
Zhang, J., Xiang, Y., Wang, Y., et al. 2012, IEEE TPDS, 24, 104